UNIVERSITY OF CALIFORNIA,
IRVINE

Investigation of protein sequence-structure dynamics using bioinformatics, molecular dynamics and machine learning

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Vy T. Duong

Dissertation Committee:
Professor Rachel W. Martin, Chair
Professor Ray Luo
Professor David Mobley

2020

DEDICATION

To my parents, Thuy Duong and Tuan P. Lam,
who are have supported me and inspire me with their strength and perseverance

to Daniel Ramirez-Guerrero,
for lighting the way during my Ph.D., providing emotional support, and everyday inspiring me to be a better person

and finally to all my family and friends,
for their continuing love, support and kindness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my appreciation to my Ph.D. advisors and mentors, Professor Rachel W. Martin, and Professor Ray Luo for being supportive and caring mentors. She provided enthusiasm, optimism, and wisdom to move projects forward from bioinformatics to protein dynamics projects. From a computational background, Professor Ray Luo nurtured my computational skills, providing in-depth mentoring regarding analytics, molecular dynamics simulations, and reminding me to take care of my health by drinking ginseng or exercising. Both advisors were great mentors whose aid helped me grow as a scientist. With their guidance and mentoring, I was able to produce this thesis as well as publish the work described herein.

I am also thankful to my other dissertation committee member, Professor David Mobley, In addition to my advancement committee members, Professors Elizabeth Read, Markus W. Ribbe, James Nowick, and Andrej Luptak for detailing constructive commentary regarding my projects. I am also grateful to Martin, Luo, Butts, Vanderwal, and Wang lab members for collaborative ideation, brainstorming, and input on projects. These groups were great colleagues and I appreciate their help and patience. I am also thankful for project and analytical input from Professor Carter T. Butts. I would also like to thank my collaborators Professors Wenqi Wang and Chris Vanderwal, it was great obtaining their experimental input and the dialogue/ideatino I had with these two groups. I would also like to thank Dr. Yibo Wang and Dr. Richard P. Donovan for providing computational resources at the Calit2 Think Tank, and aiding me whenever I needed help.

Previous and current postdocs and graduate students have been an invaluable helpful resource. For lighting the path at the beginning of my Ph.D. I would also like to give thanks to Andrew Schaub for his mentorship, friendship, and project input. From providing mentorship, input, and encouragement, I am grateful to Dr. Gianmarc Grazioli for his helpful input and mentorship, he was essentially a third unofficial advisor to me and I am confident he will be an amazing advisor at SJSU. I also can't thank my lab members from Martin and Luo lab enough for their support and advice, Ruxi Qi, Haixin Wei, Shiji Zhao, Edward King, Erick Aitchinson, Dr. Changhao Wang, Dr. D'artagnan Greene, Terry Lambros, Jan Bierma, Kyle Roskamp, Megha Unhelkar, Jessica Kelz, and Marc Piercy. I would also like to give thanks to undergraduates who have dedicated their time and efforts in helping me, Amal El Ali, Zihao (Henry) Chen, and Danessa Yip.

I would also like to thank Elsevier Inc., Oxford University Press, EMBO Press, and the American Chemical Society for allowing me to reproduce Chapters 2-5. In addition, I would like to thank the MCSB department and administrators Karen Martin, Cely Dean, and Naomi Carreon for all of their help and advice during my academic training. In addition, funding for two years was provided by the Mathematical, Computational and Systems Biology Pre-doctoral Training Grant T32 EB009418-08. Funding for several projects is sourced from NSF award DMS-1361425.

VITA

Vy T. Duong

EDUCATION

Doctor of Philosophy in Chemistry **2020**
University of California Irvine *Irvine, CA*

Bachelor of Arts in Integrative Biology **2013**
University of California Berkeley *Berkeley, CA*

RESEARCH EXPERIENCE

Graduate Student Researcher **2015–2020**
Research Advisors: Dr. Rachel W. Martin, Dr. Ray Luo
University of California, Irvine *Irvine, California*

Investigations of a variety of protein systems, namely plant metabolic proteins and intrinsically disordered proteins, using bioinformatics, molecular dynamics simulation, and machine learning.

Undergraduate Student Researcher **2011–2013**

Research Advisors: Dr. Maya DeVries, Dr. Roy Caldwell
University of California, Berkeley *Berkeley, California*

Investigations of stomatopod memory and behavioral patterns.

JOURNAL PUBLICATIONS

1. Duong, V.T., Grazioli, G., Butts C.T., Martin, R.W, 2020. Neural upscaling from coarse protein structure networks to atomistic structures. *To be submitted 2020.*

2. Vargas, R.E., Duong, V.T., Han, H., Ta, A.P., Chen, Y., Zhao, S., Yang, B., Seo, G., Chuc, K., Oh, S. and El Ali, A., 2019. Elucidation of WW domain ligand binding specificities in the Hippo pathway reveals STXBP4 as YAP inhibitor. The EMBO journal. `https://doi.org/10.15252/embj.2019102406`

3. Duong, V.T., Chen, Z., Thapa, M.T. and Luo, R., 2018. Computational Studies of Intrinsically Disordered Proteins. The Journal of Physical Chemistry B, 122(46), pp.10455-10469. `http://dx.doi.org/10.1021/acs.jpcb.8b09029`

4. Duong, V.T., Unhelkar, M.H., Kelly, J.E., Kim, S.H., Butts, C.T. and Martin, R.W., 2018. Protein structure networks provide insight into active site flexibility in es-

terase/lipases from the carnivorous plant Drosera capensis. *Integrative biology*, 10(12), pp.768-779. http://dx.doi.org/10.1039/C8IB00140E

5. Ellis, B.D., Milligan, J.C., White, A.R., Duong, V., Altman, P.X., Mohammed, L.Y., Crump, M.P., Crosby, J., Luo, R., Vanderwal, C.D. and Tsai, S.C., 2018. An oxetane-based polyketide surrogate to probe substrate binding in a polyketide synthase. Journal of the American Chemical Society, 140(15), pp.4961-4964. http://dx.doi.org/10.1021/jacs.7b11793

6. Unhelkar, M.H.†, Duong, V.T.†, Enendu, K.N., Kelly, J.E., Tahir, S., Butts, C.T. and Martin, R.W., 2017. Structure prediction and network analysis of chitinases from the Cape sundew, Drosera capensis. Biochimica et Biophysica Acta (BBA)-General Subjects, 1861(3), pp.636-643. http://dx.doi.org/10.1016/j.bbagen.2016.12.007 († = co-first author)

POSTERS/PRESENTATIONS

Posters: Center for Complex Biological Systems Conference Annual Retreat (2017), The 2nd annual Southern California Theoretical Chemistry Symposium (2017), Biophysical Annual Meeting (2018), UCI Chemistry Recruitment Symposium (2018), MCSB Recruitment Symposium (2019); Vertex Day (2020)

Presentations: Molecular Dynamics Seminar (2019), UCI Calit2 Think Tank Invited Speaker (2020), Virtual Synthetic and Chemical Biology Club Seminar (2020)

TEACHING EXPERIENCE

Teaching Assistant – General Chemistry Laboratory (CHEM 51LC) 2016
Teaching Assistant – Organic Chemistry Laboratory (CHEM 1LE) 2017

ABSTRACT OF THE DISSERTATION

Investigation of protein sequence-structure dynamics using bioinformatics, molecular dynamics and machine learning

by

Vy T. Duong

Doctor of Philosophy in Chemistry

University of California, Irvine, 2020

Professor Rachel W. Martin, Chair

As genomic repositories increasingly grow with a variety of data from a multitude of organisms, the need to approach extracting and interpreting data also becomes increasingly difficult. Recent advances in protein annotation and structure prediction have improved, however the variety and sheer amount of data requires unique approaches from multiple different disciplines. Bioinformatics yields important functional sequence information and classification. Molecular dynamics (MD) simulation allows for the interrogation of biochemical systems at the atomistic level. Combined with machine learning, these disciplines can be equipped to investigate the complex functions and relationships of proteins within the current abundant genomic landscape.

The objective of this dissertation is to outline complementary methodologies from various fields - bioinformatics, molecular dynamics simulation, and machine learning - that together, can investigate vast genomic repositories, functional protein data.

Aim 1: The development of the bioinformatics and *in silico* maturation pipeline consists of gene annotation, MD simulation to equilibrate predicted proteins, and statistical methods adopted from graph theory in collaboration with the Butts lab. Proteins can be represented in graph theoretic terms allowing for the exploration of diverse protein structural features.

Aim 2: Molecular dynamics simulation gives rise to atomic level details of complex systems. A variety of protein systems - HIV Rev, short intrinsically disordered peptides, STXPB4, YAP-1 WW domain - explored are intrinsically disordered. MD simulations were used to simulate the complexities and difficulties encountered within these proteins as well as plant metabolic proteins.

Aim 3: After the aforementioned bioinformatics pipeline and *in silico* molecular dynamics-based maturation of predicted proteins, methods to extract useful atomistic information from coarse protein structure networks (PSNs) were developed. A multi-layer perceptron was used to essentially upscale coarse PSNs into atomistic models. The significance of this technique permits for the simulation of coarse PSNs, and the exploration of complex protein structural conformations.

Chapter 1

Introduction

1.1 Intrinsically disordered proteins/regions

It was once strongly presumed proteins required rigid secondary structure to function. More recently, the scientific community has largely discounted this and accepted the overall prevalance of intrinsically disordered regions (IDRs) and fully intrinsically disordered proteins (IDPs). These proteins are found in all three domains of life – archaea, bacteria, eukarya – as well as all viruses studied to date [293]. Viruses in particular function with minimial protein production, requiring adaptive proteins to bind to a multitude of different targets and perofrm a variety of functions [283, 11]. Recent computational research has also suggested with increasing organismal complexity, the presence of IDRs/IDPs also increases [328, 316, 76]. Among archaea and eukarya, IDRs with lengths of approximately >30 amino acids are comparably similar, however these computational studies suggest much more is present in eukaryotic organisms [213, 316, 76, 329, 195]. IDRs/IDPs participate in wide variety of important cellular actiivities to maintain functions encompassing recognition, assembly, and modification of other proteins/molecular compounds [73, 74]. Their folding

properties also make these proteins elusive targets for structural characterization. Many IDPs/IDRs exhibit coupled folding and binding properties, only forming a more defined structure upon binding to a specific partner, thus also making apo conformations difficult to structurally characterize [192].

These proteins are also implicated in a wide variety of diseases such as cardiovascular diseases, diabetes, neurodegenerative diseases, cancer, etc [293]. Examples such as α-synuclein, p53, amyloid-β, and tau protein are proteins of considerable interest in the scientific community and public health [293]. Computationally, a combination of sequence and structural studies are required to tackle the arduous investigation of these elusive, complex proteins.

1.2 Diversity and complexity of plant metabolic proteins

Plant metabolic proteomes remain largely unexplored in a wide variety of species. However, these contain potentially useful proteins that can be used in agriculture, biomedical purposes, and a plethora of other applications. Plant proteins from sources such as soy, wheat, and corn also have a lower likelihood of inducing immunogenic responses in the human body compared to animal-based proteins. For example, the commonly used bovine collagen used in most medical procedures have been reported to cause negative reactions [212]. Therefore, the investigation of proteins involved in plant metabolism can yield potentially useful biomolecular tools in generating useful compounds such as flavonoids.

Herein, this dissertation focuses on mainly *Drosera capensis* and its unique metabolic biomolecular machinery. Bioinformatics, MD simulation, and statistical techniques are combined to investigate two proteins classes in *Drosera capensis*.

1.3 Bioinformatics and genomic studies

In the Uniparc database, there is approximately 250 million protein sequences, very few of which have been fully structurally characterized [150]. A wide spectrum of functional and structural diversity is present amongst these vastly unexplored data. With the rapid development of high-throughput techniques, the availability of sequence data has thus quickly outpaced the production of structural data in recent decades. The sheer expansion of genomics, proteomics, and transcriptomics data has also motivated researchers to develop technologies to obtain meaningful interpretation of these repositories. Both sequence and structural approaches are imperative in investigating the underlying biochemical machinery of the the vast number of unexplored proteins found in nature.

Of the proteins that require IDPs are abundant in specific amino acids compositions, specifically polar and unstructured residues (Gly, Pro, Arg, Glu, Gln, Ser, Lys) [293]. Bioinformatics analysis facilitates the prediction of IDRs/IDPs as shown by the DISOPRED3 prediction server [128]. In combination with large plant proteome repositories, bioinformatics also facilitates the exploration of *Drosera capensis* explored in this dissertation.

1.4 Molecular dynamics simulation

Experimental characterization techniques (e.g. X-ray crystallography, NMR, etc.) are integral in furthering current understanding of complex protein dynamics. However these methodologies capture mainly rigid snapshots or average approximations of complex systems with a plethora of different conformations and behavior. To obtain a more expansive range of conformations and behavior, molecular dynamics (MD) simulations have been extensively utilized to explore systems such as hinge movement in active site opening and closing [62], tRNA flexibility [106], ligand binding in heme proteins [47], and a multitude of

other systems. MD simulations have developed rapidly from the first picosecond simulation of bovine pancreatic trypsin inhibitor in 1977 to current capabilities [186].

Molecular dynamics (MD) simulations approximate the forces acting upon atoms via numerical solution of the classical equations of motion: $F = -\nabla U(r^N)$. The forces acting upon an atom are derived from a potential energy function, $U(r^N)$, where 3N coordinates are represented as $r^N = (r_1, r_2, ...r_N)$. Ranging between femtoseconds to microseconds for most simulations, MD simulations approximate protein dynamics on timescales inaccessible to traditional structural biology techniques. In this thesis, AMBER [46, 45, 243] MD software suite is the primary tool used to generate simulations and investigate multiple protein systems, ranging from plant metabolic proteins to IDPs. Other popular MD simulation software alternatives consist of CHARMM [31], GROMOS [248], and NAMD [218].

Although MD simulations are useful for exploring timescales from nanosecond to microseconds, they are limited by the incapacity to simulate beyond microsecond timescales, with few research groups having the resources to simulate in millisecond timescales. In the next section, I explore the incorporation of machine learning in the investigation of complex protein dynamics.

1.5 Machine learning

Machine learning algorithms have rapidly expanded as an essential utility across a variety of different fields. In the field of structural biology, machine learning has established prominence in topics ranging from structure prediction to modeling functional properties of proteins. In 2018, Deepmind's AlphaFold convolutional neural network (CNN) model won the CASP18 competition by a wide margin, demonstrating the potential of incorporating neural network models in structure biology [251]. Two major implementations of machine learning described

herein consist of unsupervised and supervised learning. Unsupervised learning encompasses two possible tasks, either clustering data or learning potential groupings. These are standard, common techniques used by computational chemists/biologists to group sequence data, MD simulation data, or other structural biology data. Algorithms range from classical algorithms such as hierarchical, DBScan, K-means, etc. or more advanced techniques such as autoencoder neural networks. Supervised learning consists of mainly regression (prediction of a dependent variable from one or more independent variables) or classification (prediction of qualitative labels from one of more input variables). Algorithms range from classical techniques such as random forest to neural networks models (e.g. multilayer perceptons, CNN, RNN, LSTM, etc.)

This thesis also focuses primarily on the implementation and results from various unsupervised and supervised learning on MD simulation and bioinformatics data. These techniques greatly expand the ability to extract and learn new features of protein dynamics or commonalities between protein sequence data.

1.6 Objectives of dissertation

Each methodology – bioinformatics, MD simulation, or machine learning – has their individual advantages and disadvantages. Proteins with high sequence identity can often have drastically different functions. For instance, ovalbumin, the most abundant protein in egg white for instance belongs to the serpin protein superfamily. Despite sharing high sequence and structural similarity to other serpins, ovalbumin however lacks the standard serpin function to inhibit serine proteases [115]. This example demonstrates the need of a multifaceted approach regarding the investigation of protein sequence, structure, and function.

This dissertation probes the elucidation of protein sequence-structure-function relationship

via the merge of bioinformatics, MD simulation, and machine learning techniques. Chapter 2 delves into the sequence and structural features of *Drosera capensis* chitinases, and the unique findings of these proteins used to either break the chitinous exoskeletons of insect prey and/or defend itself against fungal pathogens. In the subsequent chapter 3, *Drosera capensis* esterase/lipases are explored using analyses comparable to chapter 1 with the exception of principal component analyses of network-based structural data. Chapter 4 investigates multiple WW domain proteins using proteomic anaysis, MD simulation, and unsupervised learning. Chapter 5 compares the performance of ff14SB and IDP-specific force field ff14IDPSFF in their abilities to recapitulate experimental measurements of multiple IDP systems. Chapter 6 explores the merger of MD simulation, graph-based networks, and machine learning to backmap/upscale contact adjacency matrices to atomistic coordinate models.

Chapter 2

Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*.

2.1 Summary

Background: Carnivorous plants possess diverse sets of enzymes with novel functionalities applicable to biotechnology, proteomics, and bioanalytical research. Chitinases constitute an important class of such enzymes, with future applications including human-safe antifungal agents and pesticides. Here, we compare chitinases from the genome of the carnivorous plant *Drosera capensis* to those from related carnivorous plants and model organisms.

Methods: Using comparative modeling, *in silico* maturation, and molecular dynamics simulation, we produce models of the mature enzymes in aqueous solution. We utilize network analytic techniques to identify similarities and differences in chitinase topology.

Results: Here, we report molecular models and functional predictions from protein structure networks for eleven new chitinases from *D. capensis*, including a novel class IV chitinase with two active domains. This architecture has previously been observed in microorganisms but not in plants. We use a combination of comparative and de novo structure prediction followed by molecular dynamics simulation to produce models of the mature forms of these proteins in aqueous solution. Protein structure network analysis of these and other plant chitinases reveal characteristic features of the two major chitinase families.

General Significance: This work demonstrates how computational techniques can facilitate quickly moving from raw sequence data to refined structural models and comparative analysis, and to select promising candidates for subsequent biochemical characterization. This capability is increasingly important given the large and growing body of data from high-throughput genome sequencing, which makes experimental characterization of every target impractical.

Highlights:

We report eleven new chitinases from the carnivorous plant *Drosera capensis*. A novel two domain class IV chitinase similar to those found in microbes was found. Protein structure prediction and comparison to other carnivorous plant chitinases reveals commonalities. Sequence and structural motifs are conserved among carnivorous plant chitinases. Protein structure networks reveal structural differences and predict functionality.

2.2 Introduction

Chitin, a polymer of β-(1,4)-N acetylglucosamine (GlcNAc), is the second-most abundant biopolymer [136]. Chitinases (EC 3.2.1.14) are ubiquitous even among organisms that do not produce chitin, with the latter employing them for purposes of digestion and/or defense. These enzymes cleave chitin at the β-1,4 linkage of N-acetyl glucosamine units, although substantial variation in activity and substrate specificity exists. Some chitinases can also cleave peptidoglycans at β-1,4 linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine, and chitodextrins between N-acetyl-D-glucosamine units. Plant chitinases sometimes have multiple functionalities; some display lysozyme activity [229], while others have a calcium storage function [183]. In humans, chitinases are produced in response to fungal infections, a feature of the innate immune system that is suppressed in immunocompromised individuals, including AIDS patients, transplant recipients, and burn victims [298]. These enzymes and related chitin-binding proteins are expressed in human lung tissue, where they are dysregulated in cystic fibrosis and asthma [169].

In plants, these enzymes are expressed in response to environmental stress and pathogen or pest infestation [33], driving efforts to overexpress particularly effective examples in transgenic crop plants [132]. Carnivorous plants use chitinases as part of the prey capture response: active chitinases have been found in the pitcher fluid of *Nepenthes* [79, 238], and in the digestive fluids of the Venus flytrap [206]. However, the extent to which chitin is used as a nitrogen source remains controversial. *Drosera capensis* plants fed on chitin incorporate its nitrogen into their leaf tissue; however nutrient uptake is less efficient than for plants fed on protein [207]. Examination of insect carcasses after digestion reveals that 40-60% of the total nitrogen is unused [7, 208], consistent with the observation that the remains of insect exoskeletons appear mostly intact [130]. However, chitinase expression is upregulated in the presence of prey in the related species *Nepenthes alata*. In *Drosera rotundifolia*, an increase in both expression of chitinase mRNA and chitinase activity was induced by addition of

crustacean chitin with mechanical stimulation of the traps [185]. The prey-induced induction of chitinase activity, despite the low efficiency of chitin use, may indicate that chitinases primarily function to inhibit fungal growth in the traps, just as cytotoxic peptides discourage microbial growth in the fluid of *Nepenthes* pitchers [108, 32].

Here, we compare novel chitinases recently discovered from the genome of the Cape sundew (*Drosera capensis*) [36], to those from other carnivorous plants in order Caryophylalles. The conservation of the overall protein folds and active site architectures suggests that many of the *D. capensis* chitinase sequences form functional enzymes. We use sequence analysis, comparative modeling with all-atom refinement followed by *in silico* maturation [38], and investigation of protein structure networks to identify structurally distinct subgroups of proteins for subsequent expression and biochemical characterization.

2.3 Results and Discussion

2.3.1 Two Distinct Families of Carnivorous Plant Chitinases Are Found

Gene sequences annotated as coding for chitinases using the MAKER-P (v2.31.8) pipeline [40] and a BLAST search against SwissProt (downloaded 8/30/15) and InterProScan [225] were clustered by sequence similarity, along with chitinases previously identified from *Dionaea muscipula* [206] and various species of *Drosera* and *Nepenthes* [232]. Annotated sequence alignments of the Family 18 and Family 19 chitinases are shown in Supplementary Figures A.1 and A.2, respectively. We have identified four fragments ranging from 41%-100% identity to the DcChit1_1 fragment previously found by Renner and Specht in *D. capensis* genomic DNA [232] (Supplementary Figure A.3). Several well-characterized reference sequences (e.g chitinases from *Vitis vinifera*, *Brassica napus*, and *Hordeum vulgare*) are also included for

comparison. Using the characterization scheme of the carbohydrate-active enzymes (CAZy) database [42, 162], the chitinases investigated here belong to Family 18 (orange) or Family 19 (green). Overall, the sequence identity among the Family 18 chitinases from Caryophylalles carnivorous plants is much higher than that of Family 19, as illustrated in Figure 3.2A and B. These two types of chitinases have different folds and are thought to have evolved independently, [189, 166], consistent with their separation into separate clusters (Figure 3.2C). Family 18 contains types III and V, while types I, II and IV belong to Family 19 [206].

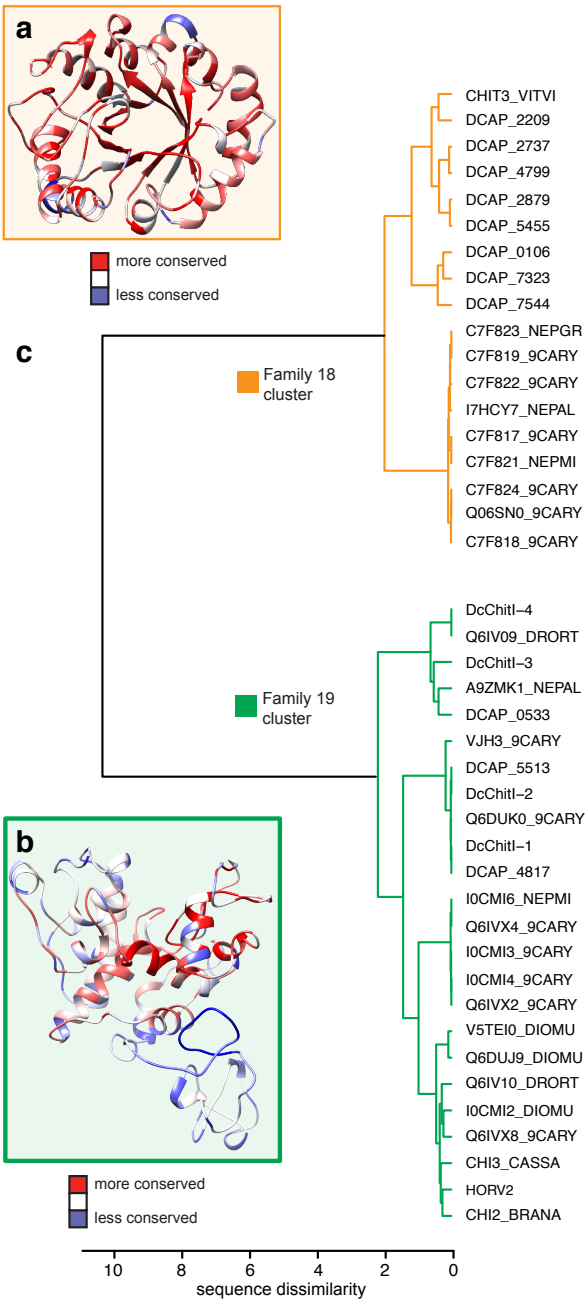


Figure 2.1: Clustering of chitinases identified from the *D. capensis* genome, compared with those from other Caryophylalles carnivorous plants and well-characterized reference sequences. All of the sequences examined belong to GH Families 18 or 19. The sequence dissimilarity used here is the e-distance metric of Székely and Rizzo [273] (with $\alpha = 1$). This parameter is a weighted function of within-cluster similarities and between-cluster differences with respect to a user-specified reference metric, defined here as the raw sequence dissimilarity (1 - (%identity)/100).

2.3.2 *D. capensis* Chitinases are Predicted to Adopt Folds Consistent with Active Enzymes

Family 18 chitinases, which retain the β-anomeric carbon stereochemistry from the substrate to the product, adopt the $(\alpha\text{-}\beta)_8$ triosephosphateisomerase (TIM)-barrel fold [134, 294], shown for DCAP_0106 in Figure 2.2A. The *in silico* maturation process, which we have previously described for cysteine proteases [38], is illustrated in Supplementary Figure A.4. The active site (Figure 2.2B), consists of a characteristic DXXDXDXE motif [134, 294]. The "tunnel" containing the active site is shaped by an unusual structural feature, two non-proline cis peptide bonds that are highly conserved, although the particular residues involved are somewhat variable [285, 183]. The cis peptide bonds (shown in black in Figure 2.2C), are captured by the molecular models for all full-length Family 18 chitinases examined here. The shape of the tunnel and the surface formed by the aromatic rings opposite the catalytic D and E residues acts to guide the chitin polymer chains into the active site, leading to processive activity [112]. The ability of Family 18 chitinases to keep the strand that is currently being degraded from re-encountering solid substrate is thought to be a key determinant of their ability to hydrolyze crystalline polysaccharides [305].

The Family 19 chitinases, all of which are characterized by an anomeric inverting mechanism [274], have diverse structural features. Much of the structural and functional diversity results from two highly variable regions, the C-rich chitin-binding domain and the P-rich hinge [188, 198], each of which may vary in length or be absent altogether. We have identified two class I chitinases (DCAP_4817 and DCAP_5513) and one class IV chitinase (DCAP_0533) from the *D. capensis* genome. Most of the sequences in this set contain N-terminal secretion signals, however two *D. spatulata* sequences (Q6IVX2_9CARY and Q6IVX4_9CARY) and the reference sequence CHI2_BRANA contain short C-terminal extensions indicating targeting to the vacuole, consistent with their playing a purely defensive role. One sequence each from *D. capensis* (DCAP_5513), *D. rotundifolia* (Q6IV09_DRORT), and *D. spatulata*

(Q6DUK0_9CARY) is missing one or more critical active site residues; in other organisms, enzymatically non-functional chitinase homologs are often present and can serve as chitin-binding proteins [222]. The predicted structure after *in silico* maturation for a representative chitinase, VF-1 from *D. muscipula* (Figure 2.2) is in good agreement overall with the homology model of Paszota et al. [206], with the active site residues positioned in a shallow cleft on the surface of the active domain. The two models do differ in the relative orientations of the domains; however examination of the other models in this set suggests that the P-rich hinge is highly flexible (Supplementary Figure A.5).

Because sequence identity between our targets and proteins with solved structures is only moderate (in the range of 30-50 %), comparative modeling with all-atom refinement was used. The starting structures are predicted using the Robetta implementation [139] of Rosetta [228]. This approach uses a combination of fragment homology and de novo structure prediction, and is regularly validated via CAMEO [99]. Our modeling approach, in which the starting Rosetta structures are subjected to in silico maturation, was previously validated experimentally when the x-ray structure of a cysteine protease we had previously predicted was solved. The crystal structure of Dionain 1 (PDB ID 5A24) [234], shows excellent agreement with our predicted structure, with the prediction capturing all major secondary structural elements and exhibiting only minor deviations in the flexible loop regions [38]. For the chitinases, fragment homology was the primary method used. Sequence alignments for the target molecule with all of the template sequences used by Rosetta are shown for representative members of Family 18 and Family 19 in Supplementary Figures A.6 and A.7, respectively. For DCAP_2209 (Family 18), excluding the N-terminal signal sequence, 100% of the sequence aligns with homologous regions in the 11 template sequences (tabulated in Supplementary Table A.1). For DCAP_5513, excluding the N-terminal signal sequence, only one 6-residue stretch of the P-rich region is not directly homologous to at least one of the template sequences (tabulated in Supplementary Table A.2). As a further validation, a blind structure prediction was performed for the reference sequence HORV2, in

which the actual pdb structure of this molecule (PDBID 1CNS, 2BAA) [264] was excluded from the template set. The predicted and experimental structures are shown overlaid in Supplementary Figure A.8. After equilibration, the backbone RMSD between these structures was 1.01 Å. All major secondary structure elements are reproduced, with only minor differences in relative orientation as well as some deviation in the loops and termini.

2.3.3 The Class IV Chitinase DCAP_0533 Has Two Functional Domains

We have identified a new class IV chitinase from *Drosera capensis*, DCAP_0533. A class IV chitinase has previously been described as one of the most abundant proteins in the pitcher fluid *N. alata* [108], where it preferentially hydrolyzes small GlcNAc oligomers over larger polymeric substrates [126]. Unlike other known plant chitinases, DCAP_0533 contains two class IV catalytic domains. The N-terminal domain appears to be fully active, while the C-terminal domain lacks one of the active residues but containts a full complement of substrate-binding residues(Figure 2.2E, Supplementary Figures A.9-A.10). Multidomain chitinases containing dedicated substrate-binding domains have previously been observed in microbes [276]. For example, ChiA from the thermophilic archeon *Pyrococcus kodakaraensis*, has two chitinase domains and three catalytically inactive substrate binding domains, allowing separate optimizaton of substrate binding and catalytic function [279]. AFM data suggests the binding is mostly determined by interaction of the aromatic residues in the binding site (orange in Figure 2.2E) with the pyranose rings of the substrate [137]. This type of functionality has not been previously observed in plants; we hypothesize that it is an adaptation associated with carnivory, perhaps related to more effective breakdown of small oligosaccharides to components that can be used as a nitrogen source.

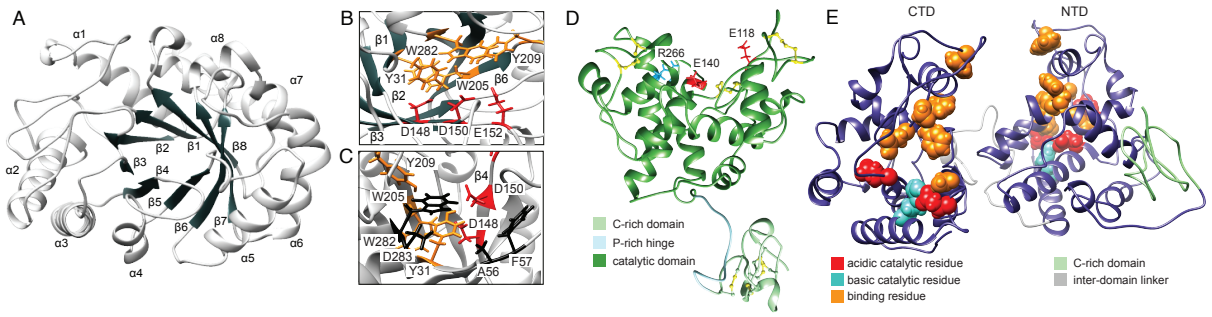


Figure 2.2: Equilibrated structures of the mature sequences of chitinases from carnivorous plants. A. DCAP_0106, a representative Family 18 chitinase, after *in silico* maturation. Numbering of secondary structure elements follows the convention of Si et al. [257]. B. Notably, the tunnel containing the active site has two surfaces with different chemical properties; the aromatic rings (orange) hold the more hydrophobic face of the chitin polymer in place, while the acidic residues (red) perform hydrolysis of the glycosidic linkages. C. Two conserved non-proline cis peptide bonds (black) are critical to shaping the active site tunnel in Family 18 chitinases. D. Chitinase VF-1 from *Dionaea muscipula* V5TEI0_DIOMU [206], with important sequence features and active site residues labeled (red: acidic active residue. blue: basic active residue. yellow: disulfide bond). E. The two-domain chitinase DCAP_0533. Color coding is as in D, with the addition of substrate-binding residues in orange.

2.3.4 Network Analysis Shows Substantial Topological Differences by Family and within Proteins

When selecting potential targets for biophysical characterization, it is useful to consider general patterns of structural similarity or difference within and between families that may correlate with functional differences. Protein structure networks are useful for this purpose, as they directly encode the potential for direct physical interaction between functional groups (rather than representing detailed structure through properties such as side chain dihedral angles that can often vary substantially and dynamically without impacting protein function). Here we employ the PSN representation of Benson and Daggett [21], where vertices represent small moieties and edges represent the potential for direct interaction (as determined by moiety-specific proximity constraints). Given two or more such PSNs, we may compare their

topology by the structural distance method of [37], identifying the smallest number of edge changes (i.e. altered inter-moiety interactions) needed to make one PSN isomorphic to the other. Figure 2.3 depicts respective hierarchical clusterings of the Family 18 (panel A) and Family 19 (panel B) chitinases based on this notion of structural similarity, with distances normalized by the number of vertices to yield a metric with units of average changed interactions per moiety. For Family 18, the pattern of topological similarity is strikingly close to the pattern of sequence similarity, although somewhat more diversity can be seen among structures than among sequences (compare with Figure 3.2). By contrast, topological clustering of Family 19 chitinases shows substantial differences from the sequence-based clustering. For instance, while DCAP_0533, A9ZMK1_NEPAL, and Q6IV09_DRORT belong to an outlying but internally cohesive cluster with respect to sequence similarity, the three show markedly different topologies (and, indeed, are split between the two large structural clusters characterizing the family). More broadly, we find that the Family 19 chitinases divide structurally into two primary clusters (rather than the four obtained from sequence similarity), both of which are internally heterogeneous and neither of which maps cleanly onto the clusters found by sequence similarity. The relationship between sequence and structure is thus much more tightly coupled for Family 18 than Family 19.

Further insight into the structural differences between the two families can be obtained by considering variation in the properties of their respective PSNs. Here, we examine four basic graph-level indices (GLIs) related to protein network organization. *Transitivity* [318] is defined as the fraction of (i, j, k) two-paths for which there exists an (i, k) edge, and is a standard measure of triadic closure; in the PSN context, higher levels of transitivity are associated with structures that are closely and uniformly packed, with few cavities or extended regions. *Degree* is defined as the number of edges incident on a given vertex; for a PSN, this corresponds to the number of other moieties with which a given chemical group is in contact. The standard deviation of the degree distribution within a PSN then provides a measure of the level of heterogeneity in local packing around chemical groups, and we employ

it here as a second GLI. At a somewhat less local level, the (degree) *core number* of a given vertex [250] provides a measure of the extent to which that vertex is embedded in a region of high cohesion within the graph. More precisely, the k-th core (or k-core) of a graph is defined as the maximum set of vertices having at least k neighbors within the set. The core number of a vertex is then the number of the highest-order k-core to which it belongs. Although each k-core is not necessarily cohesive as a whole, cores with $k \geq 2$ are composed of *unions* of cohesive subgraphs, such that all vertices with high core numbers necessarily belong to highly cohesive subgroups. In a PSN context, cohesive subgroups of moieties are joined by multiple, redundant paths and cannot be pulled apart without severing large numbers of edges. At the level of the entire PSN, then, the standard deviation of the core number serves as an indicator of the degree of heterogeneity in structural cohesion, and distinguishes between highly organized structures and structures that combine rigidly and loosely bound regions. Finally, we consider an indicator of the global path structure within the PSN, which we call *M-eccentricity*. The *eccentricity* of a vertex is the maximum geodesic distance from that vertex to any other vertex in the graph [323]; we here refer to the corresponding mean geodesic distance as the M-eccentricity. Vertices with high M-eccentricity are on average peripheral to the graph structure, while those with low M-eccentricity are relatively centrally located. At the level of the PSN as a whole, the standard deviation of the M-eccentricity distinguishes between uniformly globular structures and structures with deformations or other elongations, and we employ it as our fourth GLI.

Panel C of Figure 2.3 shows the distribution of the above GLI values for both chitinase families. All GLIs were calculated using the `sna` library [35]; to facilitate visualization, each GLI was standardized across the combined set of PSNs by subtracting the mean and dividing by the standard deviation prior to analysis. As is clear from Figure 2.3, the two families differ markedly on these four characteristics. On average, the Family 18 structures are substantially more homogeneous with respect to extended structure, local packing, and cohesion, while also being less transitive ($p < 0.001$ for all measures, two-tailed t-test).

With respect to variation within family, the Family 18 structures show significantly less variability in eccentricity heterogeneity and transitivity (permutation test of logged IQR ratios, respective p values $< 1e-5$ and 0.015), but more comparable variability with respect to heterogeneity in local packing and cohesion (respectively $p = 0.073$ and $p = 0.066$, not significant).

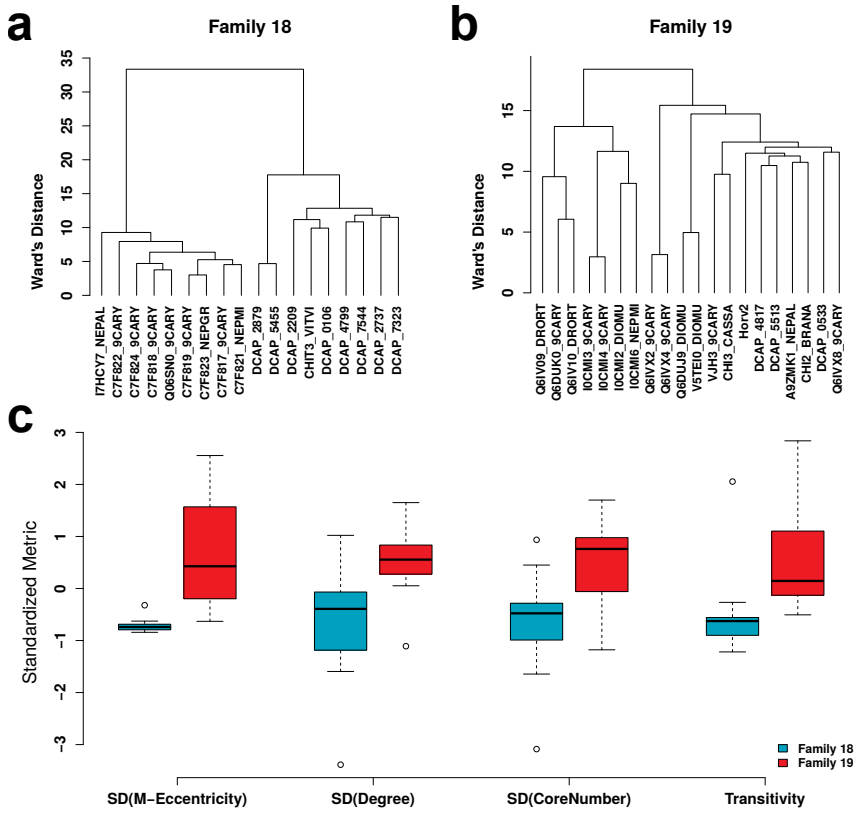


Figure 2.3: (a)-(b) Within-family clustering of chitinases by normalized structural distances. Ward's method (in the generalization of [273]) was employed to construct a hierarchical clustering of Family 18 (a) and Family 19 (b) chitinases based on topological dissimilarity. Sequence similarity is broadly recapitulated by the structural distances in Family 18, while Family 19 shows distinct patterns of variation. Differences between families are large, as illustrated in (c), which shows distributions of M-eccentricity variation, degree variation, core number variation, and transitivity by family. Family 19 chitinases tend to be markedly more internally heterogeneous, with chemical groups whose local structural environments vary far more than their counterparts in Family 18. Family 19 chitinases also show a higher overall level of triadic closure, as captured by transitivity.

To provide an intuition for how these patterns play out in specific cases, Figure 2.4 shows

vertex-level core numbers and M-eccentricity scores for the structures of CF821_NEPMI (Family 18) and DCAP_5513 (Family 19). These structures have low median distance to each other structure in the family, and are hence broadly representative of the classes in question. The core number visualizations of panels (a) and (b) clearly show that CF821_NEPMI is dominated be a large and uniformly cohesive core region, with few vertices in the outer region (i.e., lower cores). By contrast, the highly irregular structure of DCAP_5513 has numerous areas of low cohesion (including much of the C-rich domain) as well as the highly cohesive region associated with the central helices (compare with Figure 2.2). Differences in global structure are brought into sharp relief by the M-eccentricity visualizations of panels (c) and (d). The uniform and tightly connected topology of CF821_NEPMI results in a large number of vertices with short path distances to nearly all other chemical groups in the protein, and relatively little overall variation. Moieties in DCAP_5513, on the other hand, may be at an average distance of more than 9 steps from the rest of the protein, with large differences between the relatively central vertices in the helical region and those in the outer portions of the C-rich domain or the P-rich hinge.

Taken together, these findings suggest substantial structural differences in the basic organization of the Family 18 and Family 19 chitinases, with the former having more internally homogeneous structures, and with structural differences being more closely related to differences in sequence. Family 19 is on the whole more diverse, and contains members that are on average less internally homogeneous. The presence of a higher volume of low-cohesion regions in the Family 19 chitinases suggests that these enzymes may be more prone to thermal denaturation than those in Family 18 (since low-cohesion regions require fewer disrupted edges to pull apart), but may also have functional significance (e.g., by allowing enhanced flexibility). Such structural insights from PSN topology complement those gained by studying specific features, and are more easily extended to analyzing large numbers of sequences.

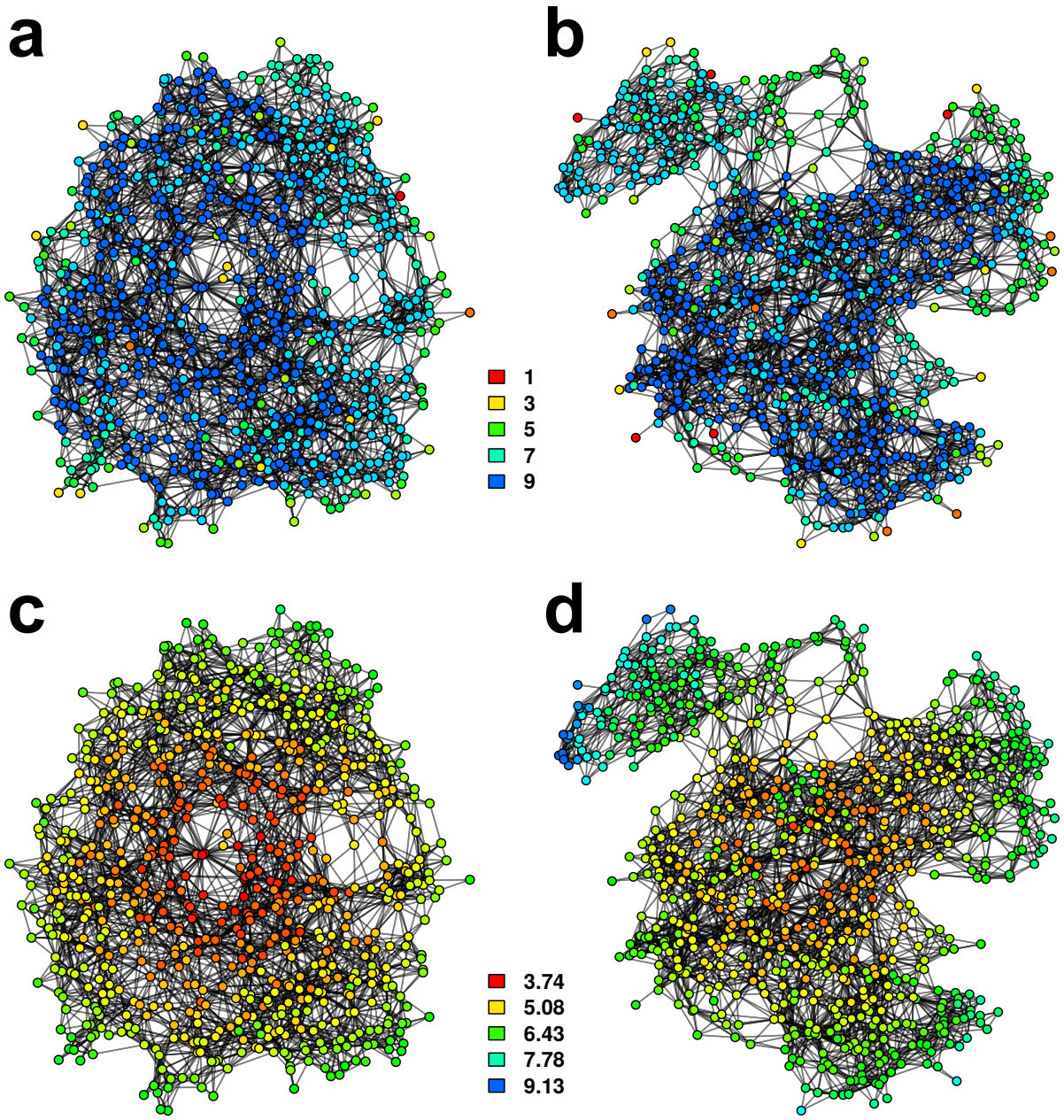


Figure 2.4: PSN Visualizations for family-representative structures C7F821_NEPMI (Family 18, (a) and (c)) and DCAP_5513 (Family 19, (b) and (d)). In panels (a) and (b), vertices are colored by k-core number; vertices with higher core numbers are embedded in more strongly cohesive local structures. Panels (c) and (d) show vertices by M-eccentricity (with higher values indicating a higher mean distance to other vertices in the network). The much higher level of internal heterogeneity in DCAP_5513 versus C7F821_NEPMI is immediately evident, with the former containing complex and irregular structure that subjects some vertices to higher levels of both cohesion and proximity than others.

2.4 Materials and Methods

2.4.1 Sequence Alignment and Prediction of Putative Protein Structures

Network Modeling and Analysis

We mapped each equilibrated protein structure to a protein structure network (PSN) as defined by the representation of [21] using software tools from [36]; these in turn make use of VMD [119] and the `statnet` toolkit [102, 34] within the R statistical computing system [226]. To compare PSNs, we use the structural distance approach of [37], which defines a metric on graph pairs that is in our case equal to the number of edges in one graph that would need to be altered in order to make it isomorphic to the other. (Isolate addition was performed when comparing graphs with differing numbers of vertices.) To remove size effects, the raw distance between each pair of PSNs was normalized by the number of vertices, yielding a metric corresponding to edge changes per vertex. These normalized structural distances were analyzed using hierarchical clustering using R. Additional network analysis and visualization was performed using the `network` and `sna` libraries within `statnet` [34, 35].

2.5 Conclusion

Modeling and analysis of Family 18 and 19 chitinases from *D. capensis* and several related species reveal a number of novel enzymes that present promising targets for subsequent expression and biophysical characterization. These include what is to our knowledge the first plant chitinase found with multiple active domains, as well as several proteins that differ in more conventional ways from others in their class. Comparative network analysis of

these structures reveals within- and between-family differences in structural properties, with Family 18 chitinases tending to be substantially more homogeneous in internal structure and Family 19 chitinases showing variation in cohesion and packing with possible implications for both function and thermal stability. These results also demonstrate the potential of *in silico* pipelines to move rapidly from genomic DNA to predictions of tertiary structure and comparative analysis thereof. As the "genomic revolution" makes such data available at an ever-increasing rate, such pipelines will become critical to our ability to exploit this scientific resource.

2.6 Acknowledgments and Contributions

This work was made possible, in part, through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01; this research was also supported by NSF award DMS-1361425. K.N.E was supported by the UCI UROP program. S.T. and R.W.M. acknowledge the California State Summer School for Math & Science (COSMOS) and NSF grant CHE-1308231.

R.W.M. chose the protein set and oversaw the structural biology aspects of the study. C.T.B. performed the cluster analysis, molecular dynamics simulations, and network visualization and analysis. R.W.M., M.H.U., V.T.D., K.E., S.T., and J.E.K. performed sequence annotation and structural analysis. M.H.U., V.T.D, C.T.B. and R.W.M. wrote the manuscript.

Chapter 3

Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*

3.1 Summary

In plants, esterase/lipases perform transesterification reactions, playing an important role in the synthesis of useful molecules, such those comprising the waxy coatings of leaf surfaces. Plant genomes and transcriptomes have provided a wealth of data about expression patterns and the circumstances under which these enzymes are upregulated, e.g. pathogen defense and response to drought; however, predicting their functional characteristics from genomic

or transcriptome data is challenging due to weak sequence conservation among the diverse members of this group. Although functional sequence blocks mediating enzyme activity have been identified, progress to date has been hampered by the paucity of information on the structural relationships among these regions and how they affect substrate specificity. Here we present methodology for predicting overall protein flexibility and active site flexibility based on molecular modeling and analysis of protein structure networks (PSNs). We define two new types of specialized PSNs: sequence region networks (SRNs) and active site networks (ASNs), which provide parsimonious representations of molecular structure in reference to known features of interest. Our approach, intended as an aid to target selection for poorly characterized enzyme classes, is demonstrated for 26 previously uncharacterized esterase/lipases from the genome of the carnivorous plant *Drosera capensis* and validated using a case/control design. Analysis of the network relationships among functional blocks and among the chemical moieties making up the catalytic triad reveals potentially functionally significant differences that are not apparent from sequence analysis alone.

3.2 Introduction

In land plants, tissues that are exposed to air are protected by the cuticle, a composite biomaterial comprising a cross-linked polyester scaffold interpenetrated by wax components [252]. The cuticle provides a barrier that minimizes water loss and protects the plant from pathogen infection. The relative quantities of hydrophilic and hydrophobic components must be appropriately balanced and spatially located to adhere to the underlying cell walls while presenting a hydrophobic surface to the air interface [55]. Numerous enzymes are involved in producing the polymer components of this material, including esterases, lipases, and GDSL esterase/lipases. Herein we focus on the GDSL esterase/lipases, characterized by the proximity of the active serine residue to the N-terminus, as well as by its surrounding

residues (canonically GDSL) [8]. Esterase/lipases belong to the large α/β hydrolase enzyme superfamily, in which the catalytic triad consists of a nucleophile, an acid, and a stabilizing histidine (in this case Ser-Asp-His). In plants, these enzymes are often localized to the cuticle matrix, where they catalyze the reverse reaction (biosynthesis of polyesters) rather than acting as hydrolases [93]. This biosynthetic activity in the waxy cuticle is consistent with *in vitro* results indicating that esterase/lipases are highly tolerant of hydrophobic environments, where they catalyze the formation of polyesters rather than performing hydrolysis reactions [78].

Esterase/lipases present attractive targets for biotechnology applications because of their potential for producing robust yet ultimately biodegradable polyester materials and hydrophobic surface coatings [174, 141, 300]. Several microbial GDSL proteins have been characterized as relatively promiscuous enzymes that serve a variety of purposes (e.g. protease, lysophospholipase, thioesterase, arylesterase) [160, 184], and accomodating a wide range of substrates [149]. Microbial cutinases, a subclass of serine esterases found in fungi and bacteria, catalyze esterification and transesterfication and can hydrolyze both hydrophobic and lipid substrates in solution or emulsion [254]. In a chemical biology or biotechnology setting, enzymes with different degrees of specificity may be preferred for different applications; for example, promiscuous enzymes are useful for generalized hydrolysis, while those catalyzing a specific reaction are more useful for biosynthetic reactions. Harnessing the potential of these enzymes, given the enormous number of uncharacterized sequences available, requires methodology for predicting their functional characteristics.

Plant GDSL esterase/lipases may provide a rich source of particular chemical functionalities. Many such enzymes have been discovered from genome and transcriptome data [59, 138]; however their specific functions and substrate preferences remain relatively unexplored despite their potential commercial and technological importance. 114 esterase/lipases have been identified from the genome of rice (*Oryza sativa*) alone [57], and a survey of 12 plant

proteomes found that each plant has many esterase/lipase isoforms, including multiple unique genes as well as splice variants [306]. In genomic terms, the large number of GDSL esterase lipases found in plants results from several gene duplication events, followed by selection for novel functions and/or neutral drift [304]. Although in many cases their precise catalytic activities are yet unknown, esterase/lipases are associated with developmental processes [43], pollen exine formation [71], salt tolerance [196], and stress responses [111, 145]. Many of these functions appear to be related to the biosynthesis and metabolism of cutin and waxes [203, 275]. A recent investigation by Zhang et al. demonstrated the first plant GDSL (BS1) to exhibit polysaccharide esterase activity, which is vital for maintaining secondary cell wall acetylation levels and homeostasis [337]. In the oil palm (*Elaeis guineensis*), oil yield correlates with expression of genes for GDSL esterase/lipases and expression of these genes in transgenic *Arabidopsis* plants increases their fatty acid production as well [340].

Much of what is known to date about the specific enzymatic activities of proteins in this family comes from studies of either model systems such as *Arapidopsis thaliana* or crop plants that produce large fruits [14]. For example, in the tomato (*Solanum lycopersicum*), the GDSL1 enzyme is required for cuticle formation; knockdown of expression of the GDSL1 enzyme (also called CD1) using RNAi results in porous fruit cuticles. On the molecular level, both a decrease in the density of cutin monomers and a reduction in ester bond cross-links between the polymer chains were observed [93], consistent with the phenotype of the *cd1* mutant, in which this gene is interrupted by a stop codon. Cutin deficiency caused by the *cd1* mutation reduces the thickness of the cuticle, decreases its mechanical flexibility, and increases its susceptibility to water loss, unlike some other cutin-deficient mutants [124]. GDSL1 (CD1) acts as an acyltransferase, building up the polyester oligomers of the cuticle [332]. This finding highlights the importance of characterizing esterase/lipases in plants; studies in *A. thaliana* have shown that multiple enzymes are required to form a functional cuticle [220], and technological applications will likely also require a series of enzymatic reactions. The esterase/lipases from carnivorous plants have the potential to be

particularly useful from a biotechnology standpoint because of the unique challenges faced by their leaf surfaces, which must withstand the harsh chemical environment associated with their digestive fluids for extended time periods.

Here, we present molecular modeling and functional analyses of 26 esterase/lipases recently discovered from the genome of the Cape sundew (*Drosera capensis*) [36]. The conservation of active site residues, key functional sequence blocks, and overall protein folds suggests that many of the *D. capensis* esterase/lipase sequences form functional enzymes; however the diversity of sequence and structural features indicates a range of potential molecular targets and enzymatic activities. We use sequence analysis, comparative modeling with all-atom refinement followed by *in silico* maturation, and comparison of protein structure networks (PSNs) to identify distinct subgroups of proteins as a first step toward target selection for subsequent expression and biochemical characterization. To enable analysis of structural features with potential functional relevance, we define two novel types of *functionally-targeted protein structure networks* (FT-PSNs) generated using functional information specific to this protein class. In particular, sequence region networks (SRNs) are based on connectivity among previously identified functional sequence blocks, while active site networks (ASNs) are based on interactions among chemical moieties comprising the active site residues. Clustering of SRNs reveals several classes with distinct structural characteristics, providing a parsimonious descriptor of protein structure and a predictor of global flexibility. ASNs are used to construct a measure we hypothesize to correlate with active site flexibility and hence enzyme promiscuity. A case-control comparison with a pair of experimentally characterized esterase-lipases (one promiscuous and one specific) suggests that most of the *D. capensis* esterase/lipases have relatively rigid active sites, consistent with their having specific functionalities. This approach is readily adaptable to other incompletely characterized enzyme classes, providing a potentially useful way of selecting experimental targets based on predicted catalytic specificity.

3.3 Methods

3.3.1 Clustering, Sequence Alignment and Prediction of Putative Protein Structures

D. capensis proteins were annotated using the MAKER-P (v2.31.8) pipeline [40, 41], a BLAST search against SwissProt, and InterProScan [225], as previously described in [36]. The protein set for this study was chosen starting from all sequences identified as having esterase/lipase functionality, followed by elimination of truncated proteins for which one or more of the active site residues were in the missing regions. Clustering of sequences was performed by first aligning sequences using ClustalOmega [258], with settings for gap open penalty = 10.0 and gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, and BLOSUM weight matrix, and then computing a complete link hierarchical clustering of the resulting dissimilarity scores (one minus the ClustalOmega sequence similarity divided by 100, yielding in values on the [0,1] interval). Clustering and other data analyses were performed using the R statistical computing platform [227]. For purposes of subsequent alignment and comparison, subclusters were then made by defining a cutoff point at a sequence dissimilarity value of 0.7. The presence and position of potential signal sequences flagging the protein for extracellular transport were assessed using the program SignalP 4.1 [216], using the following settings: organism group = eukaryotes, D-value cutoff = default (optimized for correlation), and method = input sequences may include transmembrane regions. Structures were predicted from sequences using a three-stage process, following the *in silico maturation* protocol of [38]. First, an initial model was created for each complete sequence using the Robetta implementation of the Rosetta [139, 228] package. These structures were modified in the second stage of the process by removing any residues not present in the mature proteins and by correcting protonation states to reflect their predicted cellular or extracellular environments (with protonation states predicted using PROPKA 3.1 [200]).

In the third phase, each corrected model structure was equilibrated in explicit solvent; simulations were carried out using NAMD [219] with the CHARMM36 forcefield [25] and the TIP3P water model [129] at 293K under periodic boundary conditions. Solvated models were energy-minimized for 10,000 iterations before being simulated for 500ps, with the final configuration being employed in subsequent analyses. This process was performed for the 26 esterase/lipase sequences from *D. capensis* and several reference sequences from other plants. At least one reference sequence was included per subcluster. These proteins were chosen for purposes of sequence annotation: their active sites and functional regions are relatively well annotated in the UniProt database [287], enabling comparisons to the newly characterized sequences. To the best of our knowledge, no structures have yet been solved for a plant esterase/lipase, therefore we also predicted structures for the annotation reference sequences. The PDB files corresponding to the initial and equilibrated structures for all the proteins discussed in this manuscript are available in the Supplementary Information (Supplementary Tables B.1 and B.2).

3.3.2 Network Modeling and Analysis

A protein structure network (PSN) was calculated for each protein from its predicted three-dimensional structure using software tools from [38] (which also make use of VMD [119] and the `statnet` library [102, 34] for `R` [227]). Nodes and edges were defined per [21] (see Figure 3.1A), in which each node represents a chemical group and two nodes are adjacent if they potentially interact (as determined by a distance criterion). Specifically, two nodes i and j are considered adjacent if i contains at least one atom of any type that is within 4.6Å of at least one atom in j, or if i contains at least one carbon that is within 5.4Å of at least one carbon in j. These structures were then secondarily processed to construct functionally targeted PSNs (FT-PSNs) using the `sna` library [35] within `statnet`. A sequence region network (SRN) was constructed from each PSN by identifying all vertices associated

with each conserved sequence block or inter-block region (IBR, region between conserved sequence blocks) and defining two regions to be adjacent in the SRN if and only if there were more than five edges between their respective vertex sets in the corresponding PSN (Figure 3.1B). Each SRN thus encodes the non-trivial interactions among chemical groups within each functionally significant sequence region. Active site networks (ASNs) were also constructed from each PSN as follows. First, all vertices associated with active site residues were identified, as were all vertices adjacent to these vertices within the PSN. The ASN was then defined as the subgraph of the corresponding PSN induced by this combined vertex set. Thus, each ASN represents the local interactions among chemical groups in the active site and the other groups with which they are in contact, irrespective of where these groups reside within the primary sequence.

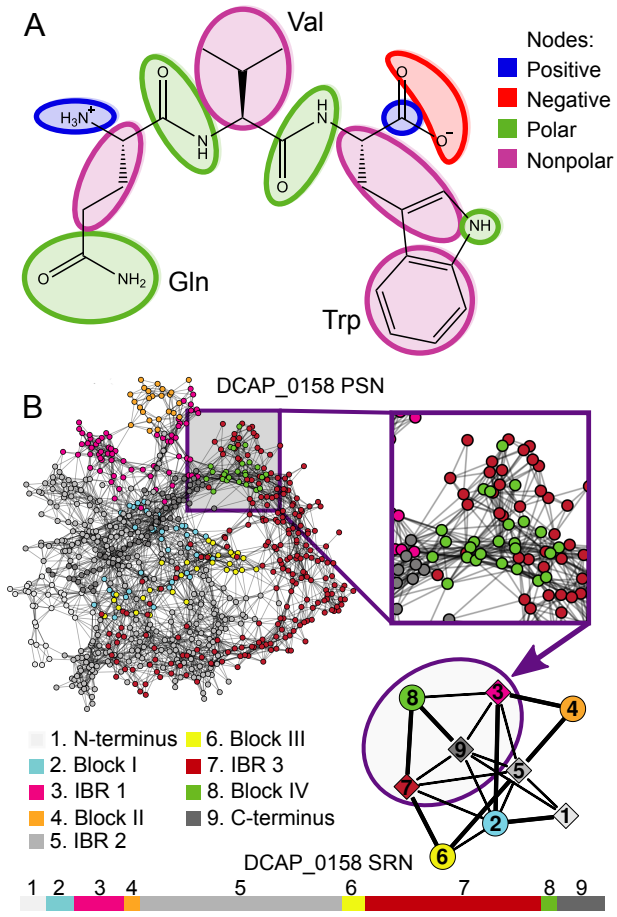


Figure 3.1: A. Node definitions for protein structure networks (PSNs).A polypeptide (here illustrated by the tripeptide QVW) is divided into chemical groups using the Benson-Daggett typology (colored ovals), each group becoming a small-moiety node in the PSN. Nodes are adjacent if at least one atom pair is within a critical radius. B. SRNs are formed from PSNs by first grouping all nodes associated with residues in each sequence region, and then defining region pairs to be adjacent if a threshold number of their respective PSN nodes are adjacent (here, > 5). Schematic shows correspondence between local structure involving the Block IV region and its SRN neighbors (IBR1, IBR3, and the C-terminal region). Shaded bar (bottom) shows relative lengths of each sequence region; although longer regions (e.g., IBR3) are often well-connected, short regions (e.g. IBR1) can also be extremely central.

Clustering of SRNs was performed by calculating the Hamming distance between SRNs (i.e., the number of edge changes needed to convert one SRN into another) and computing a complete link hierarchical clustering solution for the resulting distance matrix (all analyses performed using `statnet` and R). Inspection of the dendrogram (Figure 3.5A) indicated a

four-cluster solution, and central graphs were calculated from the networks in each respective cluster. Block image matrices showing the fraction of SRNs having each respective inter-region edge are shown in Figure 3.6.

Constraint of active site residues within ASNs was assessed as follows. For each vertex associated with a moiety in the active site, three measures were computed: the *degree,* or number of ties to other vertices; the *triangle degree,* or number of triangles (3-cliques) to which the vertex belongs; and *core number,* or number of the highest degree k-core [319] to which the vertex belongs. Physically, these respectively indicate the total number of contacts associated with the chemical group (potentially impeding its motion), the number of truss-like, triangular structures in which the group is embedded (again, restricting mobility), and the extent of local cohesion around the chemical group (found to distinguish "tighter" and "looser" packing regimes [291]). To summarize the impact of each measure over the active site as a whole, values were averaged across active-site vertices. To obtain an additional constraint measure, the number of paths between each pair of active-site vertices through neighboring (i.e., non-active site) vertices was computed, and the log of the minimum of this value over the set of active site vertex pairs was employed as a measure of *site cohesion.* Intuitively, high values of site cohesion indicate that all active site chemical groups are connected by a large number of indirect contacts, while low values suggest that at least one pair of active site moieties has few local pathways holding them together. These four indices (mean active site degree, mean active site triangle degree, mean active site core number, and site cohesion) were used to produce an omnibus index of *site constraint* via principal component analysis (PCA) of the standardized network measures. The PCA solution revealed one primary dimension, with the first principal component accounting for 75% of the total variance among the four measures (ratio of first eigenvalue to second greater than 5), and the scores on this first component scores were hence employed for subsequent analysis as the constraint index.

3.4 Results and Discussion

3.4.1 *D. capensis* Esterase/Lipases Cluster Into Distinct Subfamilies Based on Sequence Features

All enzymes from the *D. capensis* genome previously annotated as functional esterase/lipases were clustered by sequence similarity (Figure 3.2). Several annotation reference sequences from other plants were also included to facilitate identification of the active site residues and functional sequence blocks. The reference sequences (referred to by their UniProt IDs) are from the plants *Carica papaya* (GDL1_CARPA) and *Arabidopsis thaliana* (GLIP6_ARATH, GDL7_ARATH, EXL3_ARATH, APG2_ARATH). Although the active site residues and functional sequence blocks are readily found, plant esterase/lipases are relatively poorly characterized; these reference sequences lack high-resolution structures and in most cases detailed functional information, e.g. experimental data about their substrate preferences. One of the objectives of this work is to provide a starting point for approaching such studies in undercharacterized enzyme classes such as this one.

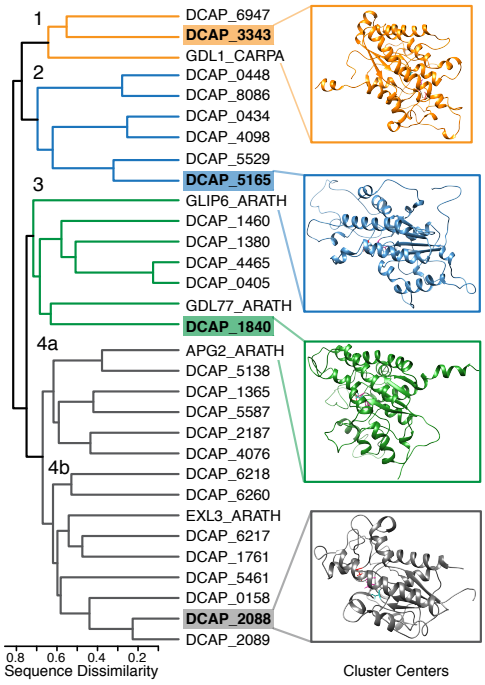


Figure 3.2: Protein sequence clustering of esterase/lipase sequences from the *D. capensis* genome, denoted by DCAP, and annotation reference sequences from other plants, which are identified by their UniProt IDs: *Carica papaya* (GDL1_CARPA) and *Arabidopsis thaliana* (GLIP6_ARATH, GDL7_ARATH, EXL3_ARATH, APG2_ARATH). Information about these annotation reference sequences found in UniProt enabled identification of functional sequence features in the novel *D. capensis* proteins via sequence alignment and comparison. Annotation details are shown in Supplementary Figures B.1-B.5.

In all the sequences examined here, the active site residues are consistent with the catalytic triad of a serine hydrolase, and the functional sequence blocks characterizing the GDSL esterase/lipase family are readily identified by comparison to the work of Akoh et al. [8] and Vujaklija et al. [306]. In most cases, SignalP 4.1 predicts the presence of a signal peptide sequence tagging these esterase/lipases for extracellular secretion. Annotated protein sequence alignments showing functional sequence features can be found in Supplementary Figures B.1-B.5. The sequence alignments are color-coded to indicate both individual amino acid properties and important sequence regions. Sequence-based clustering yields four major groups with greater than 30% sequence identity among all members. As previously observed for this protein class, each group has significant diversity among its component sequences;

only one pair in this set (DCAP_0405 and DCAP_4465) has more than 80% sequence identity. For each cluster, the central sequence (the protein having the minimum average distance in sequence space from all the others) is highlighted. Comparative models for these central sequences are shown to the right of the cluster figure, revealing variations on a common structural theme.

Cluster 1 contains sequences that have the canonical GDSL motif, as found in the reference sequence GDL1_CARPA, which was isolated from papaya latex [3] and has been proposed as a "naturally immobilized" biocatalyst for performing regioselective esterification and transesterification reactions [80]. The enzymes in cluster 2 instead have GDSN in the first functional block. Clusters 3 and 4 contain the motif GDSX, where X is usually a hydrophobic residue, but is Ser or Thr in some cases. Overall, the presence of the three active site residues in 24 of the 25 *D. capensis* esterase/lipases suggests they are functionally active enzymes.

3.4.2 Conserved Active Site Residues Suggest Functional Enzymes

In general, esterase/lipases are characterized by four moderately conserved sequence blocks of length 8-13 residues that contain the cataytic triad, the oxyanion hole proton donors, and other functionally relevant residues [292]. These blocks are always found in the same order in sequence space, though the lengths of the intervening sequences can vary substantially [71]. Functional sequence blocks I-IV are highlighted in the sequence alignments (Supplementary Figures B.1-B.5.) In Figure 3.3A, these functional blocks are represented as sequence logos, where the size of each residue label correlates with the number of instances at that sequence position within each cluster. The Ser-Asp-His catalytic triad is located within two block regions: block I (Ser) and block IV (Asp-His). The remaining two blocks contain conserved oxyanion hole residues, Gly in block II and Asn in block III [8]. Most of the proteins in this set contain the expected functional residues, as exemplified by the reference sequences

GDL1_CARPA, GLIP6_ARATH, and GDL7_ARATH, as well as the functionally characterized GDSL esterase/lipase G1DEX3_SOLLC from the tomato.

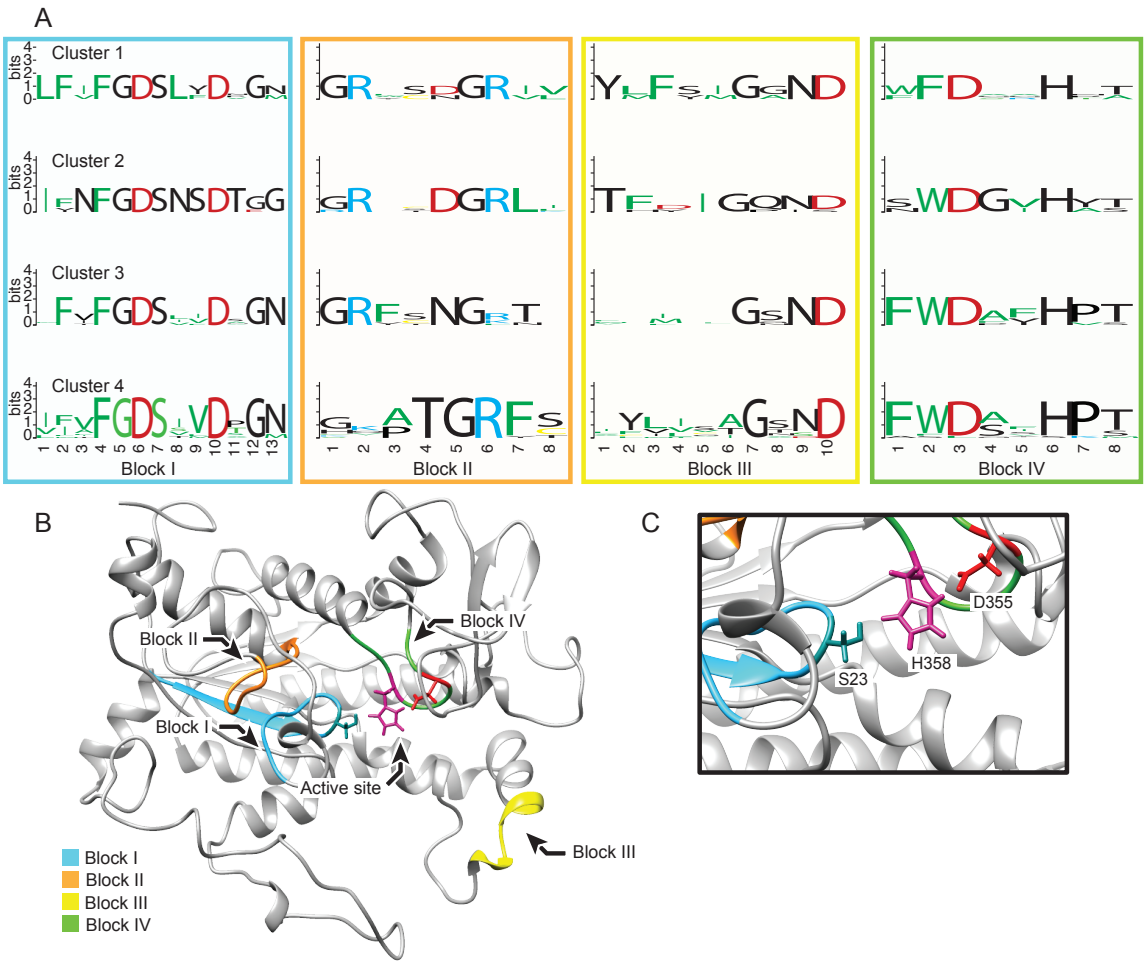


Figure 3.3: A. The sequences of the four functional blocks (inside the colored frames) are presented by sequence cluster (arranged from top to bottom as in Figure 3.2). The sizes of the residue labels correlate with the fraction of sequences in the cluster having that residue in the indicated position. Amino acid properties are color coded as follows: hydrophobic-green, positive-blue, negative-red, cysteine-yellow, other-black. B. A representative molecular model of a *D. capensis* esterase/lipase (DCAP_0434) with the four functional blocks highlighted using the color-coding of the frames in Panel A. C. Expanded view of the active site catalytic triad for a typical esterase/lipase (DCAP_0434), showing that the active site residues are positioned in a manner consistent with catalytic activity.

Some variation is observed in the oxyanion hole residues: the stabilizing Asn residue in block III is replaced by Ile in DCAP_0434, Ser in APG2_ARATH and DCAP_ 5138, and Asp

in EXL3_ARATH. These substitutions are consistent with almost all of the *D. capensis* enzymes following the canonical GDSL mechanism [231]. The two exceptions in this set are DCAP_2088, which is missing the entirety of block III, and DCAP_6260, which has substitutions to the two active site residues located in block IV (Asp to Leu and His to Ser, see Figure B.4). DCAP_6260 is the only protein in this set that does not contain all three active site residues, although it retains the canonical GDSX motif in block I and the stabilizing oxyanion residues in block II and III. The potentially catalytically inactive sequences (DCAP_6260 and DCAP_2088) were included because they do contain most of the relevant sequence and structural features; we hypothesize that these proteins may play a binding rather than catalytic role. Alternatively, they may represent pseudogenes. DCAP_4076 has a C-terminal extension not found in the other esterase/lipases, the role of which is currently not known, although it has moderate sequence similarity to transcriptional regulation proteins in *Arabidopsis thaliana* and soybean (Supplementary Figure B.8A.).

3.4.3 Molecular Modeling

The structure of a typical GDSL esterase/lipase has a 4-stranded parallel β-sheet with six α-helices arranged around it (shown for a representative example in Figure 3.3B). Due to the lack of solved structures for plant esterase/lipases, comparative modeling was used rather than traditional homology modeling. To make a standard homology model, the sequence of interest is threaded onto the known structure of a closely related protein, followed by energy minimization. In comparative modeling, the procedure is similar except that the protein is modeled piecewise using multiple template structures selected by the software (in this case Rosetta [139]) from the Protein Data Bank, followed by global minimization using a simplified force field. This methodology is regularly validated via CAMEO [99], and is the basis of well-known structure prediction systems such as Rosetta (used here) and I-TASSER [339]. All template structures used for a representative example (DCAP_0434) are tabulated

in Supplementary Table B.3 and the parent structures for each model can be found in the headers for their respective .pdb files (available for download in the SI.)

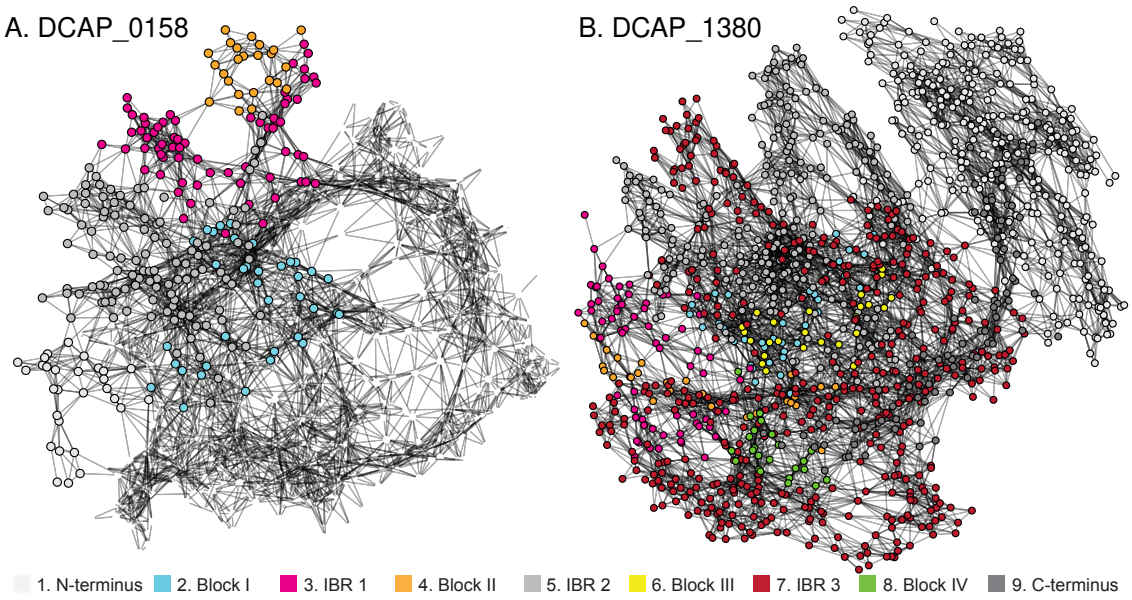


Figure 3.4: Protein structure networks of DCAP_0158 (Cluster 4a) and DCAP_1380 (Cluster 3). Each node (closed circles) represents a chemical moiety and is color coded based on its respective sequence position in a functional block, terminus or IBR. Ties (gray lines) indicate physical interactions between a set of nodes. The positioning of the nodes in this representation is optimized to show topology and does not directly correspond to three-dimensional space; proximity within the cutoff distance is solely indicated by the ties.

We used the initial models generated by the Robetta server [228] as a starting point; however as these structures are not calculated in an aqueous environment and do not account for protonation states, we modifed them to produce models that are more representative of the mature enzyme (available for download in the SI.) Signal sequences were removed and protonation states were corrected consistent with their expected functional environments. These structures were then subjected to molecular dynamics simulation in explicit solvent to generate the equilibrated structures (illustrated in Supplementary Figure B.6). The equilibrated molecular models of these proteins show that although they all have the expected overall fold, substantial diversity exists in the placement of secondary structure elements, as well as the lengths of the linker regions (Supplementary Figures B.7 and B.8B). All three

active site residues are accessible, in contrast to lipases, where only the serine is exposed due to the hydrophobic “lid" that is characteristic of that enzyme class. The positioning of the catalytic triad residues, which is consistent with catalytic competence, is shown in Figure 3.3C. The active site residues are located in loop regions, with the occasional exception of the Ser, which is part of an α-helix in some esterase/lipases (e.g. in Cluster 1). The conserved oxyanion hole residues in Block II reside in a loop region, while half of the Block III residues lie in a β-sheet and the other half in an α-helix. This mixture of structural motifs presents a challenge for coarse-grained network analysis, where a common approach is to break up the protein into discrete regions based on secondary structure. In the case of the plant esterase/lipases of this set, that classification does not align with the functional regions identified in previous studies of esterase/lipases; we have therefore used the functional sequence blocks, termini, and inter-block regions rather than secondary structure elements as the basis for constructing the FT-PSN representation of the overall enzyme folds.

3.4.4 Protein Structure Networks

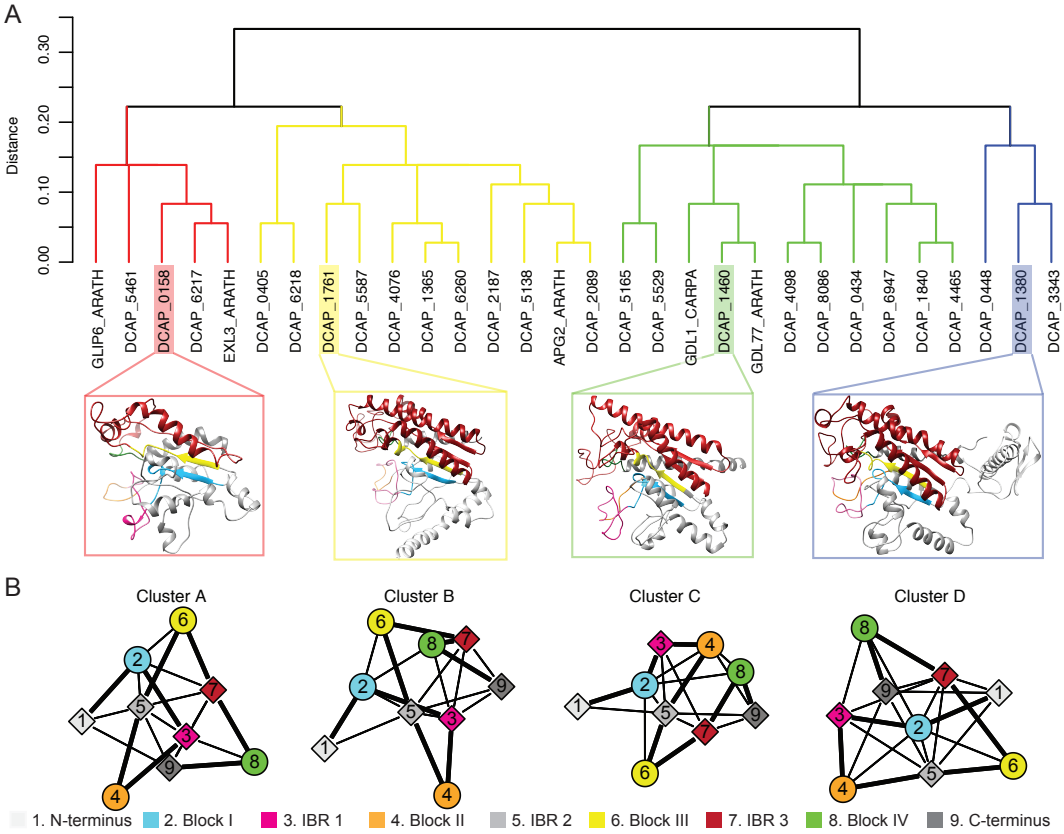


Figure 3.5: A. Clustering of sequence region networks (SRNs) for modeled esterase/lipase structures from the *D. capensis* genome and reference sequences from other plants. Inset structures depict the most central member of each cluster. B. Central graphs for the SRNs in each cluster. Colors for nodes corresponding to conserved (circular) and non-conserved (diamond-shaped) sequence regions correspond to residue colors in panel A; thick lines indicate connections along the protein backbone.

Contacts between structural regions of the esterase/lipases were analyzed using a network formalism; for each protein, full PSNs and two novel types of FT-PSNs were generated. First, full PSNs were calculated for the esterase/lipase molecular models based on the formalism of Benson and Daggett [21], where each amino acid is composed of nodes defined by chemical functionality. Two illustrative visualizations of PSNs from different sequence clusters are shown in Figure 3.4. Although we refer to the functional blocks themselves by Roman numerals I-IV as defined in the earlier literature for the sake of comparison to prior work,

for purposes of generating FT-PSNs we define nine sequence regions comprising the four functional blocks as well as the regions between them (inter-block regions, or IBRs), and the N- and C-termini. These sequence regions are numbered 1-9 in order from the N-terminus to the C-terminus for each protein. In these PSN examples, nodes (chemical moieties) belonging to the termini, functional blocks, and inter-block regions are color coded as indicated in the legend. This representation allows rapid examination of the degree of connectivity between different sequence regions, e.g. it can easily be seen that the nodes of Block II (orange) are more connected to each other in DCAP_0158 (3.4A) than in DCAP_1380 (3.4B), while many Block III nodes are connected to those from other sequence regions in both proteins. Although this representation provides a visualization of connectivity between different parts of the protein separate from the three-dimensional structure, the number of nodes and the complexity of the plots makes comparison difficult. Therefore, we define two types of specialized FT-PSNs based on functionally relevant sequence features of these proteins.

In order to further simplify the graph representations, a block model [319] was constructed for each protein by condensing all nodes within each of these sequence regions to form a coarse-grained FT-PSN whose edges represent contacts between moieties in each pair of sequence regions (each region constituting a node within the block model). These *sequence region networks* (SRNs), provide a direct representation for the structure of contacts among functionally significant components of the protein, which we hypothesize to be related to overall function. To identify distinct classes of functionally relevant structure within the *D. capensis* esterase/lipase set, we then performed a hierarchical clustering of SRNs by Hamming distance (i.e. the number of adjacency differences among sequence regions between two respective SRNs). Figure 3.5A shows the dendrogram for the clustered SRNs, along with structural models for the protein structure corresponding to the central graph for each SRN cluster. The central graphs themselves are shown in Figure 3.5B. Following clustering of SRNs by Hamming distance, clusters were summarized by forming block image matrices [319]. Within each matrix, the i, j cell value corresponds to the fraction of cluster members

whose SRN contains an edge between sequence region i and sequence region j. Schematic representations for each cluster, illustrating how the adjacency matrices for these models are constructed, are shown in Figure 3.6. In addition to showing distinct structural patterns across clusters, Figure 3.6 shows a fairly high level of consensus within clusters (with most cells having densities close to either 0 or 1). For this reason, we summarize the SRNs within each cluster by their central graph, which is equivalent to dichotomizing the image matrices at 0.5; these networks are shown in Figure 3.5B.

Clustering of the SRNs reveals important differences among esterase/lipases that are not apparent from the sequence clusters, as well as some common features of potential structural and functional significance. For example, the IBR between Blocks II and III (node 5) is highly central across all structures, being in direct contact with a large number of other sequence regions and frequently bridging regions not otherwise in contact. This suggests a key structural role for this highly variable (i.e. non-conserved) sequence region that may have been overlooked by purely sequence-based analyses. Likewise, Block III has identical neighbors in all clusters, being tied only to its sequence-space neighbors and to Block I (node 2). This highly conserved pattern of both interaction and *non*-interaction is suggestive of functional significance. By contrast, the other interaction partners of Block I vary considerably across clusters, as do e.g. the partners of IBR 1 (node 3). Such variation in interaction among conserved sequence blocks may be indicative of corresponding differences in functional characteristics.

Interestingly, clustering by structural similarity of SRNs yields a pattern that is distinct from clustering by sequence (Figure 3.2). Although sequence homology is often a good indicator of broad functional similarity at the level of protein classes, structural comparison provides a much more precise tool for functional differentiation among related proteins. As with previous applications of structure networks to study allostery, binding, inter/intramolecular interactions, and other phenomena otherwise difficult to ascertain using only sequence anal-

ysis [253, 21, 291], SRNs such as those introduced here have the potential to complement sequence analytic methods for purposes such as functional prediction and target selection.

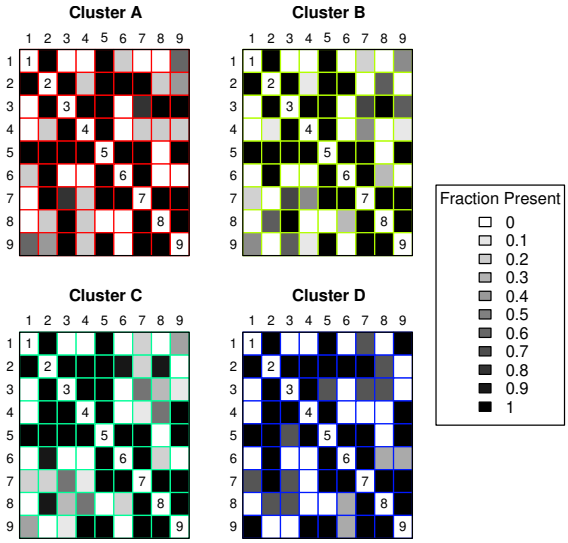


Figure 3.6: Block image matrices for the clustered sequence region networks. The i, j cell value for each matrix indicates the fraction of cluster members whose SRN contains a tie from region i to region j. Node numbers correspond to sequence regions numbered from the N-to the C-terminus as defined in the text.

The coarse-grained network representations described above provide a useful basis for comparison of overall structural properties among esterase/lipases, but they do not directly address the flexibility and accessibility of the active site itself, which is a potential indicator of enzyme specificity [12]. Most of what is known about the esterase/lipase family to date comes from the microbial esterase/lipases, which are generally regarded as promiscuous enzymes. It has been suggested that this property may generalize to plant esterase/lipases, which have so far not been extensively characterized. However, as discussed above, many plants have numerous esterase/lipase paralogs, possibly indicating that the same diversity of activity is accomplished using multiple enzymes, each with its own functionality, rather than fewer multifunctional enzymes.

Because enzyme promiscuity is strongly correlated with active site flexibility [12], we used a similar analysis of network structure to investigate the ties among nodes in the active site

regions of the *D. capensis* esterase/lipases. As before, we began by constructing moiety-level PSNs using the Benson-Daggett representation. We then formed *active site networks* (ASNs) by taking the subgraph of each PSN induced by the nodes corresponding to active site moieties together with the union of their respective network neighborhoods. Each ASN thus represents the pattern of connectivity among moieties topologically local to the active site. Structural constraints on the active site were measured using several common network properties: mean degree (the average number of ties each node has to other nodes), mean triangle degree (the number of memberships in 3-cliques or triangles), mean k-core number (where the kth core of a graph is the maximum set of nodes such that every member of the set is adjacent to at least k other nodes), and inter-node connectivity (counts of paths connecting active-site nodes via other nodes in the ASN). These properties were computed for all nodes corresponding to active site moieties, and are plotted in Supplementary Figure B.9. They were then composited by taking their first principal component, yielding a single measure of active site constraint for each network.

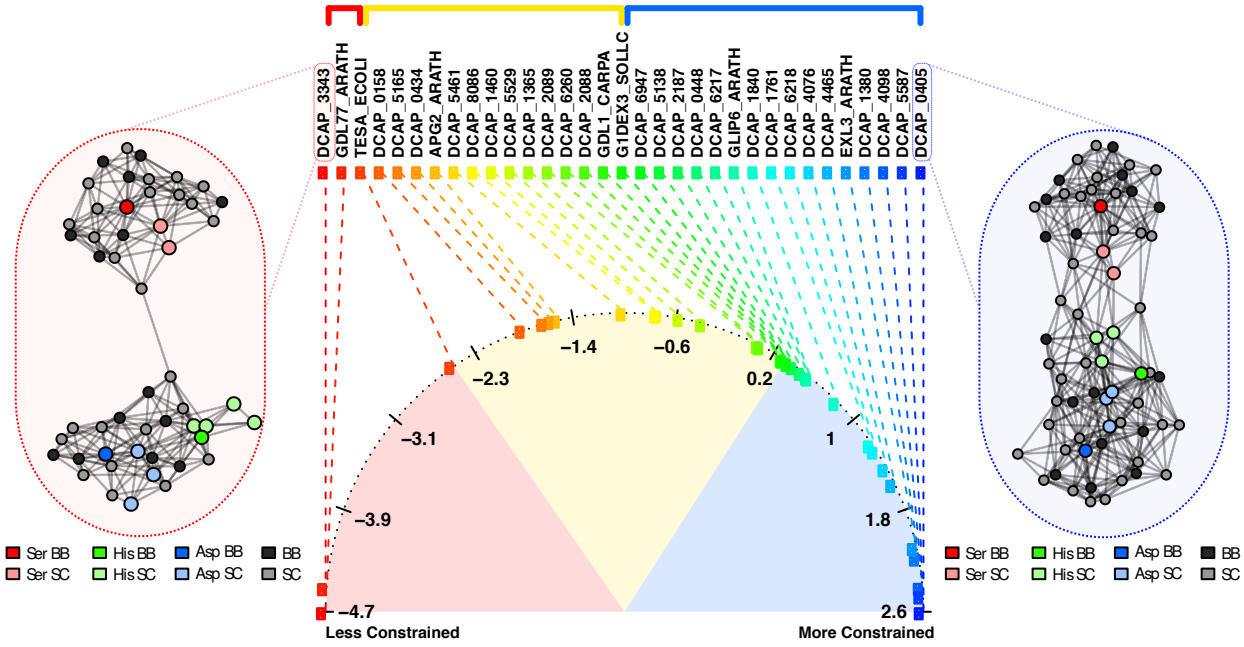


Figure 3.7: Main panel: constraint level of active site moieties within protein structure networks. Red-shaded region indicates lower constraint levels than the bacterial enzyme TesA; yellow and blue shaded regions respectively indicate levels of constraint between TesA and tomato cutinase and levels of constraint greater than tomato cutinase. Nearly all plant enzymes studied here show more active site constraint than TesA, with tomato cutinase falling near the median of these. Side panels: ASN visualizations for DCAP_3343 (left) and DCAP_0405 (right) show respective examples of low and high levels of active site constraint. Nodes correspond to moieties, with backbone (BB) and side chain (SC) moieties for the three active site residues indicated by color. Highly cohesive ASNs imply numerous constraints on the motion of active site residues, potentially leading to higher levels of substrate specificity.

Figure 3.7 shows the active site constraint measure for each enzyme in our set, as well as two enyzmes for which more detailed activity data is available. The latter two, well-characterized enzymes were selected as a "case/control" validation for the functional significance of the constraint measure: the tomato cutinase (G1DEX3_SOLLC), which is known to catalyze a specific reaction (high-specificity "case"); and *E. coli* TesA, (TESA_ECOLI), which is known to accept a variety of substrates (low-specificity "control"). Consistent with the hypothesis that the large number of esterase/lipases in typical plant genomes corresponds with a higher level of substrate specificity, we observe only two plant enzymes with a level of constraint lower than the promiscuous TesA (red-shaded area); of the remainder, roughly half showed

constraint levels between TesA and tomato cutinase (yellow-shaded area) and half showed higher constraint levels (blue-shaded area). Our analysis suggests that the majority of esterase/lipases in *D. capensis* are likely to be highly specific, with the prominent exception of DCAP_3343. This enzyme, and GDL77_ARATH from *Arabidopsis*, show extremely low levels of active site constraint implying a very high level of local flexibility. We hypothesize that these enzymes will accept a wider range of substrates than the others examined here, and that they occupy a distinct functional role (perhaps more similar to the role of microbial esterase/lipases).

Figure 3.8 shows structural models of the *D. capensis* esterase/lipases with the least (red) and most (blue) constrained active sites, as determined by the ASN flexibility metric plotted in Figure 3.7. Somewhat counterintuitively, the protein with the less flexible active site (DCAP_3343) has a better-defined secondary structure. Based on the DSSP secondary structure definitions [131], DCAP_0405 has 29.3 % α-helix, 2.9 % β-strand, and 67.8% turn/coil, while (DCAP_3343) has 43.6% α-helix, 5.3% β-strand, and 51.1% turn/coil. Although DCAP_3343 has more α-helical and β-strand secondary structure elements, the structure around the active site itself is looser and less densely connected than that of DCAP_0405, where loops and random coil regions interact to hold the active site residues more rigidly in place. Although unstructured regions are often regarded as highly flexible regions, this depends on their context in the overall structure; recent NMR dynamics measurements and MD simulations reveal that loops undergo dynamics over a wide range of timescales [204] and their motions are frequently involved in allosteric regulation [109]. Longer loops, which are more able to become mutually entangled with other structural elements are more likely to be rigid [98], which is consistent with the predicted structure of DCAP_0405.

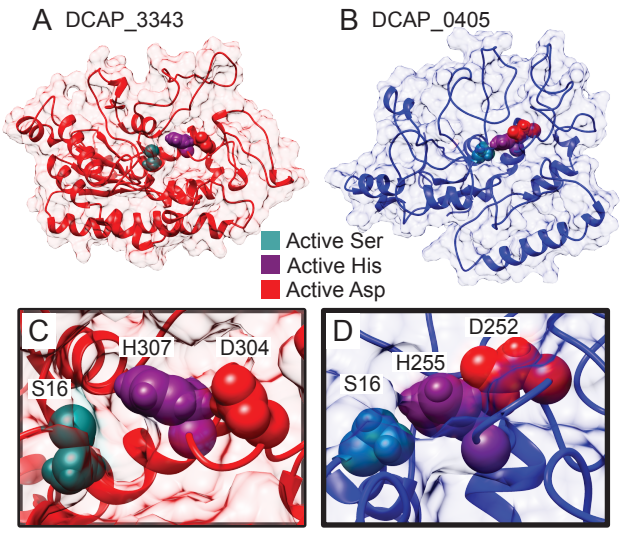


Figure 3.8: Structural models of the least and most constrained enzymes based on the ASN analysis shown in Fig 3.7. A. Surface and ribbon representations of DCAP_3343, which is the only *D. capensis* esterase/lipase with a less constrained active site than that of TesA from *E. coli*. B. Surface and ribbons for DCAP_0405, the most constrained enzyme in this set. C. and D. Expanded views of the active ites of these enzymes show the differences in active site constraint, which are not obvious from examination of the overall structural model. The active site residue side-chains of DCAP_3343 (C) are oriented out and away from each other, while those of DCAP_0405 are tightly held in a closely packed conformation.

3.5 Conclusion

In summary, molecular modeling and protein structure network analysis of 26 esterase/lipases identified from the genomic DNA of *Drosera capensis* suggest that—with the exception of one protein, DCAP_3343—the active site regions of these enzymes are less flexible than those of related microbial proteins. We hypothesize that these enzymes act (like tomato cutinase) to catalyze specific reactions, with the outlying protein behaving more like microbial esterase/lipases. Two new types of protein structure networks, seqence region networks (SRNs) and active site networks (ASNs) were defined in order to characterize overall protein flexibility and that of the active sites. Principal component analysis of active site constraint measures generated from PSNs enabled us to sort the esterase/lipases from decreasing to in-

creasing active site rigidity; case/control validation using a pair of well-characterized enzymes suggests that our index is related to substrate specificity. Clustering by SRN shows structural differences between enzymes with respect to functionally significant sequence blocks, as well as an apparently conserved structural role for a highly sequence-variable and previously unnoted inter-block region. These results may serve to guide target selection for subsequent structural or functional studies, and the analytical strategy employed may be fruitfully adapted to other protein classes.

Chapter 4

Elucidation of WW domain ligand binding specificities in the Hippo pathway reveals STXBP4 as YAP inhibitor

4.1 Summary

The Hippo pathway, which plays a critical role in organ size control and cancer, features numerous WW domain-based protein-protein interactions. However, ∼100 WW domains and 2,000 PY motif-containing peptide ligands are found in the human proteome, raising a "WW-PY" binding specificity issue in the Hippo pathway. In this study, we have established

the WW domain binding specificity for Hippo pathway components and uncovered a unique amino acid sequence required for it. By using this criterion, we have identified a WW domain-containing protein, STXBP4, as a negative regulator of YAP. Mechanistically, STXBP4 assembles a protein complex comprising α-catenin and a group of Hippo PY motif-containing components/regulators to inhibit YAP, a process that is regulated by actin cytoskeleton tension. Interestingly, STXBP4 is a potential tumor suppressor for human kidney cancer, whose downregulation is correlated with YAP activation in clear cell renal cell carcinoma. Taken together, our study not only elucidates the WW domain binding specificity for the Hippo pathway, but also reveals STXBP4 as a player in actin cytoskeleton tension-mediated Hippo pathway regulation.

4.2 Introduction

Signaling proteins often entail modular domains that facilitate protein-protein interactions to assemble functional protein complexes, control enzymatic activity and regulate protein cellular localization [60, 209]. Importantly, the recognition between domains and their peptide ligands is usually specific, thus allowing the transduction of unique information through signaling cascades [66, 114]. The WW domain is a small protein module that is defined by the presence of two tryptophan (W) residues separated apart by $\sim$25 amino acids [271]. WW domain and its cognate proline-rich peptide motif have been identified within various protein complexes widely distributed in plasma membrane, cytoplasm and nucleus. Failure of their recognition is associated with multiple human diseases including Alzheimer's disease [155, 177], Huntington's disease [85, 205], Liddle Syndrome [103], Golabi-Ito-Hall Syndrome [165, 282], muscular dystrophy [28, 83, 233] and cancers [53, 240]. These facts highlight a crucial role of the WW domain-mediated protein-protein interaction in biological processes and tissue homeostasis.

WW domain was initially uncovered by characterizing the protein sequence of YAP, a key transcriptional co-activator downstream of the Hippo pathway [127, 202, 270]. The Hippo pathway is a highly conserved signaling pathway involved in tissue homeostasis, organ size control and cancer development [101, 127, 202, 88, 243]. In mammals, the Hippo pathway is composed of a kinase cascade (two serine/threonine kinases, MST and LATS; and the adaptors SAV1 for MST and MOB1 for LATS), downstream effectors (YAP and TAZ), and nuclear transcriptional factors (TEADs). MST phosphorylates and activates LATS, which in turn phosphorylates YAP and TAZ. The phosphorylated YAP/TAZ can be recognized by 14-3-3 proteins, retained in the cytoplasm and eventually targeted by β-TRCP E3 ligase complex for degradation. When the Hippo pathway is inactivated, unphosphorylated YAP/TAZ enter into the nucleus, where they associate with TEAD transcriptional factors to promote the transcription of genes that are involved in proliferation and survival.

Notably, many Hippo pathway components and regulators contain either the WW domain or its proline-rich peptide ligand, mostly "PPxY" motif (P, proline; Y, tyrosine; x, any amino acid; hereafter named as "PY" motif) [240, 269]. YAP, TAZ, SAV1 and KIBRA, an upstream component of the Hippo kinase cascade [335], are four known WW domain-containing components of the Hippo pathway [240]. In the nucleus, the WW domain of YAP/TAZ is a requirement for their association with a group of nuclear transcriptional factors and regulators that contain the PY motif to regulate gene transcription [46, 87, 107, 156, 224, 267, 268, 45]. In the cytoplasm, the PY motif of LATS1/2 is involved in the LATS1/2-mediated YAP/TAZ phosphorylation [104, 299]; several PY motif-containing proteins can physically bind the WW domain of YAP/TAZ and promote YAP/TAZ's cytoplasmic translocation [51, 84, 159, 190, 284, 313, 314, 315, 243]. Moreover, the phosphorylated YAP/TAZ can negatively regulate Wnt pathway by forming a complex with DVL2, which is mediated by the WW domain of YAP/TAZ and the PY motif of DVL2 [296]. As a Hippo upstream component, KIBRA can similarly associate with several Hippo PY motif-containing proteins and negatively regulate YAP [284, 326]. On the other

hand, several WW domain-containing proteins have been shown to modulate the Hippo pathway activity by regulating the Hippo PY motif-containing components and regulators [6, 241, 242, 290, 308, 333]. Collectively, these facts suggest that the WW domain and PY motif-mediated protein-protein interaction plays a fundamental role in building up the major framework of the Hippo pathway.

Actually, ∼100 WW domains and 2,000 PY motif-containing peptides have been predicted in the human proteome [282], raising an issue of binding specificity for the proteins containing WW domain and PY motif. Indeed, a large scale of WW domain array screen only confirmed 10% of the tested WW domain-ligand interactions [114]. Several large-scale proteomic studies exclusively identified a group of PY motif-containing proteins (e.g., LATS1/2, AMOTs, PTPN14) as the binding partners for the Hippo WW domain-containing components [63, 15, 315]. These facts indicate the binding specificity for the Hippo WW domain-mediated protein-protein interaction, while the underlying mechanism is still largely unknown.

In this study, we demonstrated the WW domain binding specificity for the Hippo pathway proteins and uncovered a highly conserved amino acid sequence required for it. By using this criterion, we identified STXBP4 as a novel Hippo pathway regulator in human proteome. Mechanistically, STXBP4 assembled a complex with α-catenin and several Hippo PY motif-containing components/regulators to negatively regulate YAP when actin cytoskeleton tension is low. Moreover, both TCGA data and tissue array studies suggested STXBP4 as a potential tumor suppressor in human kidney cancer, whose downregulation is significantly correlated with YAP activation in clear cell renal cell carcinoma. Collectively, our study not only elucidated the WW domain binding specificity for the Hippo pathway protein-protein interaction network, but also identified STXBP4 as a Hippo pathway regulator and a potential tumor suppressor in kidney cancer development.

4.3 Results

4.3.1 Binding specificity exists for the Hippo WW domain-containing components

We re-analyzed our previously published proteomic data [315] for four Hippo WW domain-containing components YAP, TAZ, SAV1 and KIBRA (Figure 4.1A), and found that most of the known Hippo PY motif-containing proteins (e.g., AMOT, AMOTL1, AMOTL2, LATS1, LATS2, PTPN14, PTPN21, WBP2) were hardly detected in the SAV1-associated protein complex (Figure 4.1B). Moreover, proteomic analysis of the WW domains isolated from these four Hippo components (Figure C.1A) further confirmed this finding, where the WW domain of YAP, TAZ and KIBRA, but not that of SAV1, retrieved most of these known Hippo PY motif-containing proteins (Figure 4.1B). These data suggest that the WW domain of SAV1 is different from that of YAP, TAZ and KIBRA in associating with the known Hippo PY motif-containing proteins.

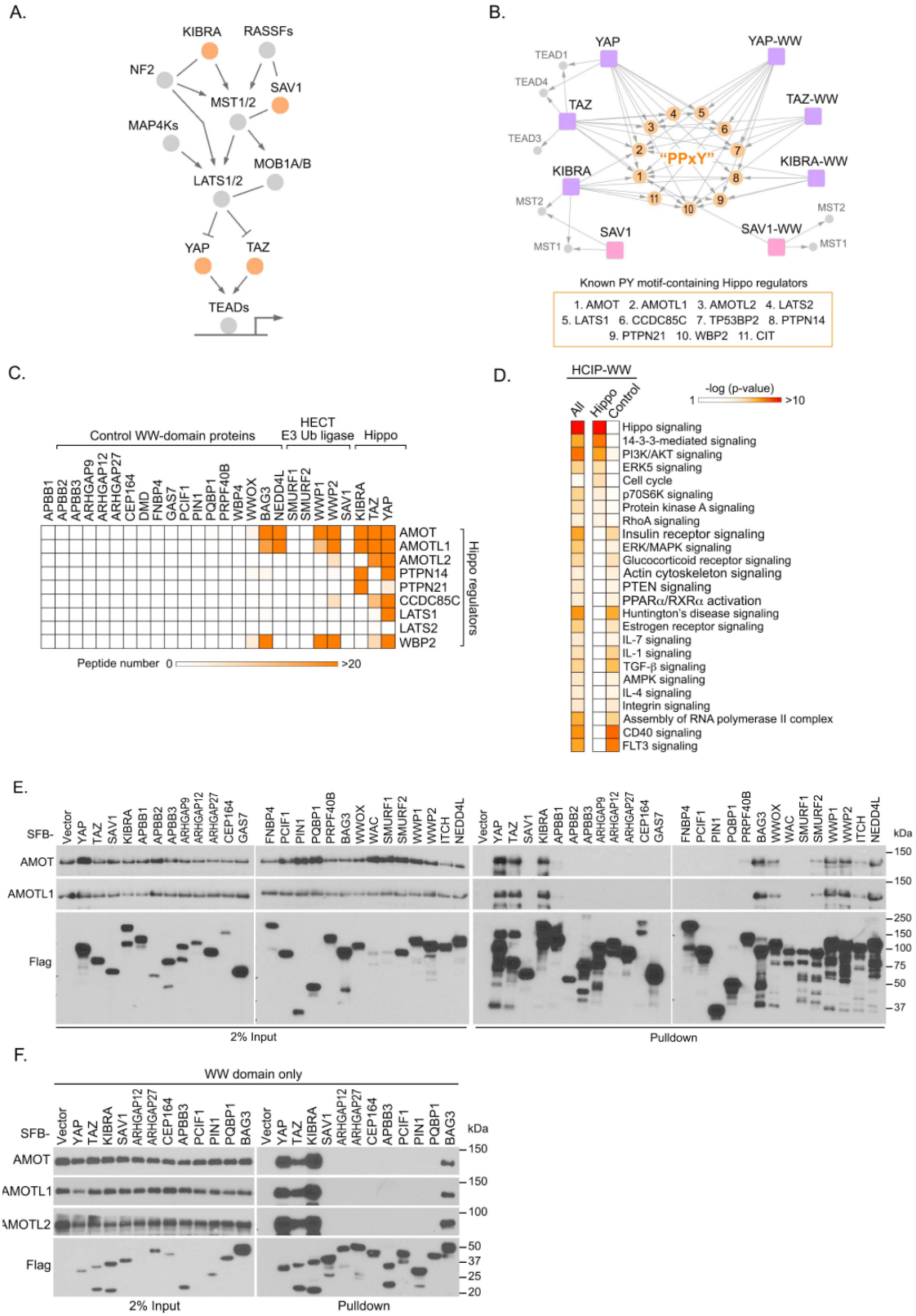

A.
KIBRA
RASSFs
NF2
SAV1
MST1/2
MAP4Ks
MOB1A/B
LATS1/2
YAP
TAZ
TEADs
B.
YAP
YAP-WW
TEAD1
TEAD4
TAZ
TAZ-WW
TEAD3
"PPxY"
KIBRA
KIBRA-WW
MST2
SAV1
SAV1-WW
MST1
Known PY motif-containing Hippo regulators
1. AMOT 2. AMOTL1 3. AMOTL2 4. LATS2
5. LATS1 6. CCDC85C 7. TP53BP2 8. PTPN14
9. PTPN21 10. WBP2 11. CIT
C.
Control WW-domain proteins
HECT E3 Ub ligase
Hippo
APBB1 APBB2 APBB3 ARHGAP9 ARHGAP12 ARHGAP27 CEP164 DMD FNBP4 GAS7 PCIF1 PIN1 PQBP1 PRPF40B WBP4 WWOX BAG3 NEDD4L SMURF1 SMURF2 WWP1 WWP2 SAV1 KIBRA TAZ YAP
AMOT
AMOTL1
AMOTL2
PTPN14
PTPN21
CCDC85C
LATS1
LATS2
WBP2
Hippo regulators
Peptide number 0 >20
D.
HCIP-WW
All
Hippo
Control
-log (p-value)
1 >10
Hippo signaling
14-3-3-mediated signaling
PI3K/AKT signaling
ERK5 signaling
Cell cycle
p70S6K signaling
Protein kinase A signaling
RhoA signaling
Insulin receptor signaling
ERK/MAPK signaling
Glucocorticoid receptor signaling
Actin cytoskeleton signaling
PTEN signaling
PPARα/RXRα activation
Huntington's disease signaling
Estrogen receptor signaling
IL-7 signaling
IL-1 signaling
TGF-β signaling
AMPK signaling
IL-4 signaling
Integrin signaling
Assembly of RNA polymerase II complex
CD40 signaling
FLT3 signaling
E.
SFB-
Vector YAP TAZ SAV1 KIBRA APBB1 APBB2 APBB3 ARHGAP9 ARHGAP12 ARHGAP27 CEP164 GAS7 FNBP4 PCIF1 PIN1 PQBP1 PRPF40B BAG3 WWOX WAC SMURF1 SMURF2 WWP1 WWP2 ITCH NEDD4L
kDa
AMOT
AMOTL1
Flag
150
150
250
150
100
75
50
37
2% Input
Pulldown
F.
WW domain only
SFB-
Vector YAP TAZ KIBRA SAV1 ARHGAP12 ARHGAP27 CEP164 APBB3 PCIF1 PIN1 PQBP1 BAG3
kDa
AMOT
AMOTL1
AMOTL2
Flag
150
150
100
50
37
25
20
2% Input
Pulldown

Figure 4.1: **The Hippo WW domain shows binding specificity with the known Hippo PY motif-containing proteins. (This figure is related to Figure C.1 and Tables C.1-C.4) (A)** Schematic illustration of the human Hippo pathway, where the Hippo WW domain-containing components are highlighted. **(B)** A summary map of cytoscape-generated merged interaction network for the Hippo WW domain-containing components and their WW domains. **(C)** The Hippo WW domain-containing proteins show binding specificity to the known Hippo PY motif-containing proteins. TAP-MS analysis of a series of WW domain-containing proteins were performed and their binding with the indicated Hippo PY motif-containing proteins was summarized in a heatmap. **(D)** The HCIPs for the Hippo WW domain-containing proteins were involved in different signaling pathways compared to those retrieved from the control WW domain-containing proteins. Gene Ontology analysis was performed. **(E)** Validation of the binding specificity for the Hippo WW domain-containing proteins. HEK293T cells were transfected with the indicated SFB-tagged constructs and subjected to the pulldown assay. **(F)** Validation of the binding specificity for the derived WW domains from the Hippo WW domain-containing proteins. HEK293T cells were transfected with the indicated SFB-tagged constructs and subjected to the pulldown assay.

Next, we expanded our proteomic analysis for additional 22 WW domain-containing proteins (Figure C.1B; Tables C.1-C.3) and examined their ability to isolate these known Hippo PY motif-containing proteins. Consistent with previous reports [6, 241, 290, 308, 333], WWOX, BAG3 and members of the HECT family of E3 ligases NEDD4L, WWP1 and WWP2 were found to form complexes with the Hippo PY motif-containing proteins such as AMOT family proteins, CCDC85C and WBP2 (Figure 4.1C). However, we failed to identify these Hippo PY motif-containing proteins as the binding proteins for other tested WW domain-containing proteins (Figure 4.1C). Moreover, the high-confident interacting proteins (HCIPs) of the Hippo WW domain-containing components were involved in different signaling pathways from those of the control WW domain-containing proteins (Figure 4.1D and Table EV4). We also performed proteomic analysis for the WW domains isolated from 13 randomly selected WW domain-containing proteins, and found that only 10.2% of the HCIPs were shared by the Hippo and control WW domains (Figure C.1C). Taken together, these results indicate that the WW domains of the Hippo pathway components YAP, TAZ and KIBRA possess a binding specificity with the known Hippo PY motif-containing proteins.

4.3.2 Validation of the Hippo WW domain binding specificity

To validate our proteomic findings, we examined the interaction between a series of WW domain-containing proteins and AMOT family proteins. Unlike YAP, TAZ and KIBRA, SAV1 failed to bind AMOT and AMOTL1 (Figure 4.1E). Consistently, we hardly detected the association between SAV1 and LATS1 in our experimental setting (Appendix Figure C.6A). Moreover, BAG3, WWOX and several members of the HECT family of E3 ligases can interact with AMOT proteins (Figure 4.1E), which is consistent with our proteomic study (Figure 4.1C). However, other tested WW domain-containing proteins as well as their derived WW domains failed to bind AMOT family proteins (Figures 4.1E, F). These results demonstrate the WW domain binding specificity for the Hippo pathway proteins.

4.3.3 A highly conserved amino acid sequence is required for the Hippo WW domain binding specificity

To further explore the underlying mechanism, we analyzed the WW domain protein sequence for the Hippo pathway components as well as WWOX, BAG3 and several members of the HECT family of E3 ligases, which can bind the known Hippo PY motif-containing proteins (Figure 4.2A). Interestingly, in addition to the two tryptophan residues, additional 9 amino acids were found to be highly conserved among these WW domains (Figure 4.2A). We hypothesized that this conserved 9-amino acid sequence could be required for the specific association with the known Hippo PY motif-containing proteins.

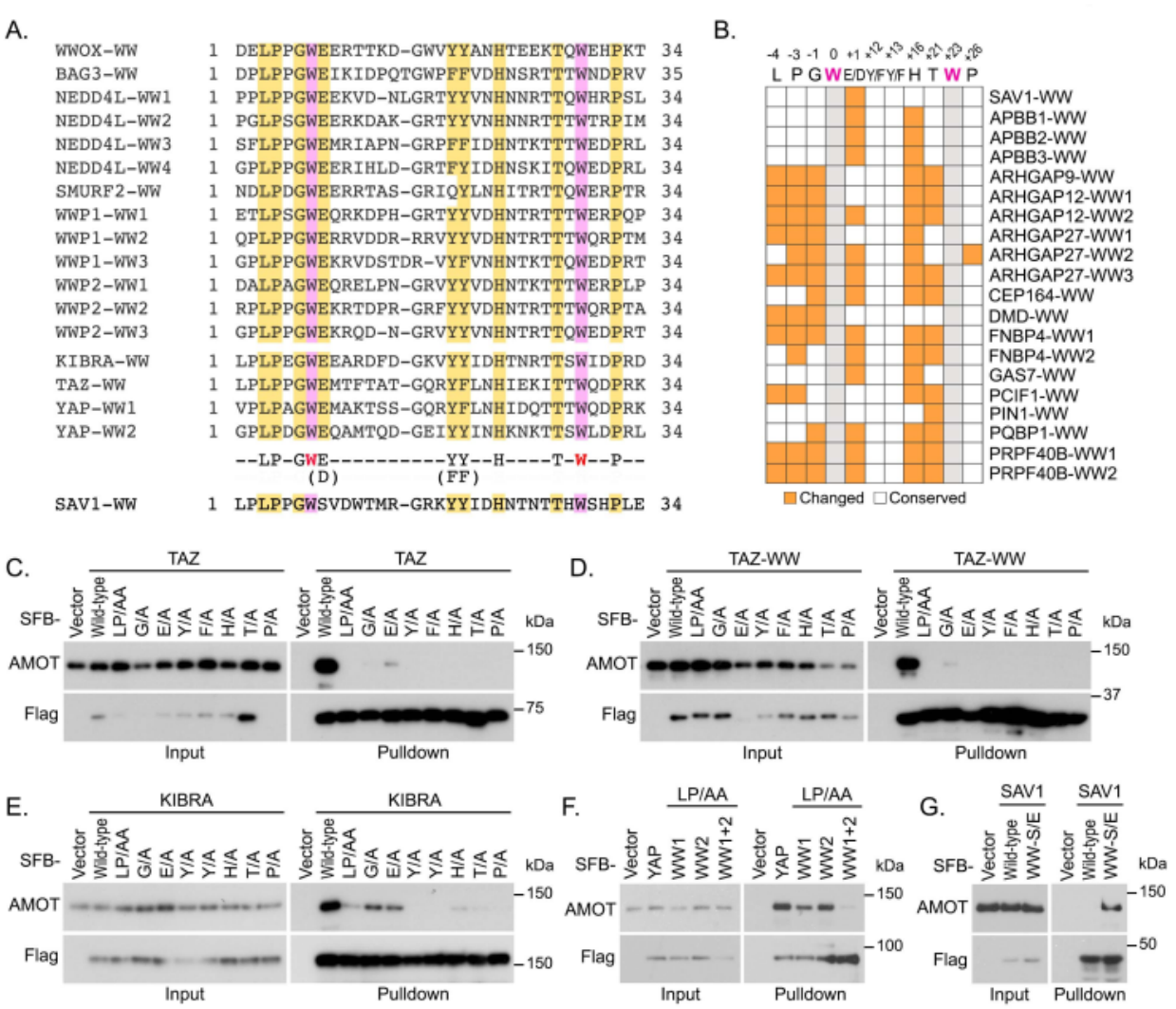


Figure 4.2: **Identification of a conserved 9-amino acid sequence that determines the Hippo WW domain binding specificity. (This Figure is related to Figures C.2 and C.3; Figures C.6-C.8; Table EV5) (A)** Sequence alignment of the WW domains derived from the WW domain-containing proteins that are known to bind the Hippo PY motif-containing proteins. The two conserved tryptophan restudies were highlighted in purple. Additional conserved amino acid residues were highlighted in yellow. **(B)** Summary of the residue difference in the identified 9-amino acid sequence for the control WW domains. The conserved two tryptophan residues are labelled in grey; the changed residues are labelled in orange; and the unchanged residues are labelled in white. **(C-G)** Validation of the identified 9-amino acid sequence in determining the Hippo WW domain binding specificity. The requirement of the identified 9-amino acid sequence for AMOT association was respectively examined for TAZ (C), TAZ-WW domain (D), KIBRA (E), YAP (F) and SAV1 (G). HEK293T cells were transfected with the indicated SFB-tagged constructs and subjected to the pulldown assay.

To test this hypothesis, we examined the identified 9-amino acid sequence in the control WW domain-containing proteins that failed to bind the Hippo PY motif-containing proteins (Figure 4.1C) and found that their WW domains have at least one of these 9 amino acids replaced by other residues (Figures 4.2B and C.2A). As for SAV1, the conserved glutamate residue within this 9-amino acid sequence was found changed to a serine in its WW domain (Figure 4.2A). Consistently, mutating either of these identified 9 amino acids to alanine dramatically disrupted the association of AMOT with TAZ (Figure 4.2C) or its WW domain (Figure 4.2D). Similar findings were also observed for both KIBRA (Figure 4.2E) and YAP (Figure 4.2F). Notably, mutations of the G and E residues among these identified 9 amino acids are less detrimental to the Hippo WW-PY interaction as compared with other identified sites (Figures 4.2C-4.2E). We also tested the conservative substitution for the "E/D", "Y/F" or "F/Y" of this conserved amino acid sequence, and found that the association of AMOT with TAZ and KIBRA was not affected by these substitutions (Appendix Figure C.6B). Interestingly, an interaction between SAV1 and AMOT was recovered when the unmatched serine residue was replaced by glutamate, allowing SAV1 WW domain to fit the 9-amino acid sequence criterion (Figure 4.2G). Taken together, these results demonstrate that the identified 9-amino acid sequence determines the WW domain binding specificity for the Hippo pathway proteins.

We also examined the Hippo WW domain-containing components in Drosophila and found that this 9-amino acid sequence was highly conserved in the WW domain of Yorkie and Kibra, while Salvador similarly contains a replacement of the conserved glutamate residue by alanine (Appendix Figure C.7). By taking YAP as an example, conservation of this 9-amino acid sequence in the YAP WW domains can be even tracked to Capsapsora owczarzaki (Figure C.2B and Table EV5), an unicellular specie that is known to contain the functional Hippo pathway components [249]. Interestingly, in Capsapsora owczarzaki, a PY motif was also identified in LATS (Figure C.2C), suggesting that this conserved 9-amino acid sequence may play a crucial role for the Hippo pathway at its premetazoan origin.

4.3.4 Role of the 9-amino acid sequence in assembly of a specific WW-PY complex involving the Hippo pathway proteins

Next, we analyzed a NMR solution structure of the YAP-WW1 domain (the first WW domain of YAP) and SMAD7-PY motif-containing peptide complex [10]. Interestingly, the identified 9 amino acids form as two functional groups.

First, together with the second tryptophan (W199 of YAP-WW1), the conserved residues E178, Y188, H192 and T197 were involved in the binding interface with the SMAD7-PY motif (Figure C.3A). Specifically, hydrogen bond (H-bond) formation was respectively paired between H192 (YAP-WW1 domain) and Y211 (SMAD7-PY motif), and T197 (YAP-WW1 domain) and P209 (SMAD7-PY motif) (Figures C.3B and C.3C). Hydrophobic contact not only existed within the intramolecular interaction between the W199 and Y188 residues of YAP1-WW domain, but also mediated their intermolecular interaction with the P208 and P209 residues within SMAD7-PY motif, respectively (Figures C.3B and C.3C). E178 (YAP-WW1 domain) functioned in sustaining the intermolecular contact between H192 (YAP-WW1 domain) and Y211 (SMAD7-PY motif) by forming both electrostatic and H-bonding interactions with H192 (Figures C.3B and C.3C).

Second, together with the first tryptophan (W177 of YAP-WW1 domain), the rest residues L173, P174, G176, F189 and P202 formed a hydrophobic cluster at the backside of the YAP-WW1/SMAD7-PY complex (Figures C.3A and C.3C). Although not directly interacted with SMAD7-PY motif, this hydrophobic cluster may maintain a unique YAP-WW1 domain structure to facilitate its binding with SMAD7-PY motif. Since these hydrophobic cluster residues are also frequently replaced by other amino acids in the non-Hippo WW domains (Figures 4.2B and C.2A), we consider them as part of the determinants for the specific Hippo WW-PY recognition.

To further determine the role of this identified 9-amino acid sequence from a structure-

based perspective, we mutated each of these conserved residues into alanine *in silico* and performed root-mean-square deviation (RMSD) analyses using the average unbound (apo) structure of YAP-WW1 domain as a reference. Interestingly, mutating either of the identified residues within the backside hydrophobic cluster significantly altered the YAP-WW1 protein structure as indicated by their relatively high RMSD values, while this was not the case for the residues within the binding interface with SMAD7-PY motif (Figure C.3D). These results further confirm the hypothesis that the backside hydrophobic cluster may play a role in maintaining a functional YAP-WW1 structure. In addition, mutating either of the conserved residues altered the complex structure (Figure C.3E and Appendix Figure C.8A) and increased the average distance between YAP-WW1 domain and SMAD7-PY motif peptide (Figure C.3F), indicating the intervention of their complex formation. As a control, we analyzed a NMR solution structure of the APBB3-WW domain (Appendix Figure C.8B). The APBB3-WW domain failed to bind the Hippo PY motif-containing proteins (Figure 4.1F), since it contains two unmatched residues (as compared to the identified 9-amino acid sequence) locating in the PY motif binding interface (Figures 4.2B and C.2A; Appendix Figure C.8B). Consistently, the average distance between APBB3-WW domain and SMAD7-PY motif peptide is comparable to that between YAP-WW1 domain mutants and SMAD7-PY motif peptide (Figure C.3F), suggesting an unstable complex formation for APBB3-WW domain and SMAD7-PY motif. Notably, the standard deviation of average distance value for both YAP-WW1 domain mutants and APBB3-WW domain complexes is relatively larger than that of the control YAP-WW domain complex (Figure C.3F), indicating a substantial movement between SMAD7-PY motif peptide and the YAP-WW1 domain mutants as well as APPB3-WW domain.

Taken together, these simulation analyses suggest that the identified 9-amino acid sequence is involved in binding PY motif and maintaining a unique WW domain structure, which both determine the Hippo WW domain binding specificity with the known Hippo PY motif-containing proteins.

4.3.5 Identification of STXBP4, a WW domain-containing protein, whose WW domain fits the 9-amino acid sequence criterion

Next, we searched all the WW domain-containing proteins in the human proteome and identified 12 WW domain-containing proteins whose WW domains fit such a 9-amino acid sequence (Figure C.4 and Table EV6). Among them, role of STXBP4 in the Hippo pathway regulation has not been fully characterized (Figure C.4). Although no STXBP4 ortholog is identified in Drosophila, this 9-amino acid sequence of the STXBP4 WW domain was largely conserved in different species (Figure 4.3A). Interestingly, STXBP4 can form a complex with several Hippo PY motif-containing regulators including AMOT, AMOTL2 and PTPN14 (Figure 4.3B). Mutating either of the conserved 9-amino acid residues diminished the interaction between STXBP4 and AMOT (Figure 4.3C). As expected, the association between STXBP4 and these PY motif-containing Hippo regulators are mediated by the WW domain of STXBP4 (Appendix Figure C.9A) and the PY motif of these Hippo regulators (Appendix Figure C.9B).

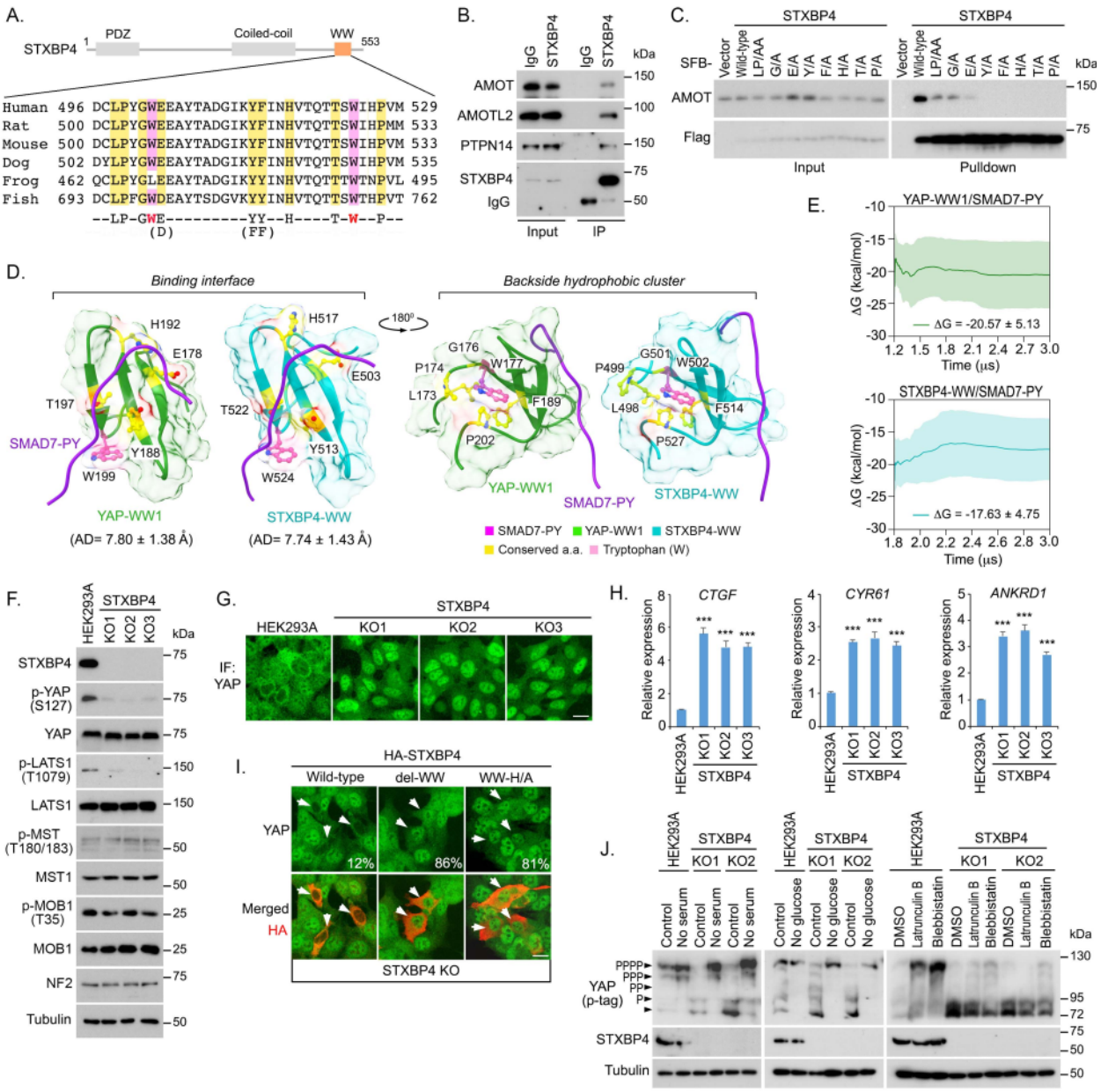

A.
STXBP4
PDZ
Coiled-coil
WW
553
Human 496 DCLPYGWEEAYTADGIKYFINHVTQTTSWIHPVM 529
Rat 500 DCLPYGWEEAYTADGIKYFINHVTQTTSWIHPMM 533
Mouse 500 DCLPYGWEEAYTADGIKYFINHVTQTTSWIHPVM 533
Dog 502 DYLPYGWEEAYTADGIKYFINHVTQTTSWIHPVM 535
Frog 462 QCLPYGLEEAYTADGIKYYINHVTQTTTWTNPVL 495
Fish 693 DCLPFGWDEAYTSDGVKYYINHVTQTTSWTHPVT 762
--LP-GWE---------YY--H----T-W--P--
(D) (FF)
B.
IgG
STXBP4
kDa
AMOT
AMOTL2
PTPN14
STXBP4
IgG
Input
IP
C.
STXBP4
SFB-
Vector
Wild-type
LP/AA
G/A
E/A
Y/A
F/A
H/A
T/A
P/A
AMOT
Flag
Input
Pulldown
D.
Binding interface
Backside hydrophobic cluster
H192
E178
T197
SMAD7-PY
Y188
W199
YAP-WW1
(AD= 7.80 ± 1.38 Å)
H517
E503
T522
Y513
W524
STXBP4-WW
(AD= 7.74 ± 1.43 Å)
180°
G176
P174
W177
L173
F189
P202
YAP-WW1
P499
G501
W502
L498
F514
P527
SMAD7-PY
STXBP4-WW
SMAD7-PY
YAP-WW1
STXBP4-WW
Conserved a.a.
Tryptophan (W)
E.
YAP-WW1/SMAD7-PY
ΔG (kcal/mol)
ΔG = -20.57 ± 5.13
Time (μs)
STXBP4-WW/SMAD7-PY
ΔG (kcal/mol)
ΔG = -17.63 ± 4.75
Time (μs)
F.
HEK293A
STXBP4
KO1
KO2
KO3
kDa
STXBP4
p-YAP (S127)
YAP
p-LATS1 (T1079)
LATS1
p-MST (T180/183)
MST1
p-MOB1 (T35)
MOB1
NF2
Tubulin
G.
HEK293A
STXBP4
KO1
KO2
KO3
IF: YAP
I.
HA-STXBP4
Wild-type
del-WW
WW-H/A
YAP
12%
86%
81%
Merged
HA
STXBP4 KO
H.
CTGF
CYR61
ANKRD1
Relative expression
HEK293A
KO1
KO2
KO3
STXBP4
J.
HEK293A
STXBP4
KO1
KO2
Control
No serum
No glucose
DMSO
Latrunculin B
Blebbistatin
kDa
YAP (p-tag)
PPPP
PPP
PP
P
STXBP4
Tubulin

Figure 4.3: **STXBP4 is a Hippo pathway regulator, which contains a WW domain that fits the criterion of the Hippo WW domain binding specificity. (This figure is related to Figure C.4, Appendix Figures C.9 and C.10; Table EV6)**
(A) Schematic illustration of STXBP4 protein, where the identified 9-amino acid sequence of STXBP4-WW domain was aligned across the indicated species.
(B) STXBP4 forms a complex with several Hippo PY motif-containing proteins. Immunoprecipitation was performed with STXBP4 antibody.
(C) The identified 9-amino acid sequence is required for the association between STXBP4 and AMOT. HEK293T cells were transfected with the indicated STXBP4 mutants and subjected to the pulldown assay.
(D) Structural comparison between the YAP-WW1/SMAD7-PY and STXBP4-WW/SMAD7-PY complexes. The identified 9 amino acid residues were indicated for both complexes.
(E) The YAP-WW1/SMAD7-PY and STXBP4-WW/SMAD7-PY complexes show similar cumulative average trend and average binding free energy (ΔG) within standard deviation (the shaded region) of one another.
(F) Loss of STXBP4 inhibits YAP phosphorylation and LATS activation. Western blotting was performed with the indicated antibodies.
(G and H) Loss of STXBP4 activates YAP. STXBP4 deficiency promotes YAP nuclear translocation (G) and YAP downstream gene transcription (mean $\pm$ s.d., n=3 biological replicates) (H). Scale bar, 20 μm.*** $p < 0.001$ (Student's t-test).
(I) WW domain is required for the STXBP4-mediated YAP cytoplasmic translocation. STXBP4 KO cells were transfected with the indicated STXBP4 constructs and immunofluorescent staining was performed. HA-positive cells (arrows) from $\sim$ 30 different views ($\sim$ 200 cells in total) were randomly selected and quantified for YAP localization. Percentage of HA-positive cells with nuclear YAP enrichment is shown. Scale bar, 20 μm.

To gain a structural insight into the STXBP4 WW domain, we compared STXBP4-WW and YAP-WW1 through ensemble molecular dynamics simulations and calculating binding free energies (ΔG) using the molecular mechanics Poisson-Boltzman surface area (MM/PBSA) method. As shown in Appendix Figure C.9C, the top 5 predicted clusters for the STXBP4-WW/SMAD7-PY complex is similar to those of the YAP-WW1/SMAD7-PY complex. By comparing the top one cluster for these two WW-PY complexes, we found that the identified 9-amino acid residues as well as the two tryptophan residues are similarly distributed within both the STXBP4-WW/SMAD7-PY and YAP-WW1/SMAD7-PY complexes, where they form as two groups to respectively involve in the binding with SMAD7-PY motif and assemble

a supportive backside hydrophobic cluster for each WW domain (Figure 4.3D). The average distance between STXBP4-WW domain and SMAD7-PY motif is close to that between YAP-WW1 domain and SMAD7-PY motif with a similarly low standard deviation value (Figure C.3F). Moreover, binding free energy (ΔG) from MM/PBSA calculations further indicates the similarity between YAP-WW1 and STXBP4-WW when they form as a complex with SMAD7-PY motif peptide (Figure 4.3E).

Taken together, these data suggest that the STXBP4 WW domain possesses the Hippo WW domain binding specificity, endowing STXBP4 a potential role in the Hippo pathway.

4.3.6 STXBP4 is a negative regulator of YAP

To test the role of STXBP4 in regulation of the Hippo pathway, we examined YAP activation in the STXBP4 knockout (KO) cells (Appendix Figure C.10). Interestingly, loss of STXBP4 significantly reduced YAP phosphorylation (Figure 4.3F), moved YAP into the nucleus (Figure 4.3G) and activated YAP downstream gene transcription (Figure 4.3H). Notably, either deleting the WW domain or mutating the histidine residue out of the identified 9-amino acid sequence to alanine failed to rescue YAP's cytoplasmic localization (Figure 4.3I), suggesting that the WW domain is required for the STXBP4-mediated YAP inhibition.

The observation that STXBP4 deficiency reduced YAP phosphorylation at S127 (Figure 4.3F) suggests that the Hippo pathway is inhibited in the STXBP4 KO cells. Indeed, as shown in Figure 4.3F, loss of STXBP4 suppressed LATS phosphorylation but did not affect that of MST or its substrate MOB1. These data suggest that STXBP4 is required for LATS activation in the Hippo pathway.

4.3.7 STXBP4 is involved in a protein-protein interaction network comprising multiple Hippo pathway components and regulators

To elucidate the mechanism by which STXBP4 regulates the Hippo pathway, we purified the STXBP4-associated protein complex and characterized its binding partners by mass spectrometry analysis. As shown in Figure 4.4A, all the AMOT family proteins were identified to form a complex with STXBP4, which is consistent with our previous findings (Figures 4.3B and Appendix Figure C.9A). Interestingly, we also identified α-catenin, a known Hippo upstream regulator [86, 230, 245, 301], as a binding partner for STXBP4 (Figure 4.4A). STXBP4 was also reciprocally identified as a binding protein for some Hippo pathway components (e.g., LATS1, LATS2, TAZ) and regulators (e.g., AMOT, AMOTL1, AMOTL2, PTPN14) [63, 122, 315] (Figure 4.4A). Collectively, these data suggest that STXBP4 involves in a protein-protein interaction network comprising a group of Hippo pathway components and regulators.

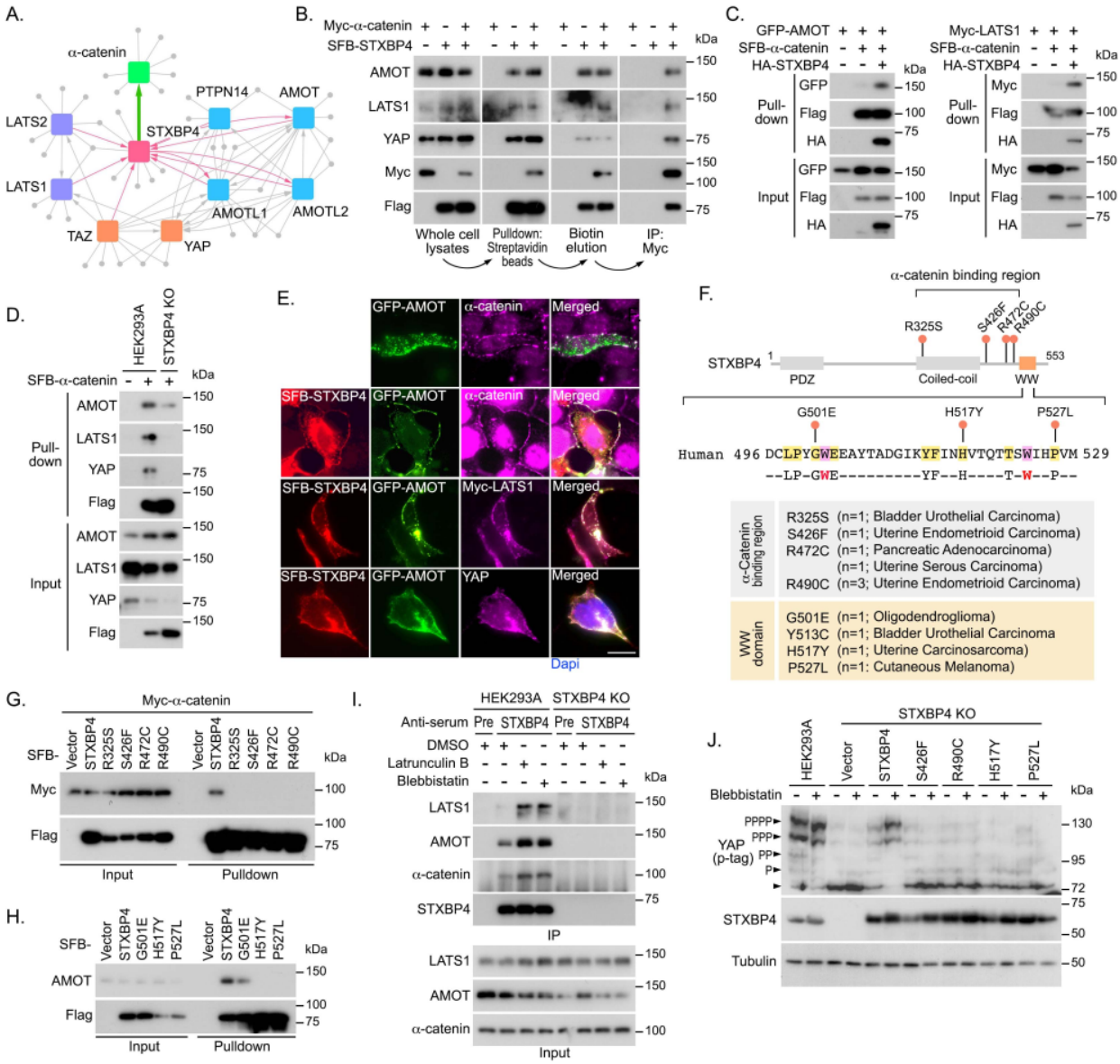

A.
α-catenin
PTPN14
AMOT
LATS2
STXBP4
LATS1
AMOTL1
AMOTL2
TAZ
YAP
B.
Myc-α-catenin
SFB-STXBP4
kDa
AMOT
LATS1
YAP
Myc
Flag
Whole cell lysates
Pulldown: Streptavidin beads
Biotin elution
IP: Myc
C.
GFP-AMOT
SFB-α-catenin
HA-STXBP4
Myc-LATS1
Pull-down
Input
GFP
Flag
HA
Myc
D.
HEK293A
STXBP4 KO
SFB-α-catenin
E.
GFP-AMOT
α-catenin
Merged
SFB-STXBP4
Myc-LATS1
Dapi
F.
α-catenin binding region
R325S
S426F
R472C
R490C
STXBP4
PDZ
Coiled-coil
WW
G501E
H517Y
P527L
Human 496 DCLPYGWEEAYTADGIKYFINHVTQTTSWIHPVM 529
--LP-GWE---------YF--H----T-W--P--
α-Catenin binding region
R325S (n=1; Bladder Urothelial Carcinoma)
S426F (n=1; Uterine Endometrioid Carcinoma)
R472C (n=1; Pancreatic Adenocarcinoma)
(n=1; Uterine Serous Carcinoma)
R490C (n=3; Uterine Endometrioid Carcinoma)
WW domain
G501E (n=1; Oligodendroglioma)
Y513C (n=1; Bladder Urothelial Carcinoma
H517Y (n=1; Uterine Carcinosarcoma)
P527L (n=1: Cutaneous Melanoma)
G.
Myc-α-catenin
SFB-
Vector
STXBP4
R325S
S426F
R472C
R490C
Myc
Flag
Input
Pulldown
H.
G501E
H517Y
P527L
I.
HEK293A
STXBP4 KO
Anti-serum
Pre
DMSO
Latrunculin B
Blebbistatin
IP
J.
Blebbistatin
YAP (p-tag)
Tubulin

Figure 4.4: **STXBP4 functions in the actin cytoskeleton tension-mediated Hippo pathway regulation by forming a complex with α-catenin and a group of Hippo PY motif-containing proteins. (This figure is related to Appendix Figures C.11 and C.12; Table EV7)**
(A) A summary map of cytoscape-generated protein-protein interaction network for STXBP4, α-catenin and a group of Hippo pathway proteins.
(B) STXBP4 forms a protein complex with α-catenin and a group of Hippo pathway proteins.
(C) STXBP4 promotes the association of α-catenin with AMOT and LATS1. HEK293T cells were transfected with the indicated SFB-tagged constructs and subjected to the pulldown assay.
(D) Loss of STXBP4 diminishes the association of α-catenin with AMOT, LATS1 and YAP. HEK293A and STXBP4 KO cells were transfected with the SFB-tagged α-catenin construct and subjected to the pulldown assay.
(E) STXBP4 induces the co-localization between α-catenin and AMOT as well as LATS1 and YAP. HEK293A cells were transfected with the indicated constructs and immunofluorescence was performed. Scale bar, 20 μm.
(F-H) Identification of several STXBP4 missense mutations that disrupt its interaction with α-catenin and AMOT. The missense mutations within the STXBP4 α-catenin-binding region and the 9-amino acid sequence of the STXBP4 WW domain were indicated and annotated (F). The identified missense mutations respectively disrupted the STXBP4 α-catenin (G) and STXBP4-AMOT (H) complex formation.
(I) Inhibition of actin cytoskeleton promotes the STXBP4-associated protein complex formation. HEK293A and the STXBP4 KO cells were subjected to immunoprecipitation using pre-immune serum and anti-STXBP4 serum under the indicated treatments.
(J) The missense mutations of STXBP4 (F) diminished the ability of STXBP4 to rescue YAP phosphorylation in the STXBP4 KO cells with low actin cytoskeleton tension. YAP phosphorylation was detected using phospho-tag gel, where the YAP phosphorylation level was indicated.

Notably, most of these STXBP4-associated proteins are PY-motif containing proteins (Figure 4.4A), suggesting that STXBP4 WW domain is required here. Since α-catenin does not contain a PY motif, we further characterized the α-catenin-binding region in STXBP4. To achieve this, a series of STXBP4 truncation and deletion mutants were generated (Appendix Figure C.11A). As shown in Appendix Figure C.11B, deletion of the 300 $\sim$ 500 amino acid residues of STXBP4, but not its WW domain, fully abolished its association with α-catenin. Moreover, we failed to further narrow down the α-catenin binding region in STXBP4 (Appendix Figure C.10C), suggesting that this identified 300 $\sim$ 500 amino acid

sequence region is required for its interaction with α-catenin.

Taken together, these data indicate that STXBP4 can form a complex with several Hippo PY motif-containing proteins and α-catenin through its WW domain and the 300$\sim$500 amino acid sequence region, respectively.

4.3.8 STXBP4 functions as a scaffold protein to assemble a protein complex including α-catenin AMOT, LATS and YAP

To test this hypothesis, we performed a sequential pulldown/immunoprecipitation assay using exogenously expressed SFB-STXBP4 and Myc-α-catenin in HEK293T cells. As shown in Figure 4.4B, we first isolated STXBP4-associated protein complex using streptavidin beads, eluted the complex with biotin, and purified the α-catenin-associated protein complex through immunoprecipitation. This sequential purification approach can help to characterize the proteins within the STXBP4/α-catenin protein complex. Consistent with our proteomic data (Figure 4.4A), AMOT, LATS1 and YAP were all identified within the STXBP4/α-catenin protein complex (Figure 4.4B).

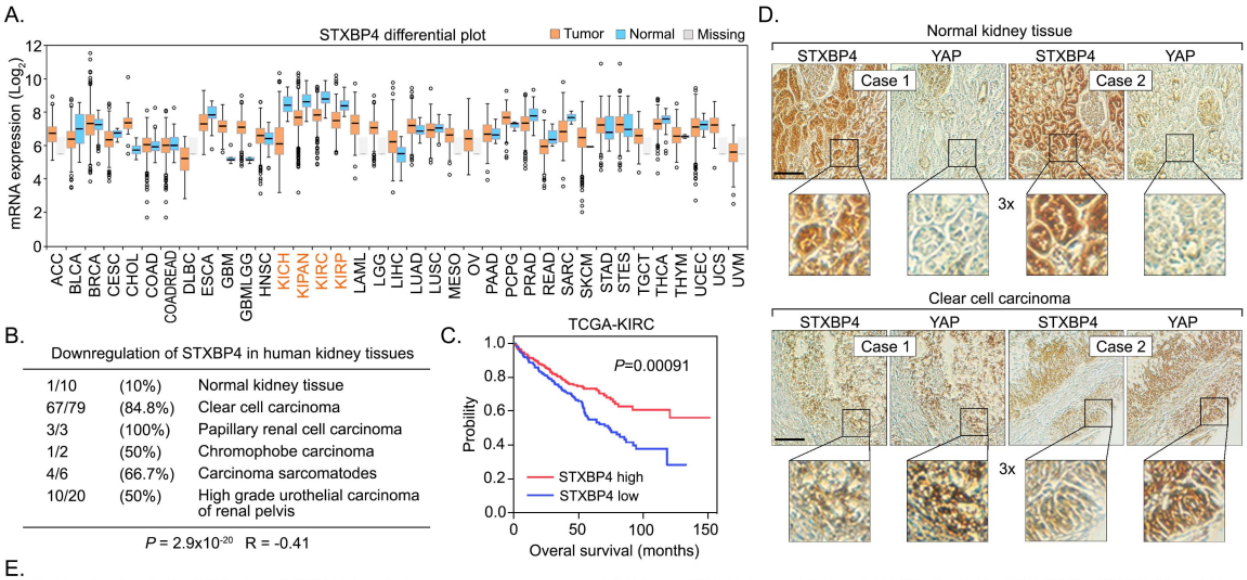
Next, we examined the role of STXBP4 in this multi-protein complex. Overexpression of STXBP4 induced the interaction of α-catenin with both AMOT and LATS1 (Figure 4.4C); while loss of STXBP4 largely attenuated the association of α-catenin with AMOT, LATS1 and YAP (Figure 4.4D). In addition, STXBP4 promoted the co-localization between AMOT and α-catenin onto cell adherens junction/membrane region, where both LATS1 and YAP were also identified (Figure 4.4E). These results suggest a scaffold role of STXBP4 in assembly of a protein complex containing at least α-catenin, AMOT, LATS and YAP at adherens junctions.

Both the WW domain and α-catenin association are required for the STXBP4-mediated

YAP regulation Given the potential tumor suppressive role of STXBP4 in targeting YAP, we next examined the genetic alteration of STXBP4 in the cBioportal database and found that STXBP4 alleles harbor a series of mutations within cancer patient samples (Appendix Figure C.12A and Table EV7). Four missense mutations that are localized in the α-catenin-binding region (Figure 4.4F) disrupted the interaction between STXBP4 and α-catenin (Figure 4.4G). As for the STXBP4 WW domain, four out of the identified 9 amino acid residues were found mutated in oligodendrogioma (G501E), bladder urothelial carcinoma (Y513C), uterine carcinosarcoma (H517Y) and cutaneous melanoma (P527L), respectively (Figure 4.4F), and they all diminished the association between STXBP4 and AMOT (Figure 4.4H). Notably, these cancer-derived missense mutations in either α-catenin-binding region or the WW domain of STXBP4 all failed to rescue YAP's cytoplasmic localization in the STXBP4 KO cells (Appendix Figure C.12B), suggesting that association with α-catenin and the Hippo PY motif-containing components/regulators is required for the STXBP4-dependent the Hippo pathway regulation.

STXBP4 functions as a potential mechano-transducer involved in actin cytoskeleton-mediated Hippo pathway regulation Notably, α-catenin is known to play a critical role in mechanotransduction [54, 334], and loss of STXBP4 significantly attenuated YAP phosphorylation upon disruption of actin cytoskeleton or inhibition of its tension (Figure 4.3J). Interestingly, depolymerization of actin cytoskeleton by latrunculin B or inhibition of its tension by blebbistatin induced the association of STXBP4 with LATS1, AMOT and α-catenin (Figure 4.4I). Reconstitution of STXBP4, but not its mutants with missense mutations at its α-catenin-binding region and WW domain (Figure 4.4F), significantly rescued YAP phosphorylation when actin cytoskeleton tension was inhibited (Figure 4.4J). These data indicate that the STXBP4-mediated protein complex formation with α-catenin and the Hippo PY motif-containing proteins plays a role in actin cytoskeleton-dependent regulation of the Hippo pathway.

STXBP4 is frequently downregulated in kidney cancer and correlated with YAP activation By analyzing the cancer database, FireBrowse, a platform developed to analyze 14,729 tumor sample data generated by The Cancer Genome Atlas (TCGA), we found that the mRNA level of STXBP4 was downregulated in all the listed kidney cancer subtypes (Figure 4.5A). This finding was further confirmed through a kidney tissue microarray analysis, where the expression of STXBP4 was found decreased in several types of human kidney cancer: 84.8% clear cell carcinoma, 100% papillary renal cell carcinoma, 50% chromophobe carcinoma, 66.7% carcinoma sarcomatodes and 50% high grade urothelial carcinoma of renal pelvis (Figure 4.5B). However, downregulation of STXBP4 was only observed in 10% normal kidney tissue (Figure 4.5B), suggesting an inverse correlation between STXBP4 expression and kidney cancer formation (P=2.9x10^{-20}, R=-0.41). Moreover, our TCGA data analysis indicated that low expression of STXBP4 was significantly correlated with the poor overall survival rate for the cancer patients with clear cell renal cell carcinoma (ccRCC) (Figure 4.5C), indicating that STXBP4 is a potential tumor suppressor in ccRCC.



E.

P=1.53x10-7 R=-0.56	STXBP4-low	STXBP4-high	Total
Normal kidney tissue	1	9	10
Clear cell carcinoma	67	12	79
Total	68	21	89

P=0.027 R=0.23	YAP-low	YAP-high	Total
Normal kidney tissue	8	2	10
Clear cell carcinoma	34	45	79
Total	42	47	89

P=0.0036 R=-0.46	YAP(N≤C)	YAP(N>C)	Total
STXBP4-low	2	36	38
STXBP4-high	4	5	9
Total	6	41	47

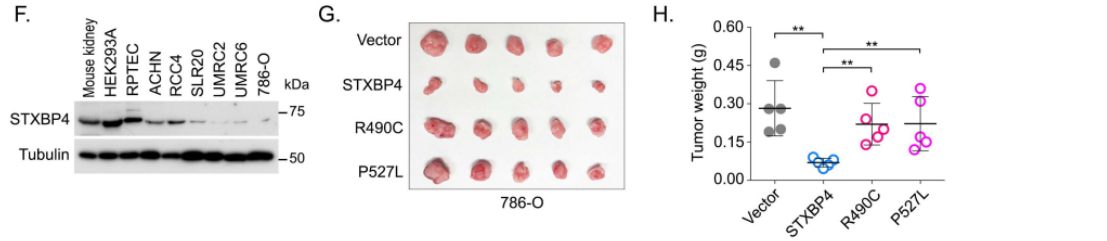


Figure 4.5: **STXBP4 is a tumor suppressor in human kidney cancer. (This figure is related to Appendix Figure C.13).**
(A and B) STXBP4 is downregulated in human kidney cancer. The mRNA level of STXBP4 is analyzed in the Firebrowse web database (http://firebrowse.org) (A), where 14,729 tumor sample data generated by TCGA were included. The first quartile, median and third quartile values were indicated as the boxplots. Outliers were plotted as individual points. Error bars indicated the standard deviation above and below the mean of the data. The expression of STXBP4 was also examined using kidney tissue microarray, where percentage of the indicated tissue samples with downregulated STXBP4 was shown (B). The p value was calculated by using the paired Student's t-test.
(C) Kaplan-Meier curves of overall survival of patients with ccRCC is stratified by STXBP4 expression level. Clinical data of STXBP4 were analyzed in TCGA-KIRC project containing total 611 patient samples. The p value was calculated by using the Log-rank (Mantel-Cox) test.
(D) Immunohistochemical staining of STXBP4 and YAP were performed in a kidney cancer tissue microarray, where the indicated regions in the box were shown three times enlarged. Brown staining indicates positive immunoreactivity. Scale bar, 100 μm.
(E) Correlation analyses between STXBP4 and YAP in human normal kidney and clear cell carcinoma samples are shown as tables. Statistical significance was determined by chi-square test. R, correlation coefficient. N, nuclear localization. C, cytoplasmic localization.
(F) STXBP4 expression is examined in a panel of ccRCC cell lines by Western Blotting.
(G and H) Both the association with α-catenin and the functional WW domain are required for the STXBP4's tumor suppressive function in 786-O cells. Overexpression of STXBP4, but not the indicated STXBP4 missense mutants, significantly suppressed the 786-O cell xenograft tumor formation. Xenograft tumors are shown in (G), and the tumor weight is quantified in (H) (n = 5 mice, mean ± s.d.). ** $p < 0.01$ (Student's t-test). Scale bar, 1 cm.

YAP is highly expressed and activated in multiple major human cancer types but genetic mutation for the Hippo pathway components is hardly detected [127], suggesting that additional oncogenic alterations could lead to YAP activation for tumorigenesis. Since loss of STXBP4 activated YAP (Figures 4.3F-4.3H), we next examined the pathological correlation between STXBP4 and YAP using a kidney cancer tissue microarray. Consistent with previous studies [44, 94, 246], upregulation of YAP was observed in 57% (45 of 79) of ccRCC tissue samples, while only 20% (2 of 10) of normal kidney tissues showed high YAP expression (Figures 4.5D and 4.5E). Moreover, an inverse correlation between STXBP4 expression and YAP nuclear enrichment was found in the tissue samples with high YAP expression

(P=0.0036, R=-0.46), where 94.7% (36 of 38) of the tested tissue samples with low STXBP4 expression had high nuclear enrichment of YAP (Figures 4.5D and 4.5E). However, there were still 10.6% (5 of 47) of the total tested specimens showing high STXBP4 expression but YAP nuclear enrichment (Figure 4.5E). These results indicate that downregulation of STXBP4 may contribute to YAP activation in a substantial fraction of ccRCC; however, YAP can still be activated in other tumors via different mechanisms.

Interestingly, although a general low expression of YAP was found in normal kidney tissues, we were still able to observe a relatively high expression of YAP in the podocytes of glomerulus region and partially in the convoluted tubule region (Figure 4.5D). Even though, these YAP highly expressed normal kidney regions still consistently showed a decreased STXBP4 expression level (Figure 4.5D), suggesting that their inverse correlation in expression could involve in normal kidney physiology.

Both the α-catenin association and functional WW domain are required for the STXBP4's tumor suppressive function in kidney cancer To investigate the role of STXBP4 in kidney cancer, we first determined the STXBP4 expression in normal mouse kidney tissue and a group of human kidney-related cell lines. Interestingly, STXBP4 had an abundant expression in mouse kidney tissue, an embryonic kidney immortalized cell line HEK293A and an immortalized human renal proximal tubular epithelial cell line RPTEC (Figure 4.5F). In contrast, STXBP4 showed moderate or low expressions in all the tested ccRCC cell lines (Figure 4.5F), where YAP was found majorly localized in the nucleus (Appendix Figure C.13A). Overexpression of STXBP4, but not its two patient-derived missense mutants (R490C and P527L) (Figure 4.4F), in a ccRCC cell line 786-O (Appendix Figure C.13B), significantly suppressed the xenograft tumor formation (Figures 4.5G and 4.5H). Since the R490C and P527L mutations can respectively disrupt the STXBP4's interaction with α-catenin (Figure 4.4G) and AMOT (Figure 4.4H), these results indicate that the association with α-catenin and a functional WW domain are both required for STXBP4's tumor suppressive function.

4.4 Discussion

In this study, we identified a conserved 9-amino acid sequence within the WW domain of the Hippo pathway components and regulators (Figure 4.2), which is required for the specific Hippo WW-PY complex formation. Notably, this identified 9-amino acid sequence has at least one residue altered in all the tested control WW domain-containing proteins (Figures 4.2B and C.2A), which could help to explain why these control WW domain-containing proteins fail to interact with the Hippo PY motif-containing proteins (Figures 4.1E and 4.1F). Since the "WW-PY" recognition is widely present in the Hippo pathway, manipulation of their recognition is likely to control the outputs of this key signaling pathway in tissue/organ growth and tumorigenesis. Thus, it would be highly exciting if this Hippo WW domain determinants could be utilized for the development of small molecules or peptides to precisely modulate YAP/TAZ activity in cancer therapy and tissue repair.

Mechanistically, the identified 9-amino acid sequence accounts for both a suitable WW domain structure and the binding interface with the PY motif peptide (Figures C.3A-C.3C), providing a structural basis for the Hippo WW domain binding specificity. Here, our study is only focused on the individual WW domain binding property. Actually, the mechanism underlying the specific "WW-PY" recognition could be more complicated given the role of WW tandem in mediating PY motif binding [154] and the potential homo- and hetero-dimer formations among WW domains [272]. Moreover, although our current study mostly focused on the WW domain, it is highly possible that its cognate PY motif ligand could also contribute to the specific Hippo "WW-PY" recognition. However, the PY motif is relatively short, flexible and could be easily buried into a higher level of protein structure, making it difficult to assess its role at a protein level. Thus, we did not further address this question from the PY motif-based perspective. Among the Hippo pathway components, the SAV1 WW domain functions differently from that of YAP, TAZ and KIBRA to bind Hippo PY motif-containing proteins (Figure 4.1). This difference may arise from the change of one

conserved glutamate residue in the identified 9-amino acid sequence for the SAV1 WW domain in both human (Figures 4.2A and 4.2B) and Drosophila (Appendix Figure C.7). Based on our E/D substitution data (Appendix Figure C.6B) and the structural analysis (Figure C.3C), the negative charge for this residue position could be essential. Interestingly, the substituted serine residue within the human SAV1 WW domain can be phosphorylated in vivo (www.phosphosite.org), suggesting that the association between SAV1 and Hippo PY motif-containing proteins could be regulated through a yet-to-be characterized phosphorylation event.

There are only a few WW domain-containing proteins, whose WW domains fit such 9-amino acid sequence in human proteome (Figure C.4 and Table EV6). Among them, STXBP4 was found as a negative regulator for YAP (Figures 4.3F-4.3H) by forming a protein complex with a series of Hippo PY motif-containing proteins and an adherens junction component, α-catenin (Figure 4.4A). Interestingly, STXBP4 serves as a scaffold protein in this network and transduces actin-based mechanical cues to regulate the Hippo pathway. Since α-catenin is known to play a role in both cell density and cytoskeleton tension-dependent regulation of YAP [86, 230, 245, 301], our findings provided molecular insights into its downstream signaling events. Under the condition with low actin cytoskeleton tension, STXBP4 recruits several Hippo PY motif-containing proteins including at least AMOT, LATS to form a complex with α-catenin at adherens junction. YAP/TAZ are also within this complex based on their interaction with AMOT and LATS (Figure C.5). In proximity, LATS phosphorylates and inhibits YAP. When mechanical cues increase actin cytoskeleton tension, both the adherens junction-associated α-catenin and the filament actin-bound AMOT would be affected in their conformation, resulting in the protein complex disassembly and YAP activation (Figure C.5). Exactly how this α-catenin-STXBP4-Hippo PY proteins axis is coordinated with other related signaling events [77, 164, 223] in regulating the interplay between actin cytoskeleton and the Hippo-YAP/TAZ pathway deserves further investigation.

Intriguingly, our TCGA database and tissue microarray studies suggested that STXBP4 is a potential tumor suppressor in kidney cancer (Figures 4.5A-4.5C) and its downregulation is significantly correlated with YAP activation in ccRCC tissues (Figures 4.5D and 4.5E). YAP has been found highly expressed and activated in human kidney cancer including ccRCC [44, 94, 246]. Here, our study identified a pathological relevance between STXBP4 and YAP, providing a potential mechanism for the YAP activation in ccRCC. Notably, a CpG island was identified in the STXBP4 promoter, suggesting that the loss of STXBP4 could occur due to its promoter methylation. In addition, STXBP4 gene alleles harbor a relative high mutation rate (13.45%) including nonsense mutation (6.92%), frameshift deletion (1.92%), in frameshift deletion (0.38%) and gene fusion (4.23%) (Appendix Figure C.12A), which could also partially explain the loss of STXBP4 in cancer.

STXBP4 is originally identified as an insulin-regulated protein involved in GLUT4-mediated glucose transport in adipocyte [39], and functions as an inhibitory protein for the SNARE complex-dependent membrane fusion [309]. Dysregualted STXBP4 expression was associated with some SNPs in breast cancer [29, 65, 16]. Recent studies also implicated the role of STXBP4 in squamous cell carcinomas, by regulating N-terminally truncated isoform of p63 (ΔNp63) [201, 236]. Together with these studies, our findings in kidney cancer suggested a complex role of STXBP4 in cancer development, which could depend on tissue context.

4.5 Materials and Methods

4.5.1 Antibodies and chemicals

For Western blotting, anti-α-tubulin (T6199-200UL, 1:5000 dilution), anti-Flag (M2) (F3165-5MG, 1:5000 dilution), and anti-AMOTL1 (HPA001196, 1:1000 dilution) antibodies were obtained from Sigma-Aldrich. Anti-Myc (sc-40, 1:500 dilution) and anti-GFP (sc-9996, 1:1000

dilution) antibodies were purchased from Santa Cruz Biotechnology. Anti-phospho-YAP (S127) (4911S, 1:1000 dilution), anti-phospho-LATS1 (Thr1079) (8654S, 1:1000 dilution), anti-LATS1 (3477S, 1:1000 dilution), anti-phospho-MST (Thr180/Thr183) (3681S, 1:1000 dilution), anti-MST1 (3682S, 1:1000 dilution), anti-phospho-MOB1 (Thr35) (8699S, 1:1000 dilution), anti-MOB1 (3863S, 1:2000 dilution) and anti-NF2 (12896S, 1:2000 dilution) antibodies were purchased from Cell Signaling Technology. The AMOT, AMOTL2, PTPN14 and YAP polyclonal antibodies were generated as previously described [313, 314]. The STXBP4 antiserum was raised against MBP-STXBP4 (the 251∼553 amino acid residues) and polyclonal antibody was affinity-purified using an AminoLink Plus Immobilization and Purification Kit (Pierce).

For immunostaining, an anti-YAP (sc-101199, 1:200 dilution) monoclonal antibody was purchased from Santa Cruz Biotechnology. Anti-hemagglutinin (HA) polyclonal antibody (3724S, 1:3000 dilution) was obtained from Cell Signaling Technology.

For immunohistochemical staining, an anti-YAP (14074S, 1:15 dilution) monoclonal antibody was purchased from Cell Signaling Technology. The STXBP4 antiserum was raised against MBP-STXBP4 (the 1∼250 amino acid residues) and polyclonal antibody (1:200 dilution) was affinity-purified using an AminoLink Plus Immobilization and Purification Kit (Pierce). Latrunculin B and blebbistatin were obtained from Sigma-Aldrich.

4.5.2 Constructs and viruses

Plasmids encoding the indicated genes were obtained from the Human ORFeome V5.1 library or purchased from Harvard Plasmid DNA Resource Core and Dharmacon. All constructs were generated via polymerase chain reaction (PCR) and subcloned into a pDONOR201 vector using Gateway Technology (Invitrogen) as the entry clones. For tandem affinity purification, all entry clones were subsequently recombined into a lentiviral Gateway-compatible

destination vector for the expression of C-terminal SFB-tagged fusion proteins. Gateway-compatible destination vectors with the indicated SFB tag, HA tag, GFP tag or Myc tag were used to express various fusion proteins. PCR-mediated mutagenesis was used to generate all the indicated site mutations and internal region/domain deletion mutations.

All lentiviral supernatants were generated by transient transfection of HEK293T cells with the helper plasmids pSPAX2 and pMD2G (kindly provided by Dr. Zhou Songyang, Baylor College of Medicine) and harvested 48 hours later. Supernatants were passed through a 0.45-μm filter and used to infect cells with the addition of 8 μg/mL hexadimethrine bromide (Polybrene) (Sigma-Aldrich).

4.5.3 Cell culture and transfection

HEK293T, ACHN, SLR20 and UMRC6 cell lines were purchased from ATCC and kindly provided by Drs. Boyi Gan and Junjie Chen (MD Anderson Cancer Center). HEK293A cells were purchased from ThermoFisher and kindly provided by Dr. Jae-Il Park (MD Anderson Cancer Center). RPTEC, 786-O, RCC4 and UMRC2 cells were purchased from ATCC and kindly provided by Dr. Olga Razorenova (University of California, Irvine). HEK293T, HEK293A, RCC4, UMRC2 and UMRC6 cells were maintained in Dulbecco's modified essential medium (DMEM) supplemented with 10% fetal bovine serum at 37°C in 5% CO2 (v/v). SLR20 and 786-O cells were grown in RPMI-1640 medium supplemented with 10% fetal bovine serum at 37°C in 5% CO2 (v/v). RPTEC cells were maintained in DMEM/F12 medium supplemented with 5 pM triiodo-L-thyronine, 10 ng/mL epidermal growth factor, 3.5 μgmL ascorbic acid, 5 μg/mL transferrin, 5 μg/mL insulin, 25 ng/mL prostaglandin E1, 25 ng/mL hydrocortisone, 8.65 ng/mL sodium selenite and 1.2 mg/mL sodium bicarbonate at 37 °C in 5% CO2 (v/v). All the culture media contain 1% penicillin and streptomycin. Plasmid transfection was performed using a polyethylenimine reagent.

4.5.4 Immunofluorescent staining

Immunofluorescent staining was performed as described previously [311] with minor modifications. Briefly, cells cultured on coverslips were fixed with 4% paraformaldehyde for 10 minutes at room temperature and then extracted with 0.5% Triton X-100 solution for 5 minutes. For α-catenin-related immunofluorescent staining, cells were pretreated with PBS solution containing 0.5% Trion X-100 and 1% paraformaldehyde for 4 minutes, and subjected to 4% paraformaldehyde fixation. After blocking with Tris-buffered saline with Tween 20 containing 1% bovine serum albumin, the cells were incubated with the indicated primary antibodies for 1 hour at room temperature. After that, the cells were washed and incubated with fluorescein isothiocyanate-, rhodamine- and Cy5-conjugated secondary antibodies for 1 hour. Cells were counterstained with 100 ng/mL 4',6-diamidino-2-phenylindole (DAPI) for 2 minutes to visualize nuclear DNA. The coverslips were mounted onto glass slides with an anti-fade solution and visualized under a Nikon Eclipse Ti spinning-disk confocal microscope.

4.5.5 Tandem affinity purification (TAP) of SFB-tagged protein complexes

HEK293T cells stably expressing the indicated SFB-tagged proteins were selected by culturing in medium containing 2 μg/mL puromycin and confirmed by immunostaining and Western blotting as described previously [315]. For TAP, HEK293T cells were lysed in NETN buffer (with protease and phosphatase inhibitors) at 4°C for 20 minutes. The crude lysates were centrifuged at 14,000 rpm for 15 minutes at 4°C. The supernatants were incubated with streptavidin-conjugated beads (GE Healthcare) for 1 hour at 4°C. The beads were washed 3 times with NETN buffer, and bound proteins were eluted with NETN buffer containing 2 mg/mL biotin (Sigma-Aldrich) for 2 hours at 4°C. The elutes were incubated with S protein beads (Novagen) for 1 hour. The beads were washed three times with NETN

buffer and subjected to sodium dodecyl sulfate polyacrylamide gel electrophoresis. Each pulldown sample was run just into the separation gel so that the whole bands could be excised as one sample and subjected to in-gel trypsin digestion and MS analysis.

4.5.6 Mass spectrometry (MS) analysis

The mass spectrometry was performed as described previously [284, 315]. Briefly, the excised gel bands described above were cut into approximately 1-mm^3 pieces. The gel pieces were then subjected to in-gel trypsin digestion [256] and dried. Samples were reconstituted in 5 μL of high-performance liquid chromatography (HPLC) solvent A (2.5% acetonitrile, 0.1% formic acid). A nanoscale reverse-phase HPLC capillary column was created by packing 5-μm C18 spherical silica beads into a fused silica capillary (100 Âţm inner diameter ÃŮ ∼20 cm length) with a flame-drawn tip. After the column was equilibrated, each sample was loaded onto the column via a Famos autosampler (LC Packings). A gradient was formed, and peptides were eluted with increasing concentrations of solvent B (97.5% acetonitrile, 0.1% formic acid).

As the peptides eluted, they were subjected to electrospray ionization and then entered into an LTQ-Velos mass spectrometer (Thermo Fisher Scientific). The peptides were detected, isolated, and fragmented to produce a tandem mass spectrum of specific fragment ions for each peptide. Peptide sequences (and hence protein identity) were determined by matching protein databases with the fragmentation pattern acquired by the software program SEQUEST (ver. 28) (Thermo Fisher Scientific). Enzyme specificity was set to partially tryptic with two missed cleavages. Modifications included carboxyamidomethyl (cysteine, fixed) and oxidation (methionine, variable). Mass tolerance was set to 0.5 Da for precursor ions and fragment ions. The database searched was UniProt. Spectral matches were filtered to contain a false discovery rate of less than 1% at the peptide level using the target-decoy

method [81], and the protein inference was considered followed the general rules [197], with manual annotation based on experiences applied when necessary. This same principle was used for isoforms when they were present in the database. The longest isoform was reported as the match.

4.5.7 Bioinformatic analysis

The full-length YAP, TAZ, SAV1 and KIBRA dataset was retrieved from a previous study [315]. The TAP-MS dataset for a group of full-length WW domain-containing proteins randomly selected from human proteome and the WW domains isolated from these proteins as well as the four Hippo pathway WW components (YAP, TAZ, SAV1 and KIBRA) were newly generated in this study. We combined these two datasets and assigned quality scores to the identified protein-protein interactions using MUSE algorithm as previously described [153], where a group of unrelated TAP-MS experiments (1,806 experiments using stably expressed TAP-tagged protein baits and 20 experiments using empty vector baits) were included as a control group. Through it, we considered any interaction with a MUSE score of at least 0.9 and raw spectra count greater than 1 to be a high-confident interacting protein (HCIP). The overall HCIP reproducibility rate was close to 85%, which increased when the cutoff peptide number increased. The full-length WW domain-containing proteins and their corresponded WW domains shared 47.5% HCIPs and only 10.2% overlapped HCIPs were identified for the WW domains isolated from the Hippo WW domain-containing components and the control ones (Figure C.1C).

The WW domain-containing proteins' interactomes were enriched in signaling pathways, biological processes and diseases using the HCIPs identified in our studies. The P values were estimated using the Knowledge Base provided by Ingenuity Pathway software (Ingenuity Systems, www.ingenuity.com), which contains findings and annotations from multiple sources

including the Gene Ontology database, KEGG pathway database, and Panther pathway database. Only statistically significant correlations ($P < 0.05$) are shown. The -log (P value) for each function and related HCIPs is listed.

4.5.8 Data availability

The MS proteomic data have been deposited in the ProteomeXchange Consortium database (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [302] with the dataset identifier PXD004649. The detailed project information is as follows:
Project Name: Human WW-domain containing proteins TAP-LC-MSMS
Project accession: PXD004649
Project DOI: 10.6019/PXD004649
Reviewer account username: reviewer38029@ebi.ac.uk
Password: eavjPdCz

4.5.9 Screen of human WW domain-containing proteins using the identified Hippo WW domain binding criterion

All the WW domain-containing proteins were retrieved from human proteome using a Simple Modular Architecture Research Tool (SMART) (http://smart.embl-heidelberg.de) and the WW domain-containing protein list was further refined in Uniprot (https://www.uniprot.org). Based on the definition, the WW domain-containing proteins are defaulted with two tryptophan (W) residues as separated by 20-22 amino acids within the sequence. All the WW domain sequences were downloaded from Uniprot and subjected to scan with the identified 9-amino acid sequence manually. The list of all the human WW domain-containing proteins and the searching result are listed in Table EV6.

4.5.10 Gene inactivation by CRISPR/Cas9 system

To generate the STXBP4 knockout cells, five distinct single-guide RNAs (sgRNA) were designed by CHOPCHOP website (https://chopchop.rc.fas.harvard.edu), cloned into lentiGuide-Puro vector (Addgene plasmid # 52963), and transfected into HEK293A cells with lentiCas9-Blast construct (Addgene plasmid # 52962). The next day, cells were selected with puromycin (2 μg/ml) for two days and subcloned to form single colonies. Knockout cell clones were screened by Western blotting to verify the loss of STXBP4 expression and their genomic editing was further confirmed by sequencing (Appendix Figure C.10).

The sequence information for sgRNAs used for STXBP4 knockout cell generation is as follows:

STXBP4_sgRNA1: AGACTTAATGTTGAGGCTTG;

STXBP4_sgRNA2: GGCTTGGTGTTGTTCCTTTG;

STXBP4_sgRNA3: TGCTTTCACCAAAGTAGCCT;

STXBP4_sgRNA4: GGAAACAGGCCTTGGCCTGA;

STXBP4_sgRNA5: AGGTACTAGGAGGAATTAAC.

4.5.11 RNA extraction, reverse transcription and real-time PCR

RNA samples were extracted with TRIzol reagent (Invitrogen). Reverse transcription assay was performed using the Script Reverse Transcription Supermix Kit (Bio-Rad) according to the manufacturer's instructions. Real-time PCR was performed using Power SYBR Green PCR master mix (Applied Biosystems). For quantification of gene expression, the $2^{\Delta\Delta Ct}$ method was used. *GAPDH* expression was used for normalization.

The sequence information for each primer used for gene expression analysis is as follows:

CTGF-Forward: 5'-CCAATGACAACGCCTCCTG-3';

CTGF-Reverse: 5'-GAGCTTTCTGGCTGCACCA-3';

CYR61-Forward: 5'-AGCCTCGCATCCTATACAACC-3';

CYR61-Reverse: 5'-GAGTGCCGCCTTGTGAAAGAA-3';

ANKRD1-Forward: 5'-CACTTCTAGCCCACCCTGTGA-3';

ANKRD1-Reverse: 5'-CCACAGGTTCCGTAATGATTT-3'.

4.5.12 Molecular dynamics simulations

All simulations were conducted using the AMBER18 molecular dynamics suite [46, 45, 243]. Initial parameterization of complexes and apo conformations was conducted with the LeAP module in AMBER18, using the protein force field ff14SB [175]. YAP-WW1 domain bound to SMAD7-PY motif-containing peptide was initially parameterized using the PDB structure, 2LTW. The SMAD7-PY motif-containing peptide structure was removed from 2LTW and docked to the STXBP4-WW domain (PDB: 2YSG) to form a complex (STXBP4-WW/SMAD7-PY). In the N-terminal sequence of STXBP4, four non-native residues (GSSG) were removed prior to docking and formation of the complex STXBP4-WW/SMAD7-PY to maintain consistent residue number with the YAP-WW1 domain. To generate the mutant complexes, all the conserved residues from 2LTW were mutated into alanine using Modeller v9.21 [88, 181, 175, 270], and initial docked poses between mutated YAP-WW1 domains and SMAD7 were generated using the HADDOCK docking program [53, 70] prior to simulation (Appendix Table C.1). This docking procedure was also repeated for the APBB3-WW/SMAD7-PY simulations. An apo form of SMAD7-PY and YAP-WW1 (wild-type domain mutants: L173A/P174A, G176A, W177A, E178A, Y188A, F189A, H192A, T197A, W199A, P202A) were also derived from PDB structure 2LTW, for simulations (Appendix Table C.1).

Neutralized with either Na^+ or Cl^- counter ions, systems were solvated using a 10 Åbuffer of

TIP3P waters in a truncated octahedron box. All complexes and apo forms were minimized in a two-step process using the PMEMD program to remove any steric clashes and overlaps. Complexes were heated to 300K for 100 ps in the canonical (NVT) ensemble and equilibrated for 10 ns at 300K in the isothermal-isobaric (NPT) ensemble. Production runs were generated using the accelerated CUDA version of PMEMD [243] in the NVT ensemble with 2-fs time steps at 300K, until MM/PBSA calculations converged. Appendix Table C.1 outlines the complete simulation conditions for each complex and apo structure.

The MM/PBSA module in AMBER18 [29, 39, 191, 65, 307, 309, 240, 164] was employed to calculate the binding free energies (ΔG) of wild-type and mutant complexes. Calculations do not take into consideration entropy; however, all complexes retain SMAD7-PY as a common binder meeting the necessary requirements for MM/PBSA calculation and comparison. Convergence of both YAP-WW1 and STXBP4-WW complex simulations was determined via cumulative average calculations of ΔG values and timeframes for all subsequent analyses (e.g. clustering, averaging, RMSD, etc.) of each complex were determined based on this metric.

Utilizing the AMBER post-processing program (CPPTRAJ) [235] module in the AMBER18 package, clustering was performed for each complex using only the Cα atoms in SMAD7-PY motif-containing peptide. We chose to cluster using SMAD7-PY motif-containing peptide that coordinates upon observation of the relative stability of both wild-type YAP-WW1 and STXBP4-WW domains. For wild-type complexes (YAP-WW1 or STXBP4-WW bound to SMAD7-PY), all frames were incorporated to generate representative clusters, and only the top 5 clusters are displayed (Appendix Figure C.9C). Conformations were clustered using the hierarchical agglomerative clustering algorithm (average-linkage), with 2.33 Åcriteria set as the minimum distance between clusters. Average structures were calculated from only converged timeframes indicated in Appendix Table C.1. Using only Cα atoms, the conformation with the smallest RMSD to the average structure was used to represent the

average conformation (Figures C.3A, C.3E and 3D). Hydrogen bonds were quantified using the Baker-Hubbard [15] criteria and the MDTraj [84] python module. Ionic salt bridge interactions were determined with a distance criterion [16] (6 Å) between centers of charged groups (positively charged atoms from basic residues Arg, Lys, His: NH*, NZ*, NE2; regions of partial positive charge from His: NE2, HE*, CE1, HD2; negatively charged atoms from acidic residues Glu and Asp: OE*, OD*). Hydrophobic interactions were also measured via a distance criterion of 3.9 Åbetween carbon atoms. Initially identified in WT YAP-WW1/SMAD7-PY simulations, four intermolecular residue pairs (P208-W199, P209-T197, Y211-H192, P209-Y188) and their Cα atoms were used to calculate the average distance (AD) values in frames outlined in the simulation conditions table (Figure C.3F and Appendix Table C.1). This AD calculation procedure was repeated for all complex simulations (SMAD7-PY bound to YAP-WW1 mutants, STXBP4-WW, and APBB3-WW), with Cα atoms of residues in equivalent positions of YAP-WW1 residues.

4.5.13 Xenograft Assays

Athymic nude (nu/nu) mouse strain was used for the xenograft tumor assay in this study. Four-week-old female nude mice were purchased from Jackson Laboratory (002019) and kept in a pathogen-free environment. The xenograft tumor experiments were followed institutional guidelines, approved by the Institutional Animal Care and Use Committee of the University of California, Irvine, and performed under veterinary supervision. The indicated 786-O cells ($2x10^6$) were subcutaneously injected into the nude mice. After 60 days' adaptation, mice were euthanized, and tumor weights were analyzed.

4.5.14 Immunohistochemical analysis

The kidney tissue array (BC07115a) was purchased from US Biomax, Inc. According to the Declaration of Specimen Collection provided by US Biomax, each specimen collected from any clinic was consented by both hospital and individual.

The kidney tissue array was deparaffinized and rehydrated. The antigens were retrieved by applying Unmask Solution (Vector Laboratories) in a steamer for 40 min. To block endogenous peroxidase activity, the sections were treated with 3% hydrogen peroxide for 30 min. After 1 hour of pre-incubation in 10% goat serum to prevent non-specific staining, the samples were incubated with an antibody at 4°C overnight. The sections were incubated with SignalStain Boost detection reagent at room temperature for 30 min. Color was developed with SignalStain 3,3'-diaminobenzidine chromogen-diluted solution (all reagents were obtained from Cell Signaling Technology). Sections were counterstained with Mayer hematoxylin. To quantify the results, a total score of protein expression was calculated from both the percentage of immunopositive cells and immunostaining intensity. High and low protein expressions were defined using the mean score of all samples as a cutoff point. Pearson chi-square analysis test was used for statistical analysis of the correlation of STXBP4 with tissue type (normal versus cancer) and the correlation between STXBP4 and YAP.

4.5.15 TCGA database analysis

Dataset for STXBP4 was downloaded from the Cancer Genome Atlas (TCGA) data portal (https://portal.gdc.cancer.gov/). The mRNA expression and clinical data of STXBP4 were analyzed in TCGA-KIRC project. The mRNA levels of STXBP4 was categorized into high and low expression groups based on the median value. The correlation between STXBP4 expression and patient survival rate was analyzed. Total 611 patient samples were analyzed.

4.5.16 Quantification and statistical analysis

Each experiment was repeated twice or more, unless otherwise noted. There were no samples or animals excluded for the analyses in this study. As for the mouse experiments, there was no statistical method used to predetermine sample size. We assigned the animals randomly to different groups. A laboratory technician was blinded to the group allocation and tumor collections during the animal experiments as well as the data analyses. The Student's t-test was used to analyze the differences between groups. Data were analyzed by Student's t-test or Pearson chi-square analysis. SD was used for error estimation. A P value < 0.05 was considered statistically significant.

4.5.17 Author contributions

W.W. conceived and supervised the study. R. L. designed and supervised the simulation analyses. R.V., H.H., A.P.T., S.Z., B.Y., G.S., S.O., K.C. and W.W. performed the experiments. V.T.D., A.E.A and R.L. performed the simulation analyses. Y.C. performed TCGA data and evolutionary analyses. J.C., X.L. and W.W. performed the proteomic and bioinformatic analyses. O.R. provided key reagents and revised the manuscript. V.T.D., J.C., X.L., R.L. and W.W. wrote the manuscript.

4.5.18 Acknowledgments

We thank Drs. Steven Gygi and Ross Tomaino (Taplin Mass Spectrometry Facility, Harvard Medical School) for help with the mass spectrometry analysis and Dr. Chao Wang (MD Anderson Cancer Center) for the insightful discussion. This work was supported in part by a NIH grant (GM126048), an American Cancer Society Research Scholar grant (RSG-18-009-01-CCG), and an Anti-Cancer Challenge pilot project from the Chao Family Comprehensive

Cancer Center (P30 CA062203) to W.W.; a NIH grant to R. L. (GM130367); and a Department of Defense Era of Hope Research Scholar Award to J.C. (W81XWH-09-1-0409). R.V. is supported by a NIH Initiative for Maximizing Student Development (IMSD) Fellowship (GM055246). V.T.D is supported by a Mathematical, Computational and Systems Biology Predoctoral NIH Training Grant (T32 EB009418-08).

Chapter 5

Computational Studies of Intrinsically Disordered Proteins

5.1 Introduction

As structural data accumulates at an ever increasingly fast pace, intrinsically disordered proteins (IDPs) have garnered widespread acknowledgment for their ubiquitous presence in biochemical pathways vital to eukaryotic systems. Although the exact correlation between disordered protein regions and function remains elusive, IDPs or proteins containing both structured and intrinsically disordered regions (IDRs) have been experimentally shown to participate in DNA binding, transcription, translation, cell signaling, and the overall regulation of the cell cycle [90, 123, 158, 266, 322, 327]. Mutations in IDPs/IDRs or expression pathways of IDPs/IDRs have been implicated in various neurological disorders, cancers, and other disease-related condition [13, 97, 297]. These proteins also vary considerably in

behavior, occupying a fully disordered state, exhibiting folding only upon binding (known as coupled folding and binding)[266], or existing in mixed states of structured/unstructured regions. Experimental methods to characterize IDPs and elucidate structure-function associations can therefore be arduous and challenging. To explore the dynamic structures of IDPs, computational methods can provide the expansive sampling to complement experimental measurements.

Widely used to simulate globular proteins, generic protein force fields (e.g. ff14SB[175] and CHARMM36[24]) have been shown to disagree with experimental observables due to biases towards structured motifs [23]. Improvements to address this bias have resulted in multiple IDP-specific force fields (CHARMM36m[117], ff99IDPs[331], ff14IDPs[263], CHARMM36IDPSFF[157]) to replicate the disordered characteristics of IDPs. The ff14IDPs force field developed by Song et al.[263] included dihedral energy corrections for only eight disorder-promoting residues (A, Q, G, P, R, K, S, E) [75, 237, 324]. Although this resulted in improved IDP sampling, several inconsistencies with experimental observables arose due to the limited number of residues corrected [263]. In 2017, Song et al.[262] extended their optimization of dihedral energy terms using grid-based energy correction maps[170, 171, 172] to all 20 amino acids resulting in the ff14IDPSFF force field. This new force field simulated chemical shift values in closer agreement with experimental values [262].

Thus, our first goal of this computational study of disordered proteins is to assess the quality of both the generic protein force field (ff14SB[175]) and its IDP-specific counterpart (ff14IDPSFF[262]). However, it is notoriously difficult to obtain adequate conformational sampling for IDPs/IDRs due to the lack of one or few dominant conformations. Since microsecond timescales and multiple independent trajectories may be required, our second goal of this study is to assess the extent of sampling that is needed for quantitative structural annotation of IDPs/IDRs and to explore how to assess the sampling convergence. Here, nine short IDP peptides of the motif EGAAXAASS (X = D, E, Q, W, Y, P, L, H, K)[64, 152]

and the RNA-binding protein, HIV-1 Rev (Rev)[18, 48, 68, 176, 178, 277] were chosen as test cases to assess the quality of MD simulations with the two Amber protein force fields. The EGAAXAASS short peptides were thoroughly characterized experimentally and were found to exhibit a combination of disordered behavior and local interactions between the 5X substituted residues and adjacent neutral alanine residues [64, 152]. The longer and more complex Rev protein is a more challenging and realistic system for assessment of sampling techniques and accuracy of the tested force fields. Composed of highly charged residues (10 arginines out of 23 residues), the Rev protein is a vital component in the regulation of the HIV-1 replication cycle [68, 176, 178]. Despite its short sequence the Rev protein has been shown to adopt a diverse array of conformations (α-helices, disordered, beta) and simultaneously bind to target proteins or RNA-substrates with high affinity [176, 178, 260, 278]. Once bound to its target, it was found to adopt a very stable conformation, providing a very interesting system to probe the binding-induced folding process.

By tackling issues of force field accuracy and sampling convergence, force field advancements in the realm of IDPs can be highly informative, revealing behaviors otherwise experimentally inaccessible or providing details potentially useful in guiding experimental studies. After careful analysis of the simulation sampling convergence and force field accuracy, we further analyzed the diverse conformational preferences of the Rev protein in both the apo and bound state to complete the computational analysis of this important protein.

5.2 Methods

5.2.1 Force Fields Tested

In this study, two Amber protein force fields (ff14SB and ff14IDPSFF) were tested to assess their quality in reproducing IDP structural properties. In the generic protein force

field ff14SB[175], dihedral modifications and validation relied primarily on comparison to crystal structures exhibiting ordered secondary structures. To address the limitations of increased structured propensity propagated by the ff14SB force field, the IDP-specific force field ff14IDPSFF was developed to address the deficiency of generic protein force fields by modification of the main-chain dihedral terms [262]. The ff14IDPSFF force field is the most recently developed AMBER IDP-specific force field, improved upon from older versions [263, 331]. Song et al.[262] provided the CMAP (grid-based energy correction map) parameters for ff14IDPSFF and a utility perl script to revise ff14SB-parameterized topology files into ff14IDPSFF topology files.

5.2.2 Molecular Dynamics Simulations

The molecular dynamics package, Amber version 16, was used to generate all trajectories [46, 45, 243, 243]. Nine short peptides with the sequence motif of EGAAAXAASS (X = D, E, Q, W, Y, P, L, H, K) were tested in this study. All 9 peptides were built in the all-trans initial conformation using the Amber LEaP module, followed by minimization with the steepest descent and conjugate gradient methods, each 500 steps. Short peptides were then simulated in the GB implicit solvent for 10 ns (time steps of 1 fs) at 450K to generate 10 random conformations per peptide per force field (Table 5.1). The randomized initial structures were solvated with explicit TIP3P waters in a truncated octahedron box, with a buffer of 10 Å(Table 5.1). Neutralization was accomplished with the addition of either Na+ or Cl- ions depending on the total charge of a peptide. All solvated structures were minimized for 20,000 steps steepest descent, heated up for 20 ps in the NVT ensemble from 0K to 298K, and were equilibrated for 20 ps in the NPT ensemble at 298K. The CUDA-accelerated PMEMD[243, 243] in Amber16 was then used to generate production trajectories in the NVT ensemble at 298K. The Langevin thermostat was used for all temperature regulation.

Force fields were also tested via simulation of a larger IDP, the HIV-1 apo Rev protein (apo Rev), by extracting the protein from its bound conformation in the crystal structure (PDB ID: 1ETF) as the initial conformation. MD preparation protocols (minimization, heating, etc.) were mostly identical to those for the nine peptides mentioned above, except that 60 random conformations per force field were generated in the GB implicit solvent. These conformations were used as the initial starting structures for two sampling strategies also outlined in Table 5.1: fifty 200ns simulations (short) and ten 1 μs simulations (long). Here we chose to simulate a total of 10 μs in the form of both short and long protocols to assess which strategy leads to faster convergence of tested NMR observables.

In addition to the apo Rev simulations, we also simulated the HIV-1 Rev protein bound to its RNA-binding partner Rev responsive element (RRE). Beginning with the full NMR solution structure (PDB: 1ETF), we repeated MD simulation protocol as mentioned previously, except that only five production trajectories of 200 ns each were collected.

Table 5.1: Summary of simulation setups.

Short peptide	Citations,BMRB,PDB	Force fields	Simulation number	Length per simulation	Ions	Waters
EGAADAASS	[64]	ff14SB	10	1 μs	1 Na+	1532-2178
		ff14IDPSFF	10	1 μs	1 Na+	1465-2569
EGAAEAASS	[64]	ff14SB	10	1 μs	1 Na+	1628-2622
		ff14IDPSFF	10	1 μs	1 Na+	1464-3151
EGAAQAASS	[64]	ff14SB	10	1 μs	1 Na+	1299-2752
		ff14IDPSFF	10	1 μs	1 Na+	1520-3668
EGAAWAASS	[64, 152]	ff14SB	10	1 μs	0	1574-2637
		ff14IDPSFF	10	1 μs	0	1876-3092
EGAAYAASS	[64]	ff14SB	10	1 μs	0	1804-2867
		ff14IDPSFF	10	1 μs	0	1888-3141
EGAALAASS	[64]	ff14SB	10	1 μs	0	1373-3224
		ff14IDPSFF	10	1 μs	0	1606-3131
EGAAPAASS	[64]	ff14SB	10	1 μs	0	1751-2713
		ff14IDPSFF	10	1 μs	0	1693-2885
EGAAHAASS	[64]	ff14SB	10	1 μs	0	1498-2675
		ff14IDPSFF	10	1 μs	0	1430-3159
EGAAKAASS	[64]	ff14SB	10	1 μs	1 Cl-	1733-2434
		ff14IDPSFF	10	1 μs	1 Cl-	1633-2399
apo Rev (23 amino acids)	($\Delta\delta C\alpha$),[48]	ff14SB	10/50	1 μs / 200 ns	9 Cl-	3727-11638
	($^3J_{HNH\alpha}$),[48]	ff14IDPSFF	10/50	1 μs / 200 ns	9 Cl-	4424-13224
RRE – Rev complex	($\Delta\delta C\alpha$),[17, 18]	ff14SB	5	200 ns	53 Na+29 Cl-	10928
	($^3J_{HNH\alpha}$),[17] PDB:1ETF[18]	ff14IDPSFF	5	200 ns	53 Na+29 Cl-	10928

5.2.3 Analyses of Simulations

Post-simulation analysis incorporated a variety of software to extract observables for comparison with experiment. NMR observables – chemical shift and $^3J_{HNH\alpha}$-coupling values – were calculated to validate the performance of both tested force fields and assess the quality of MD sampling. The Amber module, cpptraj[235], was used to remove solvent for subsequent frame-by-frame processing and analysis. All chemical shift values were calculated using the SPARTA+ package [255]. $\Delta^3J_{HNH\alpha}$-coupling constants were calculated using the Karplus equation that was programmed with the MDTraj python library[187] and coefficients from literature[303]. Experimental values (Figures 5.10C-D, 5.11B) were extracted from published figures in respective papers if raw data were not available from the authors (Table 5.1).

Time-dependent cumulative averages of both NMR observables were calculated for convergence assessment. From these cumulative average calculations, the rate of change per NMR observable (ΔNMR Observable) was calculated to assess its rate of convergence. Rate of change datasets were fitted to a biphasic exponential-decay model:

$$\Delta\text{NMR Observable} = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$$

Of the fitted parameters, the slower τ_2 values were calculated and utilized to assess the rate of convergence of the observable. Kernel density estimations (KDEs) were used to analyze the detailed distribution of each predicted observable per frame. KDE's were calculated using the python packages Scikit-Learn and Seaborn [121, 211]. Epanechnikov kernels were adopted with appropriate bandwidths (h=0.5) in KDEs [82]. Initial bandwidths were determined using Scikit-Learn's grid search and cross validation function (GridSearchCV) (h=0.1) and further rescaled to h=0.5 as it yields comparable distributions with less noise.

Secondary structure propensity estimates were calculated using the DSSP program [131]. Prior to clustering, frames were pre-sorted using DSSP secondary structure assignments.

Since DSSP default settings assign residues with three basic secondary structure assignments – H (α-helix, 3_{10}-helix, π-helix), E (beta ladder, isolated beta-bridge residues), C (hydrogen bond turn, bend, loops, irregular residues) – frames were first grouped into the following categories if they contained at least one of the 3 assignments: H only, E only, C only, EH only, CH only, CE only, CEH only. Frames for all simulations fell into only four of the categories: C only, CH only, CE only, and CEH only. Clustering was then restricted to a single secondary structure category (e.g. C only). This pre-clustering assortment permits filtering based on secondary structure and increases accuracy in the clustering step.

After pre-clustering, ϕ and ψ torsion angles were extracted from trajectories with the MDTraj[187] module as input in our clustering methodology. Torsional data was then subjected to PCA dimensionality reduction with settings specified to retain 99% of variation in torsion angle data. Clustering was performed by generating gaussian mixture models (GMM)[72] for each secondary structure category (e.g. C only), in which each frame was clustered depending on its likelihood of occupying a specific component/cluster. GMMs consist of a mixture of multi-dimensional gaussian probability distributions from which the number of components/mixtures (number of "clusters") can be estimated using cross-validation techniques such as Bayesian information criterion (BIC) [247]. The lowest BIC value was used to estimate the appropriate number of mixtures for each GMM model (Figure D.14). GMMs were created using the Scikit-Learn[211] python module and implemented using the expectation-maximization algorithm[69] to fit and achieve converged mixtures/clusters.

In RRE-bound Rev (RRE-Rev) simulations, the snapshot closest to the average was used as a representative of the average structure and implemented using cpptraj [235]. Hydrogen bond occupancies were calculated using the Baker-Hubbard[15] criteria from the MDTraj[187] python module and ionic salt bridge interactions were determined with a strict distance criterion[16] (4 Å) between centers of charged groups (positively charged atoms from residues Arg and Lys: NH*, NZ*; negatively charged atoms: OP* phosphate backbone atoms in the

RNA-binding partner RRE). Pymol was used to generate the representative structural image and TOC image.

5.3 Results and Discussion

Nine short peptides, EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) and the structurally dynamic apo Rev protein from type-1 HIV were simulated to illustrate the issues that must be addressed in computational studies of IDPs, namely both the accuracy of force fields and convergence of sampling. In the following, the convergence issue of the sampling is addressed before studying the quality of the two selected force fields in reproducing NMR observables. Finally, the structural characteristics of both disordered and ordered apo Rev protein are discussed based on the expansive MD simulations in explicit solvent.

5.3.1 Convergence Analysis

Previous studies of IDPs relied on backbone RMSD analysis and/or clustering of MD trajectories within hundred nanosecond timescales to confirm proper sampling and convergence of IDPs [48, 262]. In this study, we relied on direct analysis of time-dependent cumulative averages of specific NMR observables, a reasonable technique to investigate the convergence of simulated observables.

We analyzed time-dependent cumulative averages (Figure D.1-D.5) of simulated secondary chemical shifts and $^{3}J_{HNH\alpha}$-coupling constants to estimate the time scales at which the rates of change of the observables go to zero, an indication that convergence is achieved. A convergence decay was fitted to a biphasic exponential decay model (ΔNMR Observable = $A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$) thereby allowing for the determination of τ_2. Here, the parameter generated from the first rapid decay phase, τ_1 is discarded. The implementation of this

Table 5.2: Average τ_2 values ($\Delta\delta$Cα and $^3J_{HNH\alpha}$-coupling constants) of 9-residue EGAAX-AASS with standard deviations (SDs)

	Avg. τ_2 ± SD ($\Delta\delta$Cα (ns))		**Avg. τ_2 ± SD ($^3J_{HNH\alpha}$ (ns))**	
Protein	**ff14SB**	**ff14IDPSFF**	**ff14SB**	**ff14IDPSFF**
EGAADAASS	705 ± 134	221 ± 22	679 ± 242	761 ± 187
EGAAEAASS	389 ± 63	195 ± 25	639 ± 179	715 ± 179
EGAAHAASS	561 ± 104	508 ± 107	686 ± 193	786 ± 279
EGAAKAASS	412 ± 68	163 ± 21	570 ± 130	685 ± 183
EGAALAASS	307 ± 50	239 ± 36	692 ± 163	710 ± 185
EGAAPAASS	247 ± 31	270 ± 40	716 ± 205	581 ± 181
EGAAQAASS	435 ± 68	437 ± 74	747 ± 225	689 ± 154
EGAAWAASS	423 ± 60	343 ± 40	631 ± 113	525 ± 136
EGAAYAASS	511 ± 93	480 ± 77	641 ± 173	687 ± 250

technique allows us to quantitatively assess and compare the convergence rates of tested systems and sampling protocols.

Short Peptides Table 5.2 summarizes the average τ_2 values – derived from simulated $\Delta\delta$Cα – of the 9 short peptides. These values are further represented in boxplots detailing their ranges, medians, and lower/upper quartiles (Figure 5.1). Detailed fitting plots for all residues and simulation types are shown in the SI file (Figure D.6-D.7). Calculated average τ_2 values of EGAAXAASS simulations reveal a stark contrast between ff14SB- and ff14IDPSFF-generated simulated $\Delta\delta$Cα values, with ff14IDPSFF exhibiting lower values than the generic ff14SB force field, except the Q-substituted simulations, whose τ_2 values are quite similar between the two. The analysis suggests ff14IDPSFF simulations converge mostly faster than the ff14SB simulations for the chemical shifts monitored (Figure 5.1).

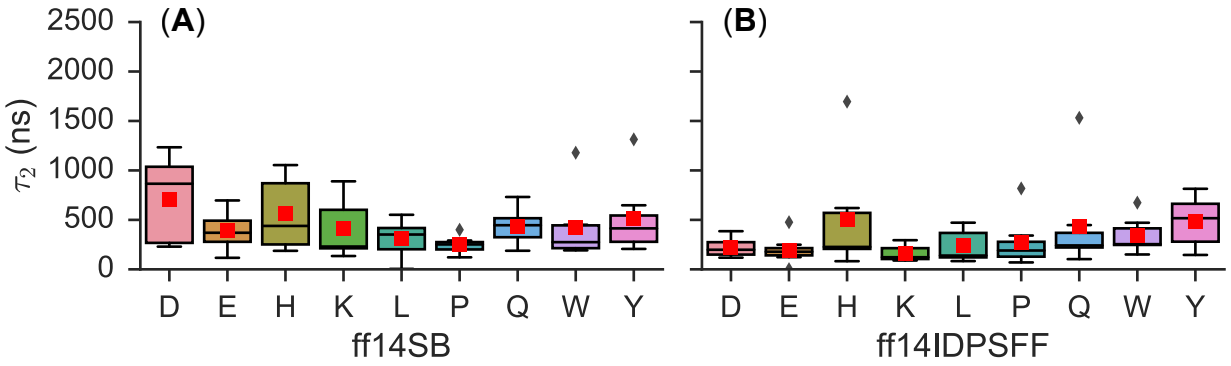


Figure 5.1: Summary of τ_2 values (medians, ranges, quartiles, outliers) for peptides of EGAAXAASS (X=D, E, H, K, L, P, Q, W, Y), derived from $\Delta\Delta\delta C\alpha$ calculations. Simulations are labeled by peptide and force field: (A) ff14SB and (B) ff14IDPSFF. Diamonds indicate outliers and a red box denotes the average τ_2 value. Fitted plots from which boxplots were derived can be found in the SI (Figure D.6-D.7).

Next, we repeated the above biphasic exponential fitting to cumulative averages of a second simulated NMR observable – $^3J_{HNH\alpha}$-coupling constants (Figure D.10-D.11). Overall, the range of calculated τ_2 values is narrow and comparable between both force fields (Figure 5.2). Upon closer inspection, the average τ_2 (indicated by red boxes) is generally higher in ff14IDPSFF simulations than those in ff14IDPSFF simulations, different from the chemical shift analysis. Interestingly, the final $^3J_{HNH\alpha}$-coupling constants are comparable between the two force fields, as the average values are within standard deviations. Peptides substituted with P, Q, or W in ff14IDPSFF simulations, exhibit lower τ_2 values in comparison to other substituted short peptides, suggesting possible conformational preferences leading to increased convergence rate. Comparison of the τ_2 values for the two NMR observables suggests that J-coupling constants in general converge slower than secondary chemical shifts in our simulations, as shown in Figures 5.1-5.2 and Table 5.2. Nevertheless, both sets of simulations are believed to be converged as far as both NMR observables are concerned, as the τ_2 values are much shorter than the cumulative simulation time scales sampled.

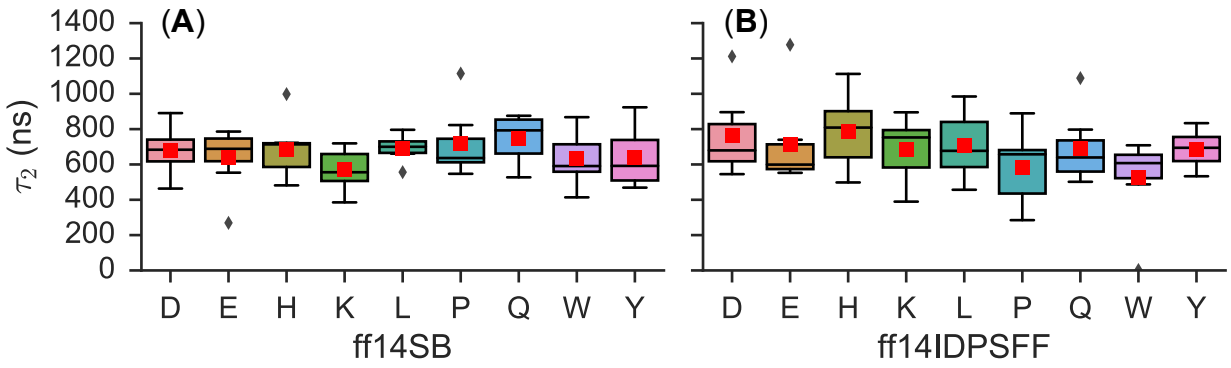


Figure 5.2: Summarization of τ_2 values (median, range, quartiles, outliers) for peptides of EGAAXAASS (X=DEHKLPQWY), derived from ${}^3J_{HNH\alpha}$-coupling constants. Diamonds indicate outliers and a red box denotes the average τ_2 value. Fitted plots from which boxplots were derived can be found in the SI (Figure D.10-D.11).

Apo Rev and RRE-Rev We extended the convergence analysis of the two tested force fields for the simulations of both apo and bound Rev. Biphasic exponential decay models were fitted (Figure D.8-D.9, D.12-D.13) as outlined in the Short Peptides subsection, using cumulative averages (Figure D.3-D.5) of simulated secondary Cα chemical shifts and ${}^3J_{HNH\alpha}$-coupling constants. A summary of τ_2 values for apo Rev in Table 5.3 reveals a consistent pattern in comparison to the short peptides: the τ_2 values for $\Delta\delta$Cα in ff14IDPSFF simulations are lower than those in ff14SB simulations and the τ_2 values of ${}^3J_{HNH\alpha}$-coupling constants in ff14IDPSFF simulations are higher than those in ff14SB simulations.

We also explored the convergence behavior of different simulation protocols in the simulations of apo Rev. Since the duration of MD simulations can significantly impact the conformational sampling, a total of 10 microseconds of MD simulation with both short (200ns x 50) and long (1μs x 10) protocols was generated for comparative analysis. Initial, qualitative inspection of cumulative averages (Figure D.3-D.4) of simulated NMR observables reveals higher fluctuations in the long protocol. Different observations in the short and long protocols suggest the two probably converged to different conformational minima, though it is clear via inspection of cumulative averages (Figure D.3-D.4) that the short protocol transitioned

Table 5.3: Average τ_2 values ($\Delta\delta$Cα and $^3J_{HNH\alpha}$-coupling constants) of apo Rev and RRE-Rev with SDs

	Avg. τ_2 ± SD ($\Delta\delta$Cα (ns))		**Avg. τ_2 ± SD ($^3J_{HNH\alpha}$ (ns))**	
Protein	**ff14SB**	**ff14IDPSFF**	**ff14SB**	**ff14IDPSFF**
Apo Rev (1μs x 10)	445 ± 75	396 ± 70	642 ± 166	710 ± 209
Apo Rev (200ns x 50)	119 ± 73	115 ± 58	422 ± 71	451 ± 67
RRE-Rev (200ns x 5)	21.8 ± 1.6	24.0 ± 2.3	3.4 ± 0.3	3.6 ± 0.3

to their minima faster.

Cumulative averages were then fitted as biphasic exponential decay models (Figure D.8, D.12, summary of fitted τ_2 in Table 5.3 and Figure 5.3). Table 5.3 and Figure 5.3 clearly show that both NMR observables converge faster in the short protocol. This is consistent with the initial qualitative inspection of apo Rev cumulative averages (Figure D.3-D.4), where it appears that the short protocol produces overall better convergence trends in all cases. The τ_2 values are also consistently distributed within narrower ranges (aka smaller SDs) in the short protocol, indicating consistent convergence of simulated NMR observables. In contrast the distributions of τ_2 values from the long protocol strongly depend on force fields and observables analyzed.

Finally convergence rates for RRE-Rev simulations in Table 5.3 also indicate comparable convergence between ff14SB and ff14IDPSFF simulations, although $^3J_{HNH\alpha}$-coupling-derived τ_2 values are much smaller than $\Delta\delta$Cα–derived τ_2, apparently due to the much more stable Rev in the bound state. Overall the convergence rate analysis shows that it is important to monitor individual observables for their convergence trends.

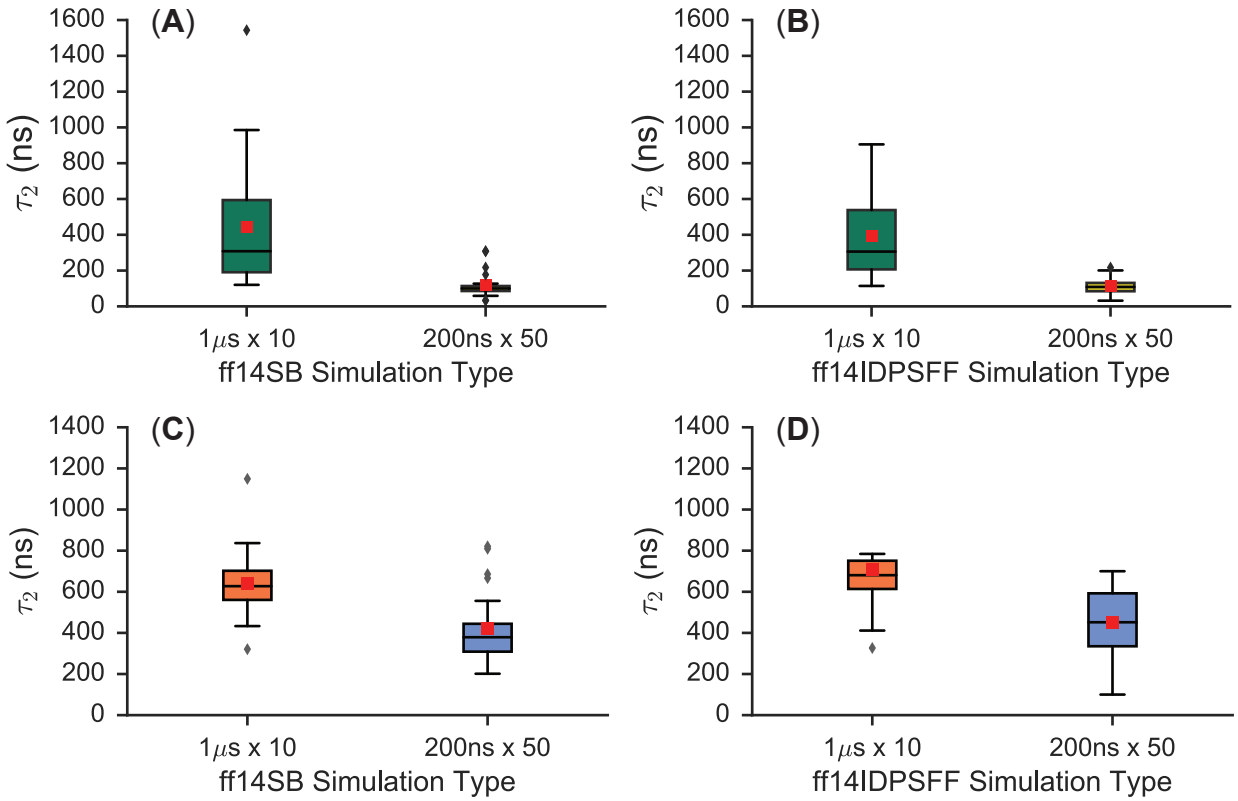


Figure 5.3: Summarization of τ_2 values derived from cumulative averages of $\Delta\delta$Cα and ${}^3J_{HNH\alpha}$-coupling constants for apo Rev. Boxplots depict median, range, quartiles, outliers, and averages (red box). (A) Details only ff14SB-parameterized simulations of $\Delta\delta$Cα-derived τ_2 values. (B) Details only ff14IDPSFF-parameterized simulations of $\Delta\delta$Cα-derived τ_2 values. (C) Only ff14SB-parameterized simulations of ${}^3J_{HNH\alpha}$-coupling-derived τ_2 values are shown. (D) Details only ff14IDPSFF-parameterized simulations of ${}^3J_{HNH\alpha}$-coupling-derived τ_2 values.

5.3.2 Distributions of Simulated Observables

We implemented the kernel density estimation (KDE) method to determine the probability density distributions of simulated NMR observables. There are two purposes in conducting this analysis. First, it provides a more detailed view of simulated observables. Second, it provides a means to cross-validate, in more detail, the different simulation protocols used in the simulations of the more challenging apo Rev.

Short Peptides Figure 5.4 shows KDE analyses for Cα secondary chemical shifts. The distribution in Figure 5.4 shows that ff14SB conformations (first/third columns) are concentrated into multiple peaks in regions characteristic of helices (3 ± 1 ppm) and random coil ($\sim$0 ppm) [265]. As an example, peptide EGAADAASS (ff14SB) exhibits multiple peaks, and a higher concentration of positive secondary Cα chemical shifts. In contrast, the ff14IDPSFF distributions (second/fourth columns) are overall narrower, more symmetrical, and more Gaussian-like centered around 0 ppm, suggesting more uniform disordered structures in the ensemble (Figure 5.4).

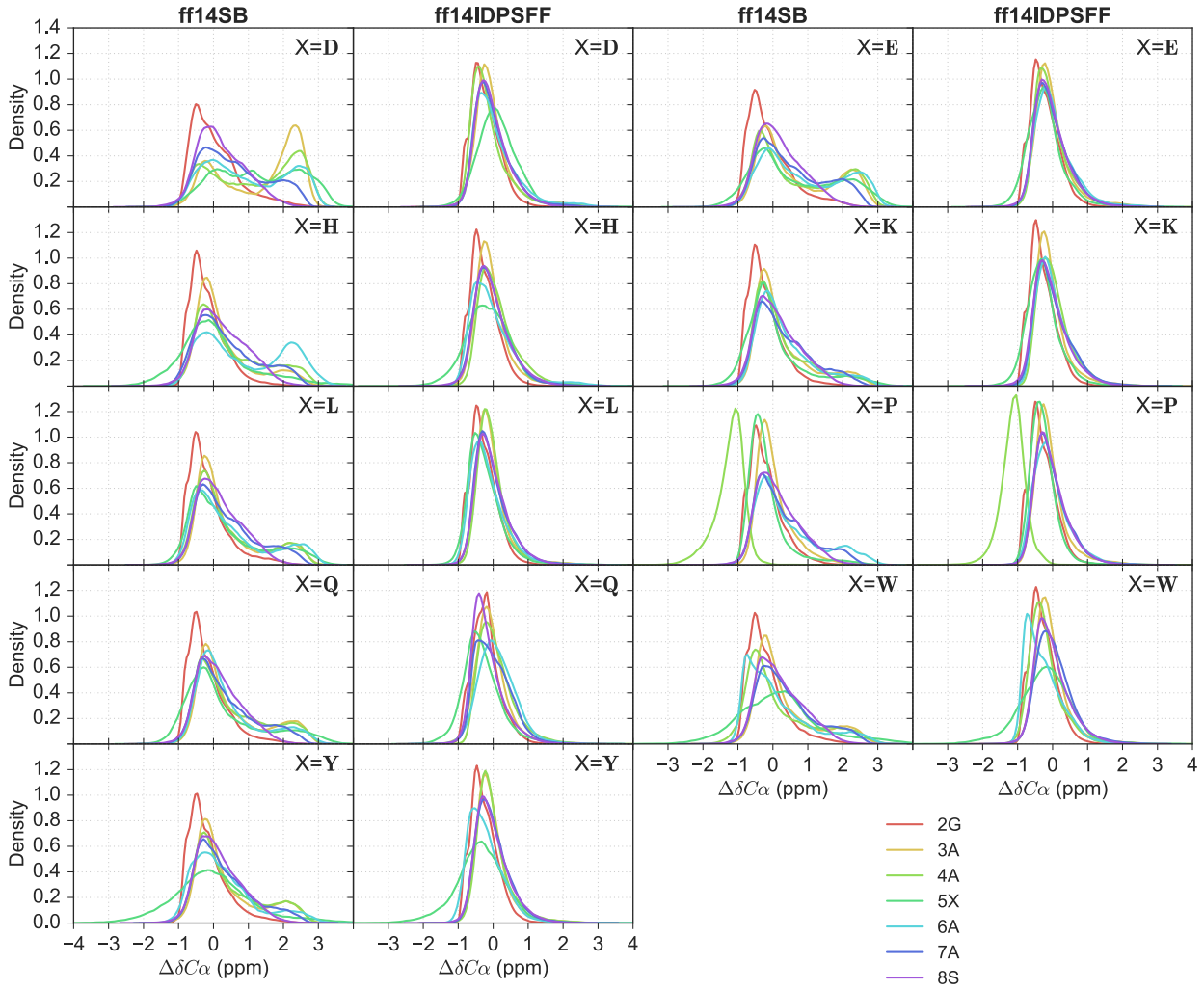


Figure 5.4: Kernel density estimations (KDEs) of secondary Cα chemical shift values for 9 short peptides of EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) and residues 2-8. Residues are colored as indicated in the legend.

KDEs of $^3J_{HNH\alpha}$ -coupling scalar coupling constants are shown in Figure 5.5. Scalar $^3J_{HNH\alpha}$-coupling constants for helical structures typically average 4.2-5.6 Hz, beta sheet conformations average 8.5-10 Hz, and random coil average 5.9-7.7 Hz [261]. In Figure 5.5, a significant proportion of residues display peaks within the helical region, from both force fields. However, distributions in ff14SB simulations display higher densities characteristic of helices than those in the ff14IDSPFF simulations for most peptides. A high concentration of peaks can also be observed in the 8.5-10 Hz range typical of beta conformations in the ff14IDPSFF simulations. However only a small fraction of conformations are within values characteristic of beta conformations in the ff14SB simulations. We supplemented the NMR observables with a more detailed secondary structure analysis based on the DSSP[131] program. The DSSP data shows, however, that beta secondary structure is nonexistent in both simulations (Figure D.17). The discrepancy is not a surprise given that the $^3J_{HNH\alpha}$-coupling constant calculation only considers the main-chain torsion angles while DSSP considers a range of different structural and energetic properties.

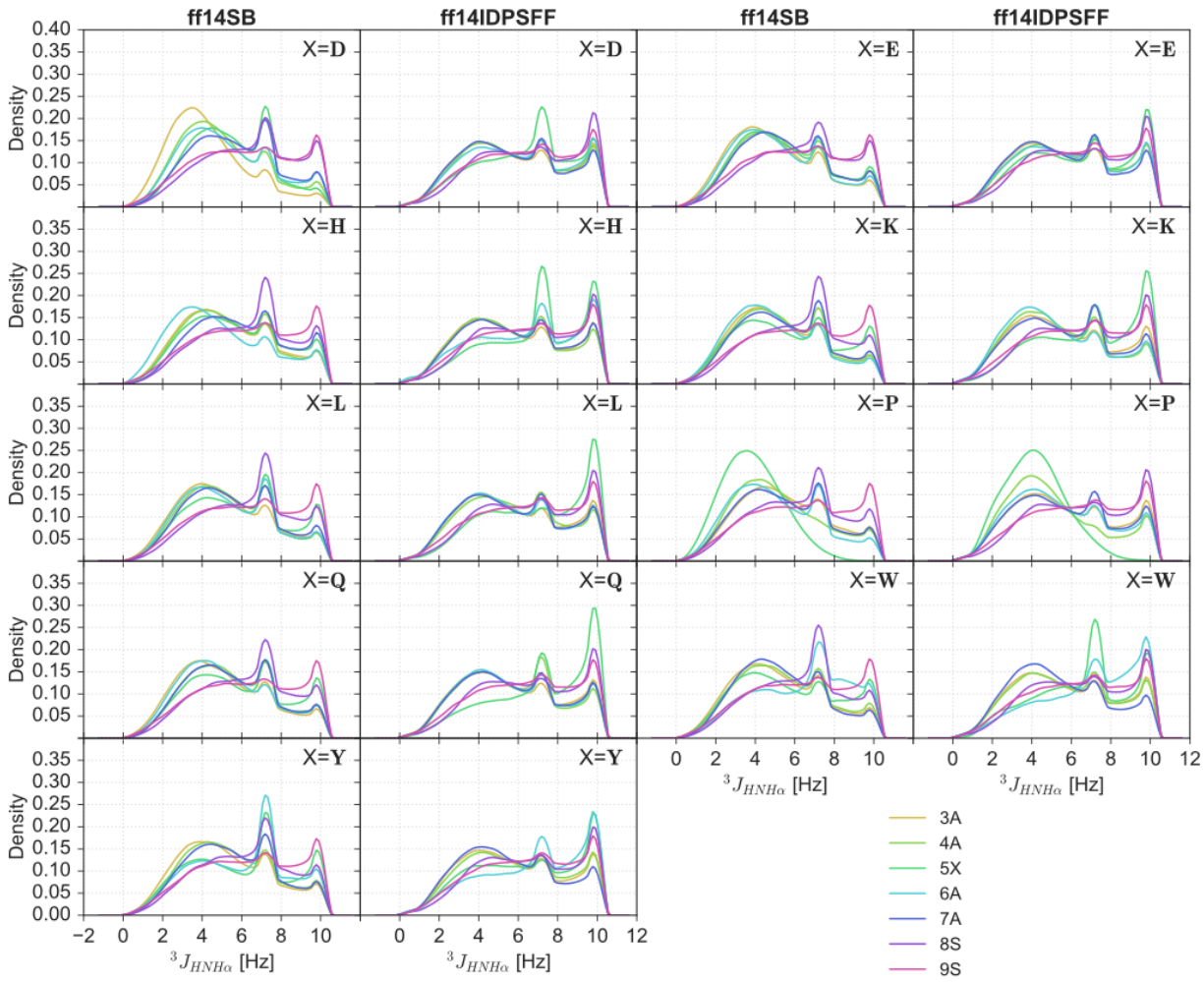


Figure 5.5: KDEs of $^3J_{HNH\alpha}$-coupling constants for 9 short peptides of EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) and residues 3-9.

Apo Rev Apo Rev simulations also display similar distributions described above – increased peak densities in the helical region in the ff14SB simulations compared to the ff14IDPSFF simulations. Juxtaposition of the two distributions displays an overall heterogeneous distribution in the ff14SB force field, with peaks in ranges typical of helical character (3 ± 1 ppm) (Figure 5.6A-B). The long-protocol simulations contain higher density peaks in the 3 ± 1 ppm range, indicating that more conformations contain helical content compared to the short-protocol simulations (Figure 5.6B). This increased helicity observed in long ff14SB simulations suggests the impact of timescales (short vs. long) is more apparent in ff14SB

simulations than ff14IDPSFF simulations. In the ff14IDPSFF simulations, both timescale types produce almost identical homogenous distributions centered ∼0 ppm (Figure 5.6C-D).

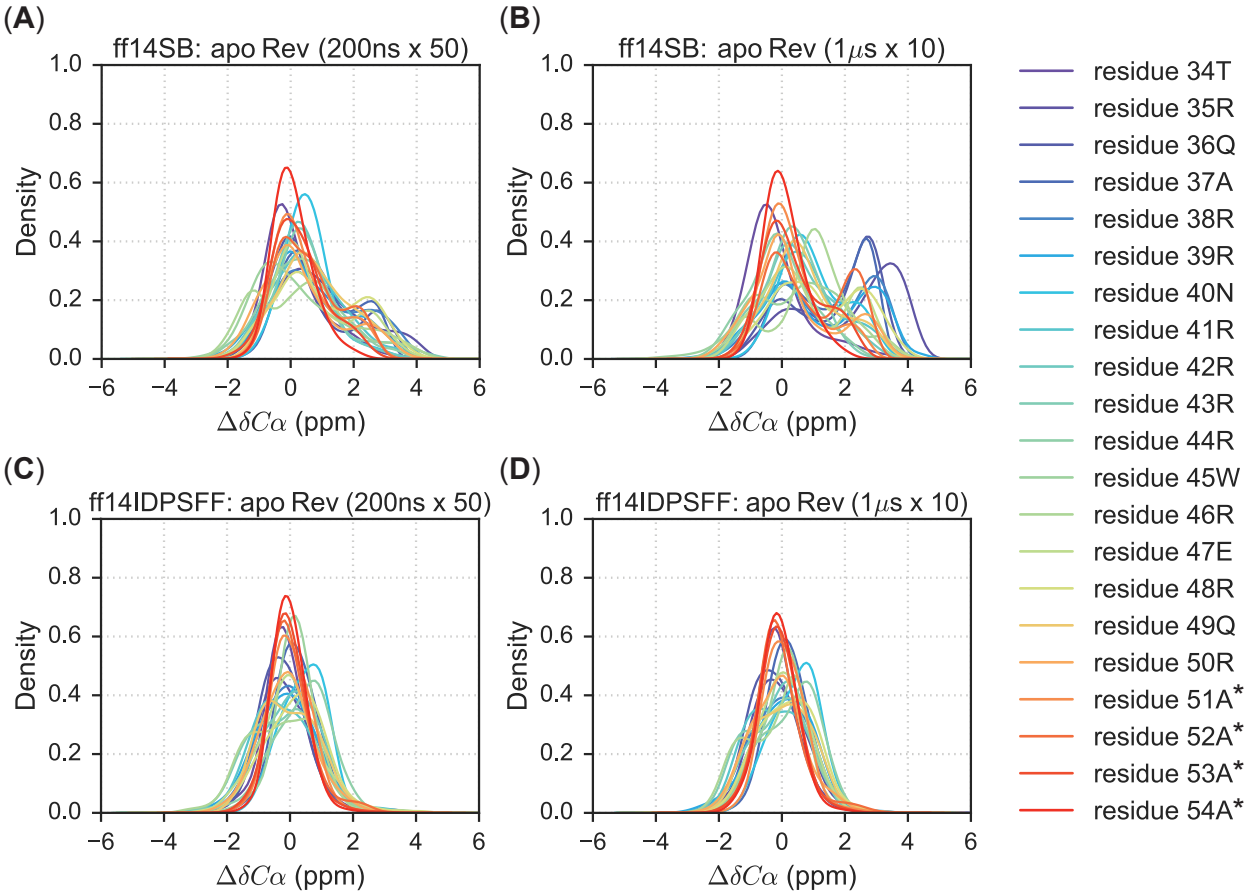


Figure 5.6: KDEs of secondary Cα chemical shift values for 1μs x 10 (long) simulations and 200ns x 50 (short) simulations. Residues are colored according to the legend and simulations are plotted according to the following combination of force field and timescale types: (A) Short simulations using the ff14SB force field. (B) Long simulations using the ff14SB force field. (C) Short simulations using the ff14IDPSFF force field. (D) Long simulations using the ff14IDPSFF force field. Asterisks (*) indicate non-native residues.

The KDE analysis was also conducted for simulated $^{3}J_{HNH\alpha}$-coupling constants. In all simulations, we observed three general regions in the KDE distributions: helical region (average 4.2-5.6 Hz), beta region (average 8.5-10 Hz), and disordered/coiled region (average values 5.9-7.7 Hz) [261]. Similar observation was also noted in experimental findings [48]. Both force fields and simulation protocols exhibit similar peaks in the helical region (broad

with densities less than 0.2), but differ in the following: 1) ff14SB simulations peaks contain higher densities, indicating more helical content than both ff14IDPSFF simulations; and 2) the long-protocol ff14SB simulations peaks are more left-shifted indicating increased helicity than its short protocol counterpart (Figure 5.7A-5.7B). In the disordered region: 1) the ff14SB simulations exhibit less disordered secondary structures as density peaks are lower than the ff14IDPSFF simulations; and 2) the peaks are similar between short and long-protocol simulations when apo Rev is modeled with ff14IDPSFF. In the beta region, density peaks in the ff14SB simulations are in general lower than those in the ff14IDPSFF simulations.

Several observations, however, are contradictory to those in the chemical-shift KDE analysis. A single peak representing residue 46R is the only density peak > 0.6 in the ff14SB simulations (long protocol), while all other peaks are $\sim$0.2 density within Figure 5.7B. The beta region is also more readily populated with high densities in the $^3J_{HNH\alpha}$-coupling distributions for all simulations whereas minimal densities were observed in the beta region (-1.48 $\pm$ 1.23 ppm)[265] in the $\Delta\delta C\alpha$ distributions for the ff14IDPSFF simulations (Figure 5.6 and 5.7). This discrepancy might result from our uses of the $^3J_{HNH\alpha}$-coupling constants to infer secondary structures as discussed in the **Short Peptide** analysis.

KDE distribution analysis of simulated NMR observables is also a useful assessment of convergence quality, supplementing the convergence rate analysis in section 5.3.1. The distribution data show that the ff14SB force field is more sensitive to simulation protocols than ff14IDPSFF. Consistently converged distributions in the ff14IDPSFF simulations allow us to use the convergence rates obtained in section 5.3.1 to compare which protocol is better. However, the rate estimations (Table 5.3 and Figure 5.3A-5.3B) show that the convergence rates between the two are quite similar, within 200ns in general, though it is clear that the short protocol converges faster than the long protocol. For ff14SB simulations, the different distributions presented here give us pause to claim that the sampling of the apo Rev

is sufficient in either protocol even if 10 microseconds worth of sampling has been collected (Figure 5.6). This indicates that enhanced sampling techniques would greatly benefit IDP simulations for systems as small as 23 amino acids such as apo Rev.

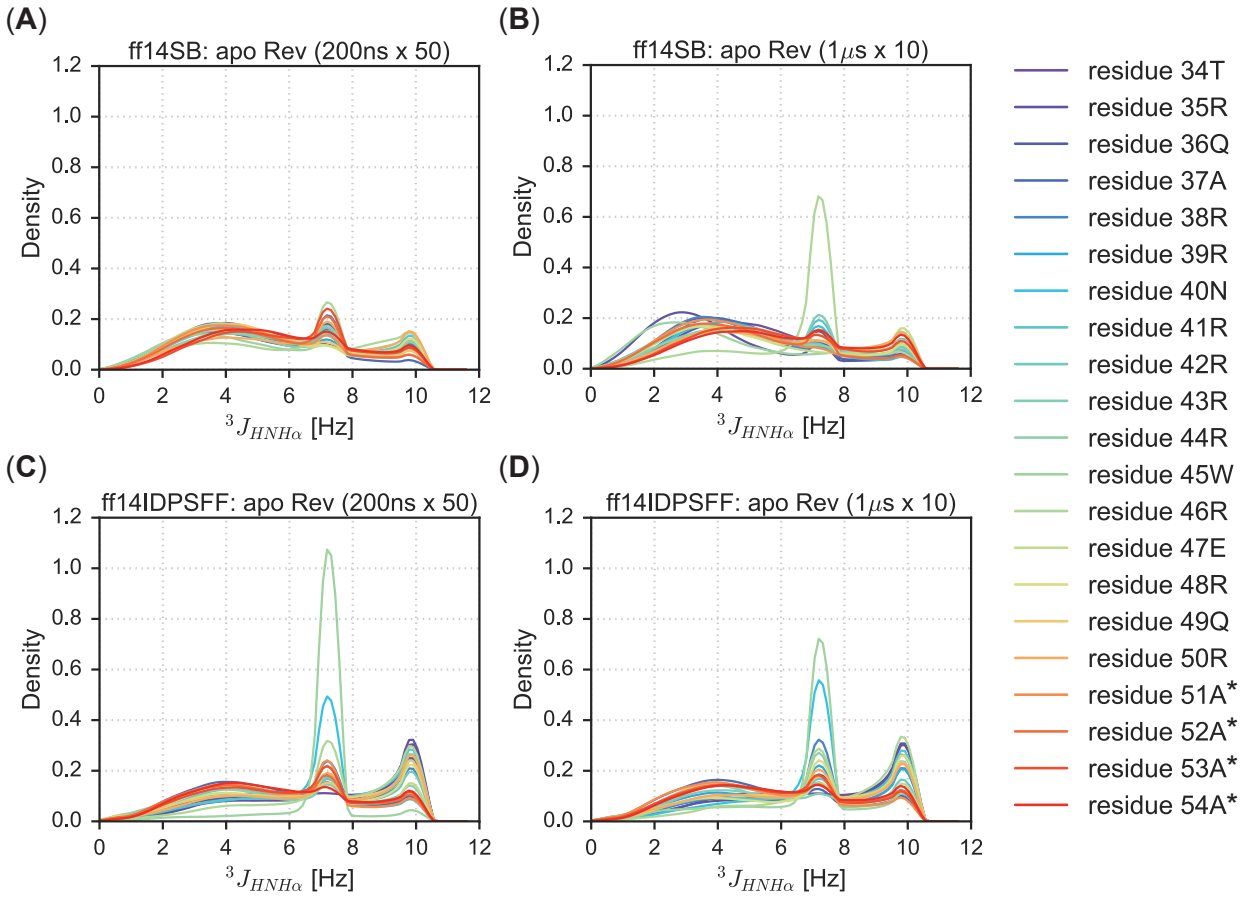


Figure 5.7: KDEs of $^3J_{HNH\alpha}$-coupling constants of short (200ns x 50) and long (1μs x 10) simulations types. Residues are colored according to the legend and simulations are plotted according to the following combination of force field and timescale types: (A) Short simulations using the ff14SB force field. (B) Long simulations using the ff14SB force field. (C) Short simulations using the ff14IDPSFF force field. (D) Long simulations using the ff14IDPSFF force field. Asterisks (*) indicate non-native residues.

5.4 Comparison of Simulated and Measured NMR Observables

Short Peptides We next calculated the final averages of secondary Cα chemical shifts for both sets of simulations and compared with experimental values (Figure 5.8). Figure 5.8 shows that experimental chemical shifts[64, 152] of the 5X-substituted residues often result in a more negative ppm shift. This suggests that the 5X-substituted residues are more disordered/extended than their adjacent residues [265]. This trend can be reproduced by both force fields, with the exception of the 5W-substituted simulations (Figure 5.8). In 5P-substituted simulation simulations, the proline residue is expected to rigidify and increase overall order in the peptide [64, 163]. Both sets of simulations agree well with experiment, replicating the expected -2 ppm shift observed for residue 4A, with ff14DIPSFF generating a slightly more negative shift (Figure 5.8). In simulations of aromatic-substituted residues (5X = W, Y), both force fields also replicate a similar observation by Dames et. al,[64] a negative -0.3 ppm shift in residue 6A. Overall, the agreement between simulation and experiment is summarized in Table 5.4, which shows improved performance of ff14IDPSFF over its generic counterpart ff14SB in modeling the tested peptides (Table 5.4, Figure 5.8).

We also compared simulated $^3J_{HNH\alpha}$-coupling constants to experimental values for these disordered peptides in Figure 5.9. Table 5.4 presents corresponding root mean square errors (RMSEs) with respect to experiment, indicating overall better agreement between experimental and ff14IDSPFF-simulated values (Table 5.4, Figure 5.9). In summary, both simulated chemical shifts and J-coupling constants demonstrates that the ff14IDPSFF simulations can better reproduce the two tested NMR observables than the ff14SB simulations in these short peptides.

Apo Rev In simulations of the more complex apo Rev, simulated secondary chemical shifts do not agree with experiment as well as those in the tested short peptides. For ff14SB sim-

Table 5.4: RMSE of calculated Cα chemical shifts and $^3J_{HNH\alpha}$-coupling constants with respect to experimental values.

	$\Delta\delta$Cα RMSE (ppm)		$^3J_{HNH\alpha}$-coupling RMSE (Hz)	
Protein	**ff14SB**	**ff14IDPSFF**	**ff14SB**	**ff14IDPSFF**
EGAADAASS	0.72	0.34	0.95	0.42
EGAAEAASS	0.54	0.2	1.01	0.61
EGAAHAASS	0.43	0.33	1.01	0.56
EGAAKAASS	0.25	0.16	0.53	0.36
EGAALAASS	0.32	0.17	0.61	0.5
EGAAPAASS	0.29	0.3	0.79	0.67
EGAAQAASS	0.36	0.18	0.88	0.57
EGAAWAASS	0.31	0.26	0.65	0.44
EGAAYAASS	0.3	0.14	0.76	0.66
Apo Rev (1μs x 10)	0.64	1.16	1.34	1.03
Apo Rev (200ns x 50)	0.68	1.19	1.17	1.02
RRE-Rev (200ns x 5)	2.35	2.62	0.9	1.08

ulations, short (200ns x 50) and long (1μs x 10) protocols overall agree with each other but not in the N-terminal portion (residues 35 to 41) (Figure 5.10A). Overall the long protocol agrees a bit better with experiment (Table 5.4). Experimental values occupy mostly positive secondary chemical shifts, indicating possible residual helical secondary structure in apo Rev and this is reproduced well in the ff14SB simulations. It is also worth noting experimental secondary chemical shifts are still within reasonable values typical of random coil, < 2 ppm. For ff14IDPSFF simulations, both short and long protocols produce nearly identical secondary chemical shift values (Figure 5.10B), lending support that the simulated observables converged very well. However, the agreement with experiment is not as good as the ff14SB simulations (Figure 5.10B and Table 5.4). Specifically, the ff14IDPSFF simulations may overestimate disordered structures in apo Rev.

Interestingly worse agreement is apparent between ff14SB-simulated $^3J_{HNH\alpha}$-coupling constants and experimental values (Figure 5.10C). Overall higher helical propensity is visible in the ff14SB simulations (average 4.2-5.6 Hz) versus higher disordered propensity (average 5.9-7.7 Hz) in the experiment (Figure 5.10C). Notably, ff14IDPSFF simulations agree closer to experiment in this regard with $^3J_{HNH\alpha}$-coupling constants in the similar range as in the

experiment. Nevertheless, both experimental and simulated $^{3}J_{HNH\alpha}$-coupling constants are still within reasonable range of disordered secondary structure. These ambiguous, sometimes overlapping secondary structure boundaries used in NMR experiments highlight the difficulty in definitively assigning secondary structures based on either chemical shifts and $^{3}J_{HNH\alpha}$-coupling constants. Multiple, independent CD experiments, however, suggest the conformational landscape of apo Rev is more populated as disordered than helical [18, 48, 67, 68]. In summary, the ff14IDPSFF simulations agree surprisingly well with both NMR and CD experiments with disordered structures dominant in its simulations of apo Rev. These observations will be highly useful in further refining IDP-specific force fields to improve simulation of complex, dynamic IDPs such as apo Rev.

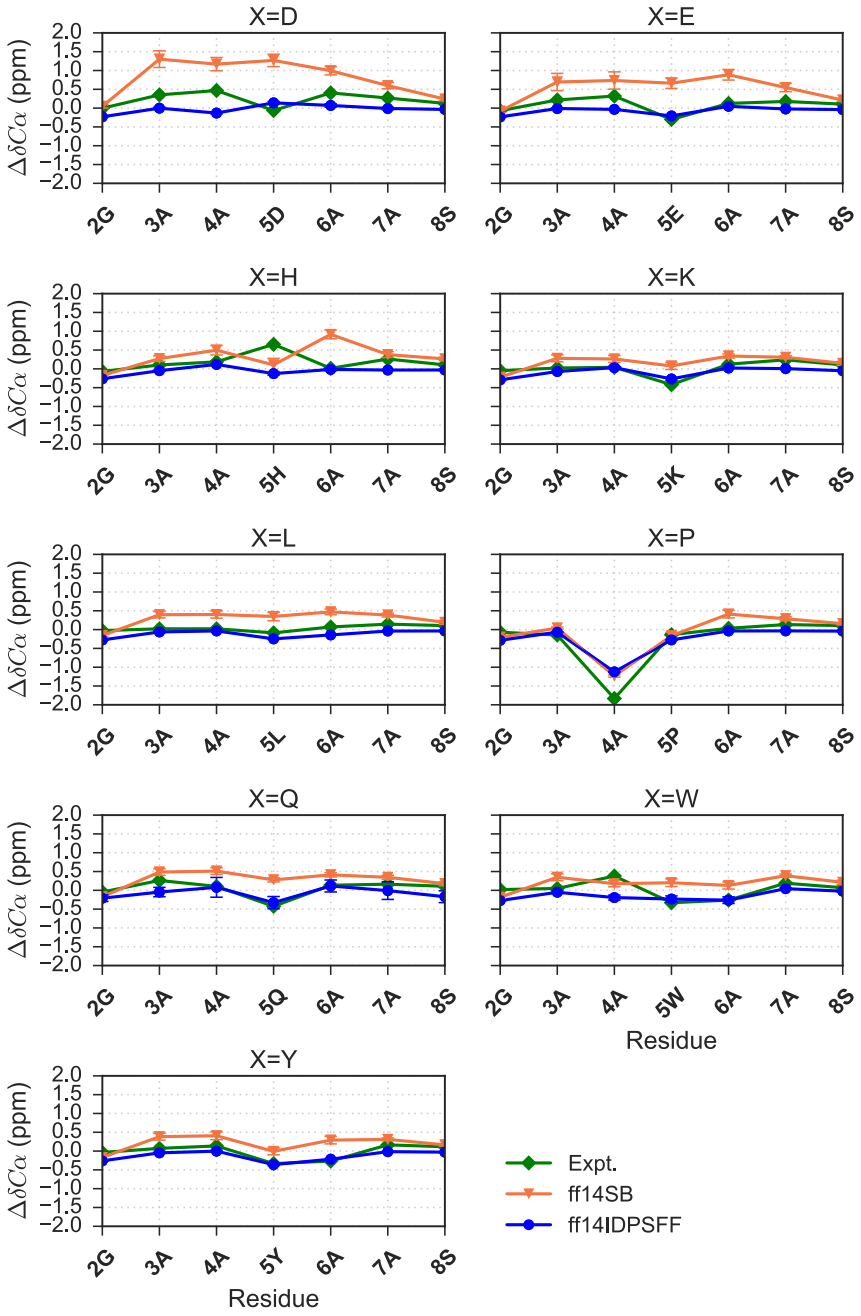


Figure 5.8: Comparison of experimental[64, 152] secondary Cα chemical shift values and simulated chemical shifts for the 9 short peptides (EGAAXAASS, X = D, E, H, K, L, P, Q, W, Y). Experimental and simulated values are colored as indicated in the legend. Standard deviation error bars are also visible for simulated values.

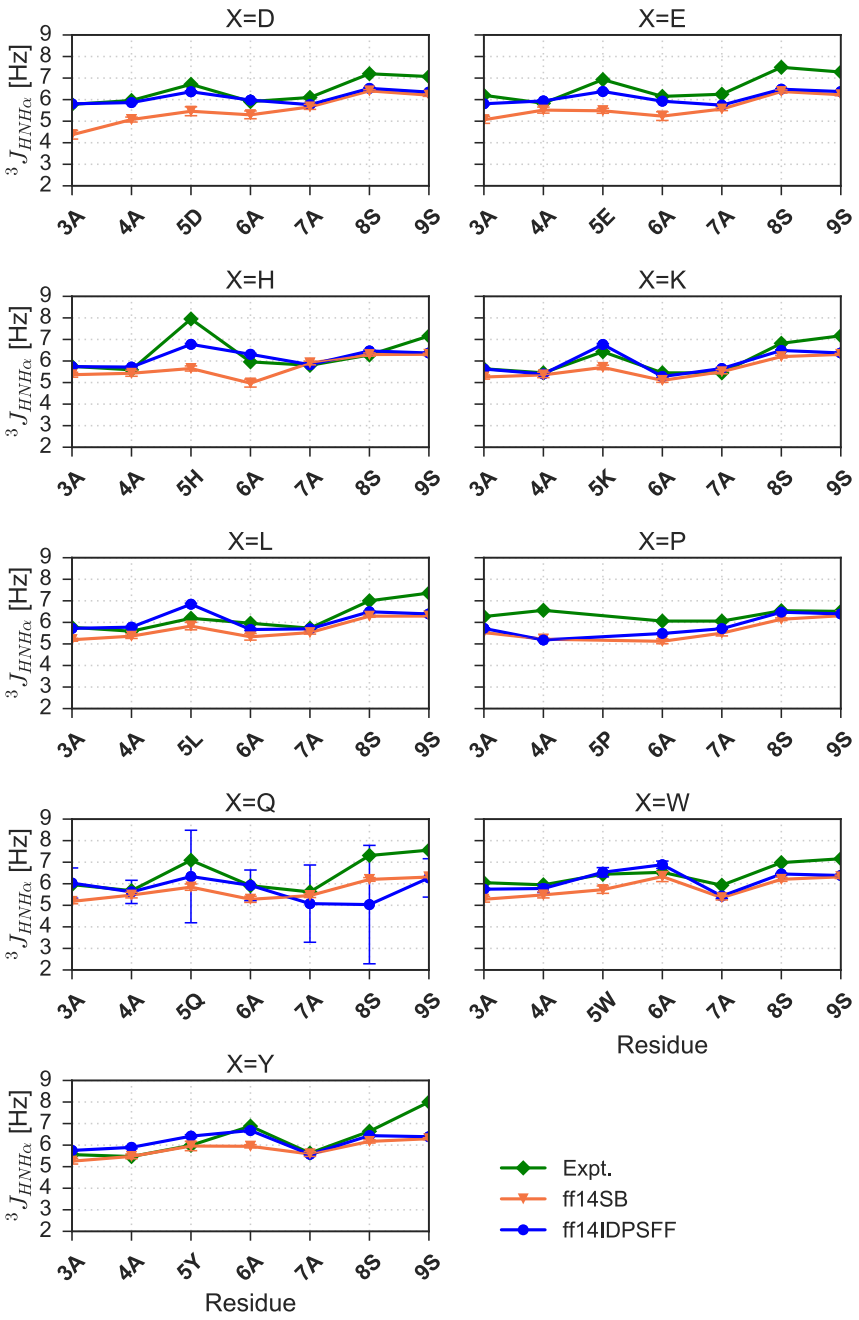


Figure 5.9: Calculated ff14IDSPFF- and ff14SB-parameterized $^3J_{HNH\alpha}$-coupling constants compared to experimentally-derived[64, 152] constants. Experimental and simulated values are colored as indicated in the legend. Standard deviation error bars are also visible for simulated values.

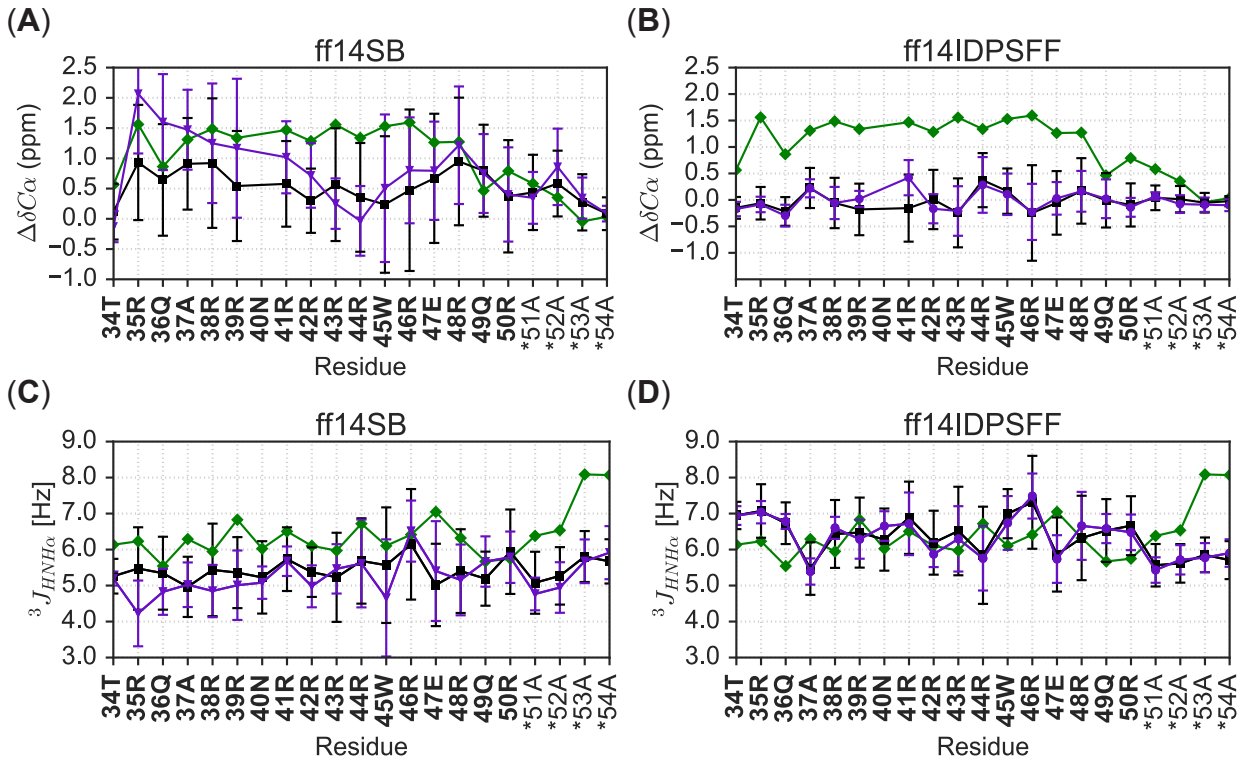


Figure 5.10: Comparison of force field and simulation types of apo Rev to experimental results. Colors are labeled according to experiment (green), short simulations (black), and long simulations (purple) and an asterisk (*) denotes non-native residues. (A) Comparison of short and long ff14SB-derived secondary chemical shifts with experiment.[48] (B) Comparison of short and long ff14IDPSFF-derived secondary chemical shifts to experiment.[48] (C) Comparison of short and long ff14SB-derived J-coupling constants with experiment.[48] (D) Comparison of short and long ff14IDPSFF-derived J-coupling constants with experiment.[48]

RRE-Rev Since the Rev protein is known to sustain a helical structure upon binding to its RNA-binding partner, Stem IIB of Rev response element (RRE), we also simulated the RRE-Rev complex (PDB: 1ETF) and compared to the apo Rev simulations. Experimental $\Delta\delta$Cα and $^3J_{HNH\alpha}$-coupling constant datasets were extracted from two separate literature sources and each source used different non-native residues in the N-terminal portion of otherwise identical Rev peptides [17, 18, 48]. The $^3J_{HNH\alpha}$-coupling dataset[48] was generated from a Rev peptide containing a 4-residue non-native extension (GAMA) at the N-terminus, while the $\Delta\delta$Cα dataset[17] resulted from a Rev peptide containing a non-native, N-terminal residue Asp. The GAMA sequence was a byproduct leftover from His6-GB1 tag, and the

Asp non-native sequence was used as an alternative to a synthetic N-terminal sequence from earlier experiments. Although we chose to simulate Rev bound to RRE with the N-terminal Asp from the literature,[17] the remaining 22 residues are identical between Rev peptides used in both experiments. Nevertheless experimental data show that both sequences from literature[17, 18, 48] exhibited RNA-binding specificity/activity in addition to disordered secondary structure in the apo state.

Although experimental chemical shifts fluctuate significantly, simulated values are stable and almost identical between the two force fields except terminal residues 49-52 (Figure 5.11). Both C-terminal experimental and simulated values seem to be decreasing to ranges characteristic of random coil (Figure 5.11). In analyses of $^{3}J_{HNH\alpha}$-coupling constants, experimental values and ff14SB-simulated values occupy typically helical ranges (< 5.6 Hz), whereas ff14IDPSFF-simulated values are almost identical to both ff14SB and experimental values until residue 49Q (Figure 5.11). The comparison shows that the beta-forming tendency is too strong for 49Q in the ff14IDPSFF simulations of the bound Rev (Figure 5.11B). Similar tendency is also noticeable in the ff14IDPSFF simulations of the apo Rev (Figure 5.10D) where the $^{3}J_{HNH\alpha}$-coupling constant is also overestimated for 49Q. This suggests further refinement is clearly required in the development of IDP force fields. RMSE differences between simulated NMR observables and experimental values are also rather close (Table 5.4), though the chemical shift agreement is not as good as those for the apo Rev simulations. This is probably because RRE was not considered in the conversion from MD conformations to chemical shifts by the SPARTA+ package [255]. Overall both ff14SB and ff14IDPSFF are adequate in the RRE-Rev simulations, with accuracy in predicted NMR observables comparable to that obtained for the NMR structure (RMSE of 2.50 ppm for $\Delta\delta$Cα and RMSE of 1.86 Hz for $^{3}J_{HNH\alpha}$-coupling constants).

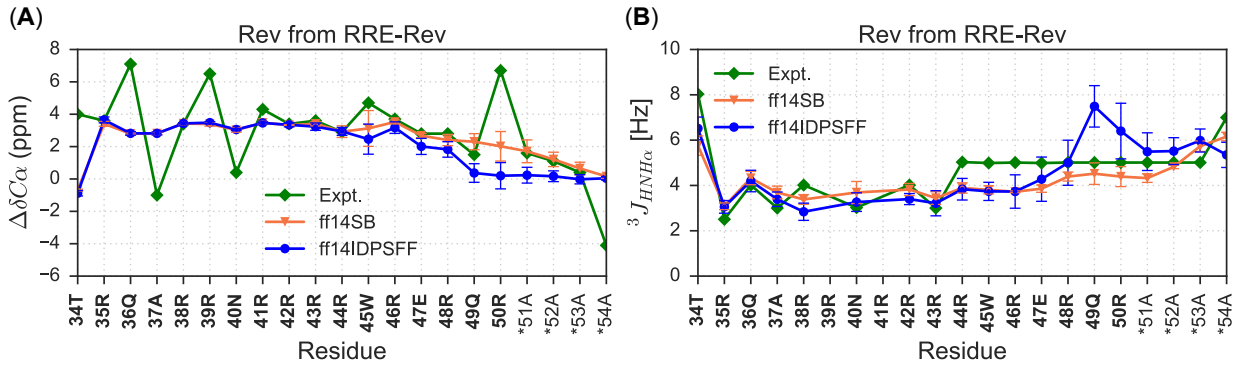


Figure 5.11: Simulated NMR observables are superimposed with experimental NMR values of Rev bound to the Stem IIB of RNA-binding partner, Rev-response element. Bold residues indicate native residues and asterisk (*) denotes non-native residues. (A) Comparison of experimental[17, 18] and average simulated $\Delta\delta$Cα values. (B) Comparison of experimental[17] and average simulated $^3J_{HNH\alpha}$-coupling constants.

5.4.1 Structural Signatures of Apo Rev Disordered State

Despite the extensive investigation of the Rev protein, as evidenced by 1647 hits from a general Pubmed search, this highly dynamic protein only occupies a monomeric state at submicromolar concentrations,[61] thus remaining elusive to structural characterization. Previous pursuits to structurally characterize the apo form of Rev encountered difficulties ranging from protein solubility to oligomerization, preventing characterization of apo Rev in physiological conditions [221]. Early circular dichroism (CD) and mutagenesis experiments suggest that apo Rev is disordered, forming helical structure depending on terminal amino acids (e.g. amidated C-terminus, C-terminal extension AAAR) [277]. Overall, attempts to characterize monomeric apo Rev have required techniques to induce ordered structure propensity, such as specific helix-inducing solution buffers (e.g. 2,2,2-trifluoroethanol), residue mutations to prevent oligomerization, or the introduction of structure-inducing binding partners [277, 244]. MD simulations thus provide a useful tool to probe the highly mobile conformations of Rev in its physiological disordered state. In previous structural modeling studies and MD simulations from Song et. al[262] and Casu et. al[48], researchers observed primarily coiled

secondary structure of apo Rev. These simulations however simulate apo Rev in nanosecond timescales. Herein we generated tens of microseconds trajectories to ensure proper sampling of disordered apo Rev conformations.

Clustering and secondary structure propensity calculations are discussed hereafter, highlighting the differences between the ff14SB and ff14IDPSFF simulations (in the long protocol). Although both ff14SB and ff14IDPSFF simulations exhibit ordered and disordered characteristics, the two force fields differ in secondary structure preferences: increased helical content observations in the generic ff14SB simulations (Figure 5.12), disordered structural preferences in the ff14IDPSFF simulations (Figure 5.13). The top ten clusters between both force fields occupy similar percentages: ff14SB at 17.87% versus ff14IDPSFF at 17.41%. Further evidence from DSSP[131] (hydrogen bond estimation algorithm) calculations also suggests the majority of ff14IDPSFF conformations exhibit coiled secondary structure, in Figure D.18. All residues in ff14IDPSFF simulations exhibit roughly equal probabilities of coiled secondary structure (average $> 80\%$) in addition to some beta contents (Figure D.18B-C, D.19B-C). DSSP (Figures D.18-D.19) and clustering results (Figures D.15-D.16) of the short protocol simulations are also provided in the supplementary information although simulations from the long protocol are the primary focus in this section. Experimental findings ranging from secondary chemical shift, $^{3}J_{HNH\alpha}$-coupling, and CD suggests apo Rev is mainly disordered when unbound [48]. Despite the observation that both force fields replicate the average coiled secondary structure as in experiment, these clustering analyses show that each force field exhibits either disordered or ordered structural bias – observations that will be useful in future refinement of IDP-specific force fields.

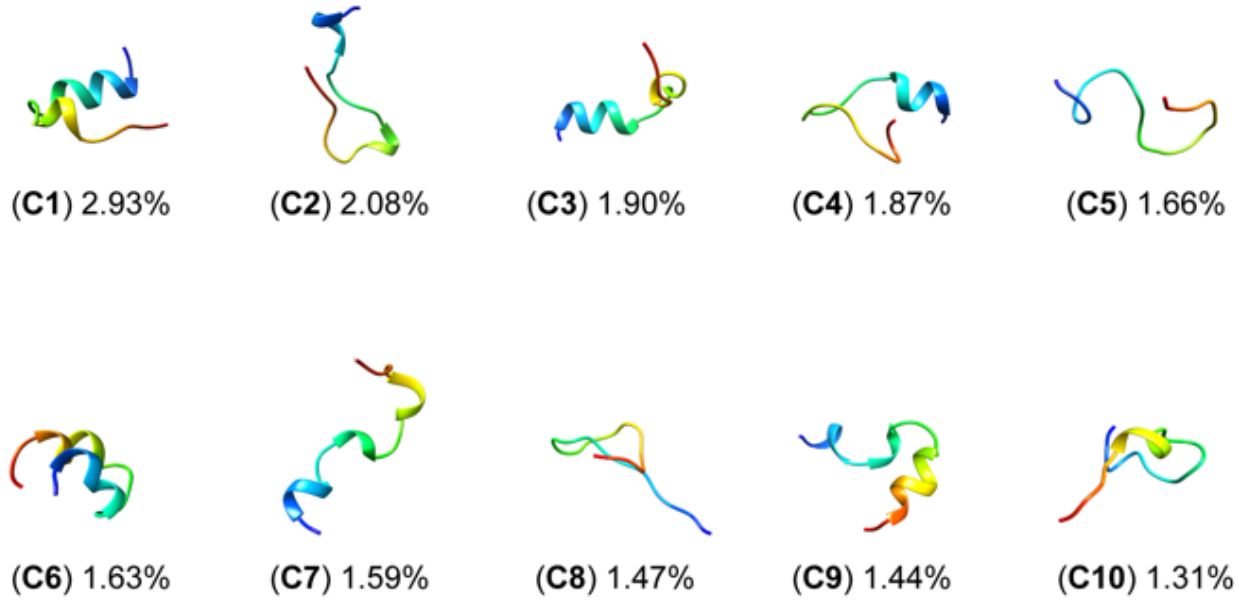


Figure 5.12: Top 10 clusters of ff14SB-parameterized simulations encompass 17.87% of all frames. Clusters are labeled C1-C10 and colored according to N- to C-termini sequence (red to blue).

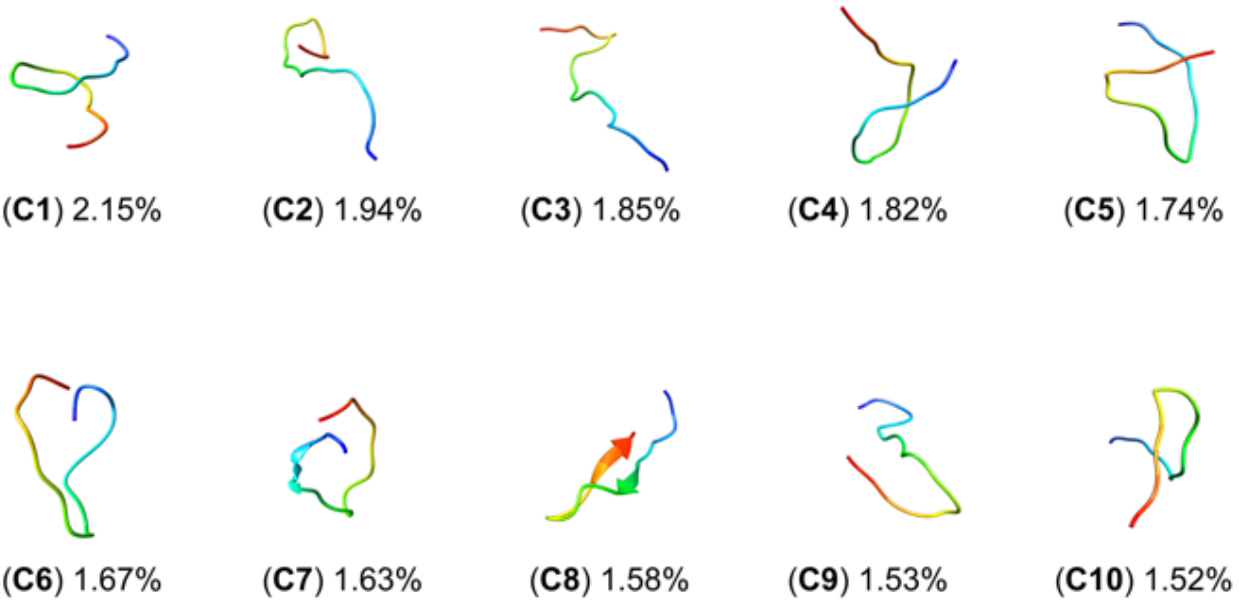


Figure 5.13: Top 10 clusters of ff14IDPSFF-parameterized simulations encompass 17.41% of all frames. Clusters are labeled C1-C10 and colored according to N- to C-termini sequence (red to blue).

5.4.2 Conformational Analysis of Bound Rev Ordered State

To supplement our apo Rev simulations above, we also simulated Rev bound to its RNA binding partner, RRE Stem IIB, to assess how our simulations perform in replicating experimentally-observed behaviors such as induced fit [61, 325]. Previous studies emphasize induced fit and proper RRE binding requires the presence of a single Rev monomer, from which more Rev monomers are recruited and oligomerize [61]. The NMR solution structure depicts an α-helical Rev situated in the major groove of RRE-Stem IIB [17]. After simulating this complex, we proceeded to align the Rev peptide from the NMR solution structure (PDB: 1ETF) to the average Rev structure extracted from RRE-Rev simulations (Figure 5.14). Simulations of Rev bound to RRE yield significantly more stabilized conformations compared to apo simulations. In the ff14SB simulations, we observed almost entirely helical content (Figure 5.14). In ff14IDPSFF force field simulations, helical secondary structure was observed in N-terminal residues, whereas coiled, disordered structure was observed in C-terminal residues (Figure 5.14). We also estimated the average secondary structure propensities of each residue for all simulations using the DSSP algorithm (Figure D.20). Despite some fluctuation in the last 4-5 C-terminal residues, most residues remain fairly stable, retaining the characteristic helical conformation found in the NMR solution structure (Figure D.20) [48].

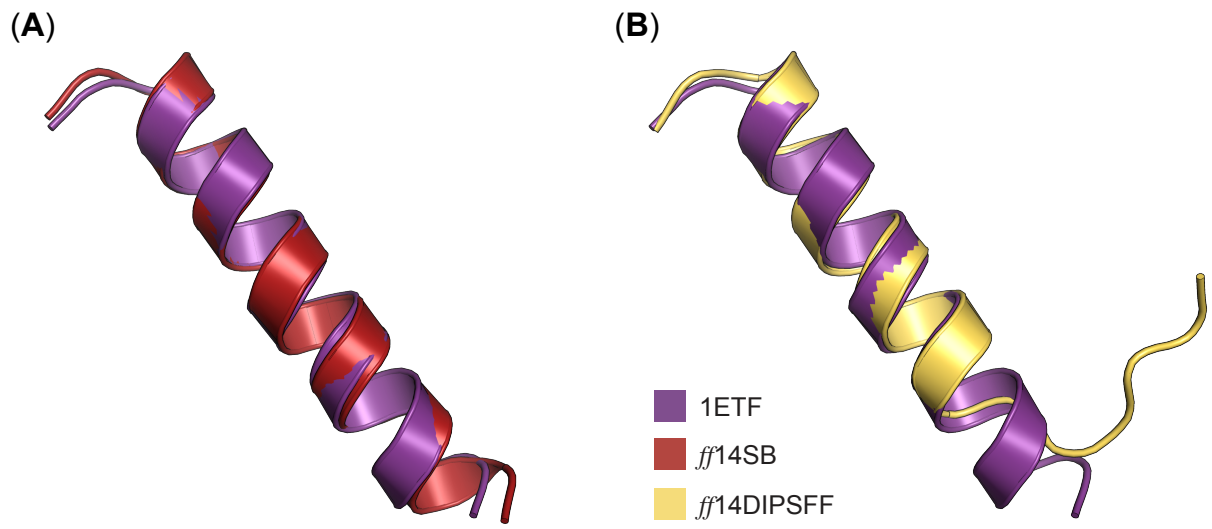


Figure 5.14: Alignment of average Rev structure from ff14SB and ff14IDPSFF RRE-Rev simulations to chain B in the NMR solution structure (PDB: 1ETF). (A) The average structure from ff14SB simulations is superimposed to Rev protein from 1ETF, with an RMSD of 0.57 (Cα atoms). (B) The average structure from ff14IDPSFF simulations is superimposed to Rev protein from 1ETF, with an RMSD of 1.14 (Cα atoms).

Unsurprisingly, ff14SB simulations yield a lower RMSD than ff14IDPSFF simulations from alignments to the experimental structure (Figure 5.14). This induced helical content is most likely attributed to inherent native-structure-biases of the generic ff14SB protein force field [22, 89, 91, 113]. Although the RMSD of the experimental and ff14IDPSFF-derived structure is larger, it is notable that the helical component is quite stable (first 16 residues), with the remaining 7 residues exhibiting multiple helix-to-coil transitions (Figure 5.14, D.20). Chemical shift and CD data of the wild-type Rev and various mutants (oligomerization-deficient mutant V16D/I55N Rev, and L60R mutant Rev bound to Stem IIB RRE), also suggests disordered content in the C-terminus [18, 48, 67, 68]. The stable N-terminal fragment found in ff14SB- and ff14IDPSFF-simulated residues contrasts sharply with the high structural fluctuation observed in apo Rev simulations, and is consistent with experimental RRE-Rev results [48]. Alignment of average simulated complexes also generated structures similar to the experimental NMR solution structure (Figure 5.15).

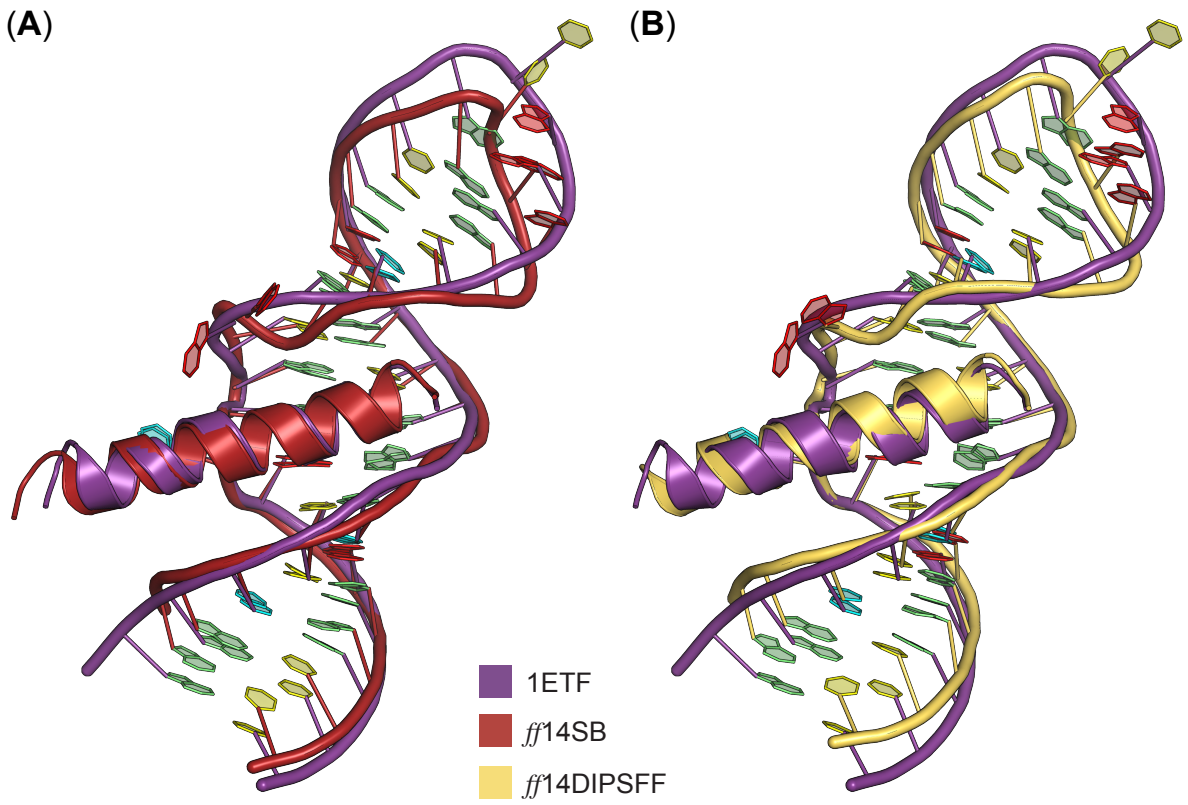


Figure 5.15: Alignment of average complex structure from ff14SB and ff14IDPSFF RRE-Rev simulations to the full NMR solution structure (PDB: 1ETF). Nitrogenous bases are colored according to Nucleic Acid Database convention: A – red, U – cyan, C – yellow, and G – green. **(A)** The average structure from ff14SB simulations (red) is superimposed to RRE-Rev from 1ETF, with an RMSD of 1.48 (backbone atoms: CA, P, O5', O3', C3', C4', C5'). **(B)** The average structure from ff14IDPSFF simulations is superimposed to RRE-Rev from 1ETF, with an RMSD of 1.9 (backbone atoms: CA, P, O5', O3', C3', C4', C5').

Fluctuation of Rev backbone atoms are further explored via root-mean squared fluctuation (RMSF) analyses for apo and bound Rev simulations. In all Rev simulations, backbone atoms (Cα) fluctuate more in ff14IDPSFF simulations than the ff14SB simulations (Figure 5.16). Comparison of apo and bound simulations shows the bound Rev fluctuates less, due to the stabilization from binding with RRE (Figure 5.16C, D.21). Unsurprisingly terminal residues display the highest fluctuation in all simulations, except the relatively stable N-terminal region in the bound Rev simulations. This is corroborated by hydrogen bonding populations of residues 34-36 (Figure 5.16, D.21, Table 5.5), which stabilizes the N-terminal

region. The observed different fluctuation trends can also be explained by the different secondary structure propensities. For instance in Figure 5.16B, residues 36-38 in the ff14SB apo Rev simulations exhibit lower RMSF values and also exhibit higher helical propensity (Figure D.18A).

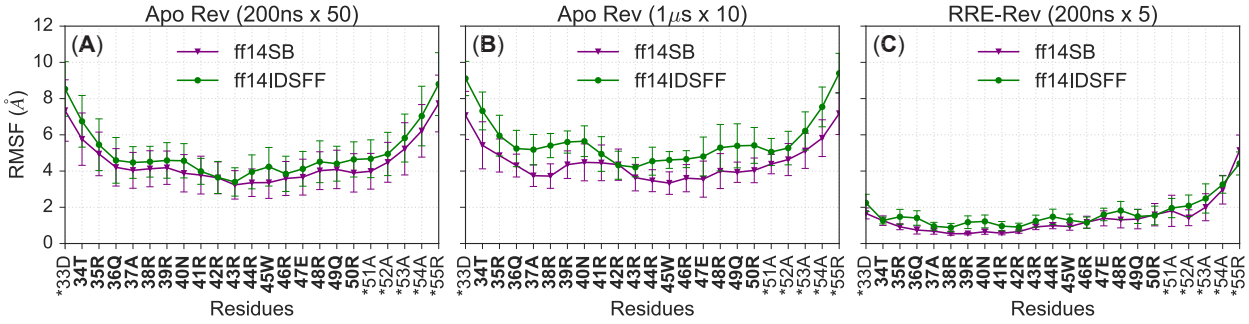


Figure 5.16: RMSF analyses of backbone Cα atoms per force field and simulation type. (A) Average RMSF of backbone atoms between fifty, 200ns apo Rev simulations. Asterisks (*) indicate non-native residues. (B) Average RMSF of backbone atoms between ten, 1μs apo Rev simulations. (C) Average RMSF of backbone atoms between five, 200ns RRE-Rev simulations.

Inspection of intermolecular hydrogen bond and ionic salt bridge occupancies (only frequencies > 0.5 is shown) in Table 5.5 and 6 reveals similar interactions between simulations of both force fields, but with slight differences (Table 5.5). Since ionic salt bridge formations are almost identical between the two force fields (Table 5.6), we chose to focus primarily on differences in hydrogen bond formation. In ff14SB complex simulations, the hydrogen bond pair ARG46-U72 dominates compared to ff14IDPSFF complexes due to the increased stability and helical propensity of the C-terminal end (Table 5.5). While retaining mostly helical character between residues 33-46, Rev contains two hydrogen bonds (GLN36-G47, ARG41-U45) in the N-terminal region in the ff14IDPSFF simulations, which are less frequent in the ff14SB simulations, an unexpected outcome considering the stability of the ff14SB simulations over that of the ff14IDPSFF simulations (Table 5.5). Co-existence of stabilized N-terminal helices and coiled C-terminal components in the ff14IDPSFF simulations of bound Rev suggests this new force field is able to simulate disordered region in an

otherwise ordered protein, while the ff14SB simulation retains more helical characteristics.

Table 5.5: Intermolecular Hydrogen Bond Occupancy (criteria: $\theta > 120°$, distance < 2.5 Å)[15]

Row Number	**Donor Residue**	**Acceptor Residue**	**Freq. (ff14SB)**	**Freq. (ff14IDPSFF)**
0	THR34	G47	0.5926	0.6576
1	ARG35	C65	0.753	0.5848
2	ARG35	U66	0.8388	0.7287
3	GLN36	G48	0.7831	0.6025
4	ARG38	U66	0.9777	0.9303
5	ARG38	G67	0.7867	0.7301
6	ARG39	G70	0.9918	0.9702
7	ASN40	G47	0.8201	0.9814
8	ASN40	G46	0.6765	0.8927
9	ARG41	G46	0.6674	0.7484
10	ARG42	G67	0.8515	0.8345
11	ARG42	A68	0.764	0.8502
12	ARG44	U45	0.7013	0.728
13	ARG46	U72	0.6373	0.4805
14	ARG48	U43	0.8294	0.7139
15	ARG48	C44	0.6949	0.6611
16	GLN36	G47	0.3891	0.5076
17	ARG41	U45	0.4667	0.5766

Table 5.6: Intermolecular Ionic Salt Bridge Occupancy (criterion: distance < 4 Å)[16]

Row Number	Acidic Residue	Basic Residue	Freq. (ff14SB)	Freq. (ff14IDPSFF)
0	U43	ARG48	0.8611	0.7314
1	C44	ARG48	0.7535	0.8062
2	U45	ARG44	0.5244	0.5146
3	G46	ARG41	0.7136	0.8019
4	C65	ARG35	0.7934	0.6226
5	U66	ARG35	0.8821	0.7189
6	U66	ARG38	0.9821	0.9527
7	G67	ARG38	0.7981	0.7406
8	G67	ARG42	0.9152	0.9513
9	A68	ARG42	0.7722	0.8712
10	U72	ARG46	0.6879	0.6017
11	U45	ARG41	0.4922	0.596

5.5 Conclusion

IDPs remain elusive by standard experimental methods due to their conformational flexibility. Molecular dynamics simulations can thus provide detailed insight into their complex structures, dynamics, and functions, if they can reproduce the available experimental observables. However, there are several issues in computational studies. First the generic force fields were found to be biased towards ordered structures in many prior simulation studies. Second the expansive conformations occupied by IDPs is often beyond typical simulation amount needed for ordered proteins.

Thus, our first goal of this computational study is to assess the quality of both a generic protein force field (ff14SB) and its IDP-specific counterpart (ff14IDPSFF) that was intended

to address the biases in the generic force field. Overall simulated average observables from ff14IDPSFF replicate experimental chemical shifts and $^3J_{HNH\alpha}$-coupling constants more accurately than those derived from ff14SB simulations for the tested EGAAXAASS peptides. DSSP analyses also suggest different secondary structural biases between the two force fields, increased helical content from ff14SB and coiled content from ff14IDPSFF, with the latter in higher agreement with experiment. When used to simulate more complex proteins such as Rev in apo and bound forms, computational models gravitate toward either ordered secondary structure (ff14SB) or disordered secondary structure (ff14IDPSFF) as the clustering analyses revealed. However simulated observables between the two force fields are roughly comparable to experiment, ff14IDPSFF simulations agree with both NMR and CD measurements slightly better.

Our second goal of this study is to assess the extent of sampling that is needed for quantitative structural annotation of IDPs and to explore how to assess the sampling convergence. This was first conducted by analyses of convergence rates of individual observables in the form of bi-phasic decays. Convergence analyses of both NMR observables show that ff14IDPSFF simulations converge slightly faster than ff14SB simulations in the chemical shift calculations for all tested systems, though they converge slightly slower for $^3J_{HNH\alpha}$-coupling constants for all tested systems. This is consistent with the observations that conformations in ff14IDPSFF simulations are more diversified, sampling a larger range of main-chain torsion angles, leading to slower convergence in $^3J_{HNH\alpha}$-coupling constants that solely depends on these torsion angles. The decay half times also show that the total sampling amount (in term of nanoseconds simulated) is adequate as they are much less the total amount collected.

In addition, simulation protocols were also tested by simulating apo Rev as either many short (50 x 200ns) trajectories or a few long (10 x 1μs) trajectories. Consistently converged distributions in the ff14IDPSFF simulations allows us to use the convergence rates to compare which protocol is better. However, the rate estimations show that differences in the

convergence rates between the two are small, within 200ns in general, though it can be said the short protocol is slightly faster than the long protocol. For ff14SB simulations, the different distributions give us pause to claim that the sampling of the apo Rev is sufficient in either protocol even if 10 microseconds worth of sampling has been collected. This indicates that enhanced sampling techniques would greatly benefit IDP simulations for systems as small as 23 amino acids such as apo Rev.

Despite the short sequence length of apo Rev, no monomeric disordered Rev protein has been structurally characterized as demonstrated by its absence in the Protein Data Bank (PDB). To compensate for this lack of structural characterization, we utilized a combination of NMR and CD data for comparison to our clustering and secondary structural analyses. Chemical shift and CD studies from various different sources of oligomerization-deficient mutants and wildtype Rev conclude that monomeric Rev is mostly disordered [18, 48, 67, 68]. These experimental findings are comparable to random coil clusters and DSSP calculations from the ff14DIPSFF simulations of and differ from the ff14SB simulations where increased helical content was found. Both force fields also generate stabilized helical structure and induced fit in RRE-REV simulations, exhibiting a coiled C-terminus as shown by the chemical shift data [17, 18, 48]. These structural computational studies of apo and bound Rev stress the importance to assign the correct secondary structural biases in both force fields.

Interesting observations were also found when Rev was simulated with its RNA-binding partner RRE, ff14DIPSFF was able to replicate the structured regions in the bound form, despite over-representation of coiled secondary structure in the apo Rev simulations. Detailed analysis of the average conformation and secondary structures of the ff14IDPSFF simulations shows that both the helical N-terminal region and coiled C-terminal region are readily observed, in agreement with experimental findings, despite coiled secondary structural preferences in the apo Rev simulations. In comparison, a more stable helical structure was observed throughout the ff14SB simulations. A natural next step is to ask a more

quantitative question: whether ff14SB is too stable or ff14IDPSFF is too unstable in the simulations of more complex IDPs such as Rev. This requires further quantitative stability analysis both experimentally and computationally.

This study articulates the difficulties of obtaining converged and expansive sampling of IDPs, though our exploration of different simulation protocols demonstrates consistent observations with the ff14IDPSFF force field regardless of the protocols used. Although successful in simulating short peptides and bound Rev, the advantages of ff14IDPSFF are not as clear-cut for the more complex apo Rev. These findings also suggest future refinements of IDP-specific force fields and reduction of force field biases are still necessary for consistent performance in modeling IDPs.

5.6 Acknowledgments

We thank Drs. Song and Chen for supplying the parameter set and the perl script to implement the ff14IDPSFF force field. V.T.D. was supported by the Mathematical, Computational and Systems Biology Pre-doctoral Training Grant T32 EB009418-08. This work was supported in part by NIH/NIGMS (GM093040 & GM079383 to R.L.).

Chapter 6

Neural upscaling from coarse protein structure networks to atomistic structures

6.1 Summary

As protein structural landscapes exhibit an increasingly diverse array of behavior and complexity, here we explore the utility of expanding exploratory methods through residue-level Protein Structure Networks (PSNs). As shown in previous work by the Butts lab, proteins can be represented as PSNs and fitted with exponential random graph models (ERGMs). An ERGM is statistical model where one attempts to fit parameters to this model such that they maximize the likelihood of observing a given network, whose energy function which is defined by a network Hamiltonian. This PSN simulation methodology can thus greatly extend the timescales accessible to computer simulations of proteins that sample diverse structural conformations over long timescales. Since PSNs represent proteins in a coarse

structural form, further information can be extracted if PSNs can be transformed into an atomistic model. Here, we use a multi-layer perceptron neural network to do exactly this with the protein amyloid-β. Amyloid-β is an intrinsically disordered protein exhibiting a dynamic range of secondary structural conformations (e.g. α-helix, β sheet, random coil). This work demonstrates it is possible to use a neural network to map from coarser PSN representations of macromolecular configurations to finer atomistic configurations. Therefore, a PSN model can possess a surprisingly minimal loss of structural information compared to classical atomistic simulations, especially considering PSN dynamics are orders of magnitude less costly to simulate than their atomistic counterparts. The trained neural network is able to reconstruct the complex conformations of amyloid-β at the atomic level from coarse binary contact adjacency matrices extracted from PSNs, thereby expanding the toolkit of protein conformation exploration.

6.2 Background

Proteins and biomolecules exhibit q wide variety of complex dynamics and interactions at varying size and time scales. Coarse-grained (CG) models offer an alternative means to traditional atomistic simulations by traversing larger timescales (e.g. beyond microsecond timescales) as well as representing biomolecules at varying degrees of freedom (e.g. residue-level, chemical moiety-level, etc.). Coarse-grained (CG) simulations can be parameterized using either a force field (e.g. MARTINI [180]) or graph-based theoretic terms [96]. Despite their temporal advantage, coarse grained models benefit from the additional step of backmapping/upscaling to atomic level in order to infer finer detailed observables. Although methods to backmap or reverse map from CG to atomistic models exist, a majority of reverse mapping methods focus primarily on force field-based CG models. Methods to backmap or reverse map force field-based CG models consist of two steps, beginning with model gen-

eration using either random placement [239], fragment-based [110, 215], or geometric-based [95, 30, 317, 168], followed by an equilibration step to relax the system. However, these methods sufficiently reverse map specifically CG force field-based models.

Previous approaches in graphed, coarse-grained modeling have focused primarily on mapping from atomistic to coarse grain networks [320, 50, 96]. One study in particular was able to simulate amyloid fibril aggregation, representing the fibril topology as network representations fitted with statistical models, exponential random graph models (ERGMs) [96]. Development of reverse mapping methods specific to these network simulation techniques can thus expand the utility of this technique to explore complex protein conformations.

Over time, the development of CG-based simulation methods has also steadily incorporated machine learning techniques [20, 27, 50, 151, 310, 320, 338]. However to our knowledge, multilayer perceptron-based (MLP) neural networks have not been incorporated with graph-based CG methods. MLP neural network architecture is a supervised learning technique capable of fine-tuning weights and biases, in this context specific to our input (contact adjacency matrices) and output (pairwise interatomic distances). Its capabilities differentiate it from a linear perceptron with its ability to interpret non-linear data.

In this work, we demonstrate the utility of multilayer perceptron neural network models to translate coarse protein structure network representations to their more finely detailed 3D coordinate structures. From coarse network representations, the trained neural network is able reproduce the conformations of amyloid-β protein to atomic-level detail while capturing its diverse secondary structure behavior. Training to contact adjacency matrices and their corresponding pairwise interatomic distances (PIDs) allows the neural network to learn detailed and specific structural information. CG network representations combined with a MLP neural network architecture can thereby capture this complex atomistic data, expanding the utility of graph-based CG modeling into applications where atomic coordinates are needed.

6.3 Methods

Data Generation Molecular dynamics simulations are the basis from which input and output data are extracted to train the model (Figure 6.1). Although PSNs can be simulated using an exponential family of random graph models (ERGMs), their starting structure is typically derived from an atomic model. Beginning with the lowest energy monomer of the PDB structure, 2LFM, β-amyloid protein was simulated for 1 μs using NAMD via the following protocol: initial monomer structure was solvated in a cubic TIP3P water box of minimum margin 25 Angstroms, and neutralized with NaCl counter-ions. This assembly was minimized for 10,000 iterations, followed by velocity initialization and 250 simulation iterations before final adjustment of the water box. A one μs trajectory was then simulated. Simulation was performed under periodic boundary conditions in NAMD, using an NPT ensemble at 300K and 1 atm pressure. Temperature control was maintained by Langevin dynamics with a period of 1/ps, with Nosé-Hoover Langevin piston pressure control. The CHARMM 36m forcefield was employed. Monomer states were sampled from the trajectory every 100ps, from which residue-level protein structure networks were constructed. Vertices correspond to individual residues, with two vertices being considered adjacent if they contain respective atoms whose distance is less than or equal to 1.1 times the sum of their van der Waals radii.

The simulation contains 11,926 total frames/conformations, of which 72% was allocated for training, 20% for testing, and 8% for validation. A 5-fold cross validation was also performed to ensure bias was not introduced during initial train-test splitting (Figure SE.1). For each frame in the amyloid-β simulation, a protein structure network (PSN) was calculated using software from [38] (in combination with VMD [120] and the `statnet` library [102, 34] for R [227]).

Monomer states were sampled from the trajectory every 100 ps, from which residue-level

protein structure networks were constructed. Vertices correspond to individual residues, with two vertices being considered adjacent if they contain respective atoms whose distance is less than or equal to 1.1 times the sum of their van der Waals radii. The input data used to train the neural network model consists of the flattened upper triangular data extracted from the residue-level contact adjacency matrix for each conformation in the amyloid-β microsecond simulation. The output data used to train the model is the flattened upper triangular of pairwise interatomic distance matrices calculated for each non-hydrogen atom (across all frames in the MD simulation) (Figure 6.1).

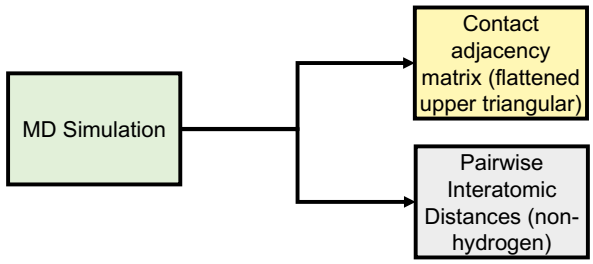


Figure 6.1: Data generation of input (upper triangular of contact adjacency matrices) and output (upper triangular of PIDs) data.

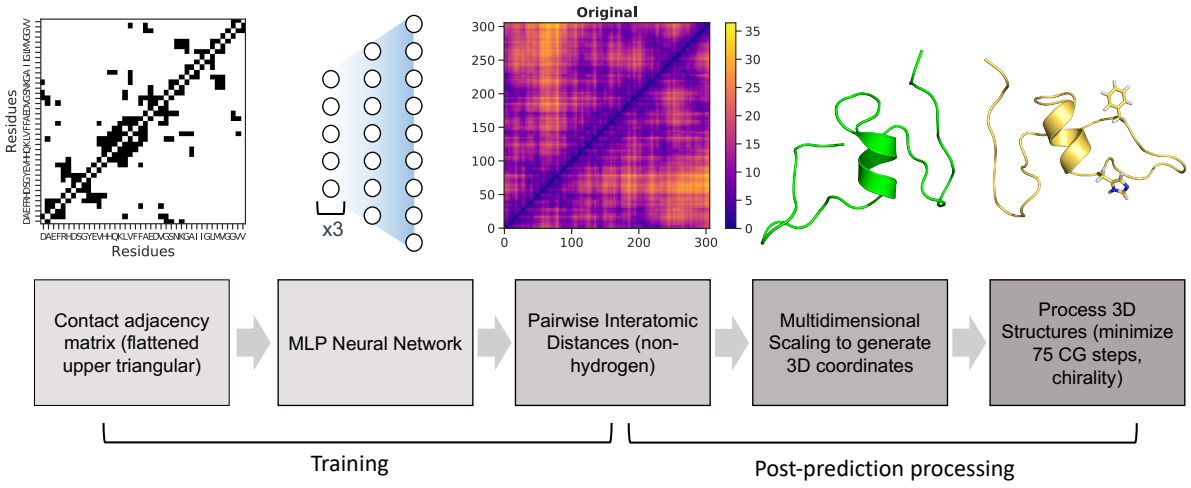


Figure 6.2: Pipeline of MLP neural network training and post-prediction processing.

Neural network architecture and hyperparameters After generation of input and

output data, a multi-layer perceptron (MLP) neural network was utilized for training as indicated in the pipeline (Figure 6.2). The neural network is based on a multi-layer perceptron utilizing the machine-learning Keras [58] and tensorflow [2] framework. The first three hidden layers consist of 2000 neurons, the fourth layer contains 8000 neurons, and the last output layer predicts the flattened upper triangular of the pairwise interatomic distance matrix for a given frame from the MD simulation (46665 neurons) (Figure 6.2). Hyperparameters were optimized using the Talos Keras tuning module [1]. A Nvidia P6000 Quadro GPU card was used to train the model with the following hyperparameters: nonlinearity = relu, dropout rate = 0.2, optimization = AMSGrad, loss = mean squared error, batch size = 50, epochs = 100. Predicted output data were initially assessed using three metrics: root-mean squared deviation/error (RMSD/RMSE), mean squared error (MSE), and mean absolute percentage error (MAPE).

Post-prediction processing The predicted output data (flattened upper triangular data of pairwise interatomic distance matrices) were then transformed into symmetric pairwise interatomic distance matrices. This was then transformed into 3D coordinate data using the multi-dimensional scaling function from scikit-learn python module and MDtraj [187] to generate PDB structures (Figure 6.2). Chimera [217] was then used to add hydrogens to predicted PDB structures, which were then further processed to remove inaccurate chiral predictions. If more than half of Cα centers were inaccurately predicted as R chiral centers (D-amino acids instead of L-aminio acids), this indicated the MDS portion predicted a reflection of the true coordinates. This was mitigated by reflecting all coordinates over the y-axis for predictions exhibiting an $\frac{R}{S}$ ratio greater than 1. If fewer than half of α-carbons exhibited R chiral centers, reflecting coordinates was unnecessary. Instead, Chimera was used to switch side chain coordinates and the α-hydrogen for all inaccurately predicted Cα chiral centers. After checking for correct chirality for each residue, all conformations were further minimized for 75 conjugate gradient steps.

The number of conjugate gradient steps was chosen by evaluating structures every subsequent 20 conjugate gradient steps for a cumulative 520 steps total. The maximum 520 conjugate gradient steps was chosen based on qualitative determination of average potential energy trends of all predicted conformations with increasing conjugate gradient minimization (Figure SE.2). Three superposition-based metrics (RMSD, global distance test, total score (GDT_TS), template modeling (TM) score) and one superposition-free metric (local distance difference test (LDDT)) were used to analyze any potential improvements in additional conjugate gradient steps between predicted 3D structure and the original, MD-generated 3D conformation. The RMSD metric analyze all heavy atoms, TM score focuses primarily on Cα atoms, and GDT_TS also focuses primarily on backbone atoms. The LDDT score calculates a comparison using all-atom pairwise interatomic distances. Average values of 500 randomly chosen structures (RMSD, TM Scores, GDT_TS, and LDDT) suggest a minimization range between 50-100 conjugate gradient steps. Thus 75 steps was chosen as the total number of conjugate gradient steps to minimize all 11,926 predicted conformations. Overall, minimization yields minimial improvement relative to no minimization, however is a necessary step to remove steric clashes and slight stereochemical errors (Figure E.3).

6.4 Results

6.4.1 Multilayer perceptron (MLP) neural network reconstructs Aβ conformations with atomistic detail

Pairwise interatomic distance (PID) predictions were made for all sets of data (train, validation, test). Predictions were evaluated against original PIDs from MD simulation using root-mean square error/deviation (RMSE/RMSD), mean absolute error (MAE), mean absolute percentage error (MAPE). The average metrics for the test set exhibit a favorable

RMSE (1.7Å), MAE (1.17Å), and MAPE (7.35%) (Figure 6.3). A 5-fold cross-validation suggests bias was not arbitrarily introduced during the initial train-test split (Figure E.1). Overall, average PID metrics for the validation and test set suggest the neural network was able to devise quality predictions.

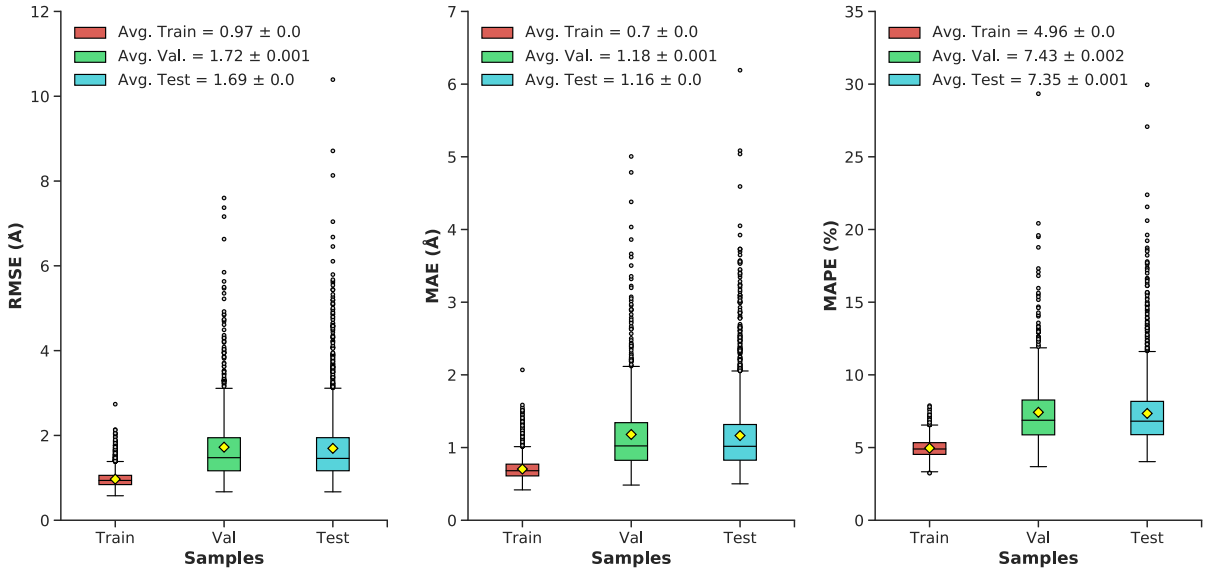


Figure 6.3: Boxplot distributions summarize the following metrics (RMSE, MAE, MAPE) for the train, validation, and test datasets: minimum, maximum, median, outliers (grey dots), average (yellow diamond) ± standard error, lower and upper quartiles.

To illustrate model performance, we assess a range of examples from the test set, beginning with frame 1133. Original and predicted pairwise interatomic distances for frame 1133 upon initial visualization, have highly comparable values (Figure 6.4A-B). A grayscale depiction of absolute value differences between original and predicted PIDs reveals white and light grey data points, denoting mostly low values (Figure 6.4D). A distribution of this data shows approximately 98% of difference values are less than 2 Å and 88% are less than 1 Å (Figure 6.4C). Within the test set, this is an example of one of best performing predictions made by the neural network model.

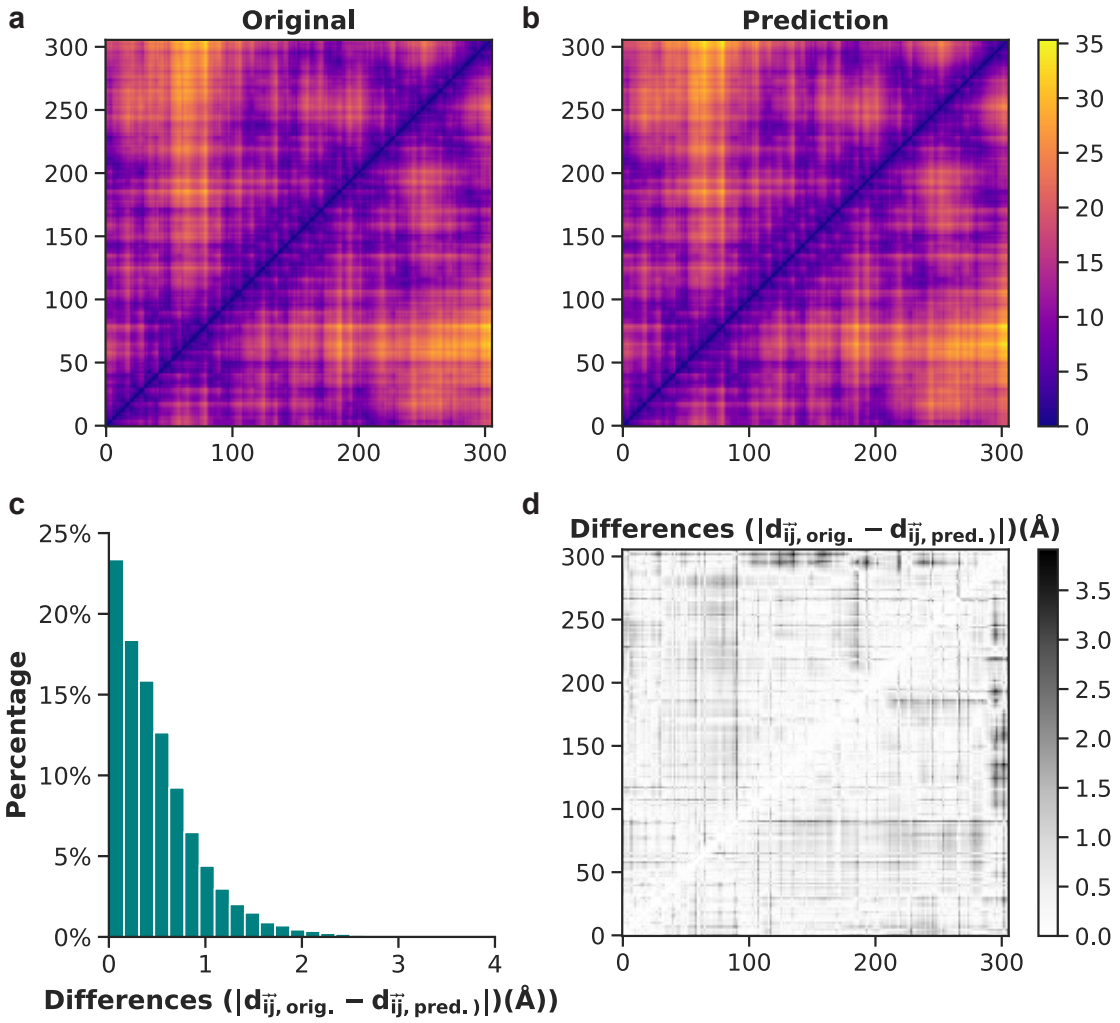


Figure 6.4: **Comparison between original and predicted pairwise interatomic distances for frame 1133 (from the test set)**. **a.** Actual distances are shown for all heavy atoms. **b.** Heavy-atom predictions of all pairwise interatomic distance. **c.** Histogram of differences between original and predicted euclidean distances. **d.** Binary plot displaying the absolute difference values between each actual and predicted distance for frame 1133.

Using RMSEs of PIDs as a basis, we show processed 3D predictions of the lowest RMSE score representation (frame 1133, Figure 6.5A), the median representation (frame 7431, Figure 6.5B), and the highest RMSE score structure (frame 7560, Figure 6.5B). The best prediction with the lowest RMSE (0.67 Å) exhibits more helical secondary structure compared to median and the worst predictions, which exhibit more random coil-like dynamics. RMSE of all heavy atoms for the median representation exhibits a fairly reasonable value of

1.46 Å whereas the worst PID prediction has a RMSE of 10.4 Å. Notably, the prediction for Figure 6.5C aligns reasonably well for the first 20 residues and the remaining residues are more poorly predicted by the neural network model. Since the protein spends the majority of its time in more compact conformations, it is not surprising the neural network model struggles to predict this specific overly extended conformation. The RMSEs according to 3D structure alignment between original and processed 3D structure and not on the basis of PIDs also contain similar values: best (0.77 Å), median (2.13 Å), and worst (12.01 Å). These values are slightly higher compared to PID-based RMSEs most likely due to introduced 3D alignment wheres PIDs report RMSEs between all heavy atoms.

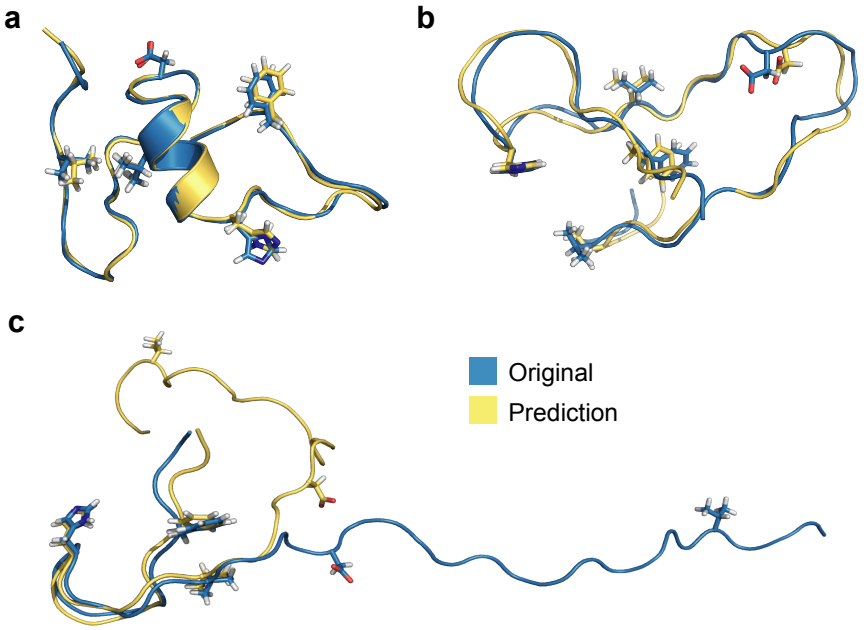


Figure 6.5: **Alignment between original and predicted and processed 3D structures** for (**a**) the best, (**b**) median, and (**c**) worst predictions based on RMSE values of PIDs.

6.4.2 Generation of 3D structures and subsequent minimization

When multidimensional scaling maps PIDs into 3D dimensional coordinates, it does so without regard to chirality. There are instances in which entire conformations are D- instead of L-amino acids, a correction that can be easily identified and fixed by reflecting coordinates

across the y-axis. We also corrected conformations that contained only a few instances of D-amino acids, a result of the neural network predicting slightly incorrect side chain and/or α-hydrogen PIDs. These chirality checks followed by minimization are necessary, computationally inexpensive processing steps required to transform PIDs into sterically reasonable 3D structures. Once corrections where fixed using Chimera, we then minimized all proteins for 75 conjugate gradient steps (a determination detailed in Methods), with a few conformations (23) requiring an additional 5 steps.

Figure 6.6 depicts a pre- and post-minimization of the best predicted conformation (frame 1133) in the test set. Here we focus particularly on residues histidine 13 (His13) and phenylalanine (Phe4). Both residues in the pre-minimized conformation are sterically incorrect and misplaced. Whereas in the post-minimized conformation, both residues have expected canonical sterics, devoid of incorrectly positioned atoms. When these optimization techniques (stereochemical corrections and minimization) are combined with the predictive power of the MLP neural network, this method yields highly effective predictive capabilities.

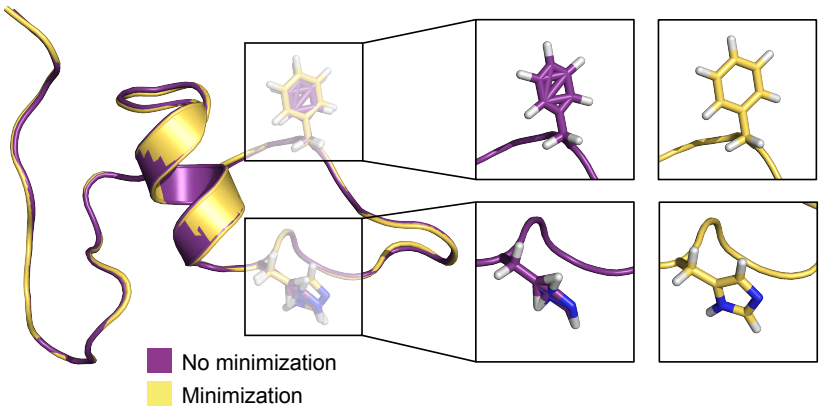


Figure 6.6: Comparison of pre- and post-minimized structures of the best prediction in the test set, frame 1133.

After minimization, it was also imperative to compare 3D minimized predictions to their original MD simulation counterparts. Three superposition-based metrics (RMSD, TM score, GDT_TS) and one superposition-free metric (local distance difference test (LDDT)) were

utilized for this evaluation. Template modeling (TM) score measures the backbone similarity between a reference protein and target protein with a range from 0 (dissimilar) to 1 (identical) [341]. RMSD is a canonical protein comparison metric and here we parameterize it to compare all heavy atoms between native and predicted structures. LDDT utilizes pairwise interatomic distances in its methodology, focusing on local intramolecular interactions and the degree (range 0-1) of their retention in the target conformation in comparison to the native reference structure [179]. Global distance test, total score (GDT_TS) is an improvement compared to RMSD designed to assess structures with the same sequence but different tertiary structure, with a higher score denoting better agreement (range 0-1) [336]. All four metrics are commonly used during the biennial Critical Assessment of Structure Prediction (CASP) structure prediction and assessment competition [142] and here we use these metrics to assess the predictive performance of the model.

Figure 6.7 illustrates these metrics for the combined validation-test set. There exists a positive correlation between LDDT vs. TM scores and GDT_TS (Figure 6.7A-B). Between RMSDs vs. TM scores and GDT_TS, predictions exhibit a negative correlation (Figure 6.7C-D). Included are also the aforementioned best (yellow diamond), median (purple diamond), and worst (red diamond) PID predictions from Figure 6.5. Since their designation as best, median and worst were on the basis of RMSEs of PIDs and not 3D structure, it is interesting to observe the surprisingly high LDDT value of frame 7560 (the worst prediction). This suggests the neural network was able to preserve more local residue interactions despite struggling with larger more regional intramolecular interactions. TM scores exhibit values in the lower range of < 0.5, whereas most GDT_TS and LDDT values occupy a range > 0.5, suggesting TM scores may not be as reliable of an assessment metric for amyloid-β. The average and 95% confidence intervals suggest predicted 3D models are predicted relatively well considering the high GDT_TS average and narrow 95% confidence interval (Figure 6.8). The best and median test cases occupy expected 3D metrics (Figure 6.7). In combination with PID metrics (Figure 6.3), the 3D metrics demonstrate the model's ability to reasonably

reconstruct the complex protein conformation of amyloid-β from coarse contact adjacency matrices.

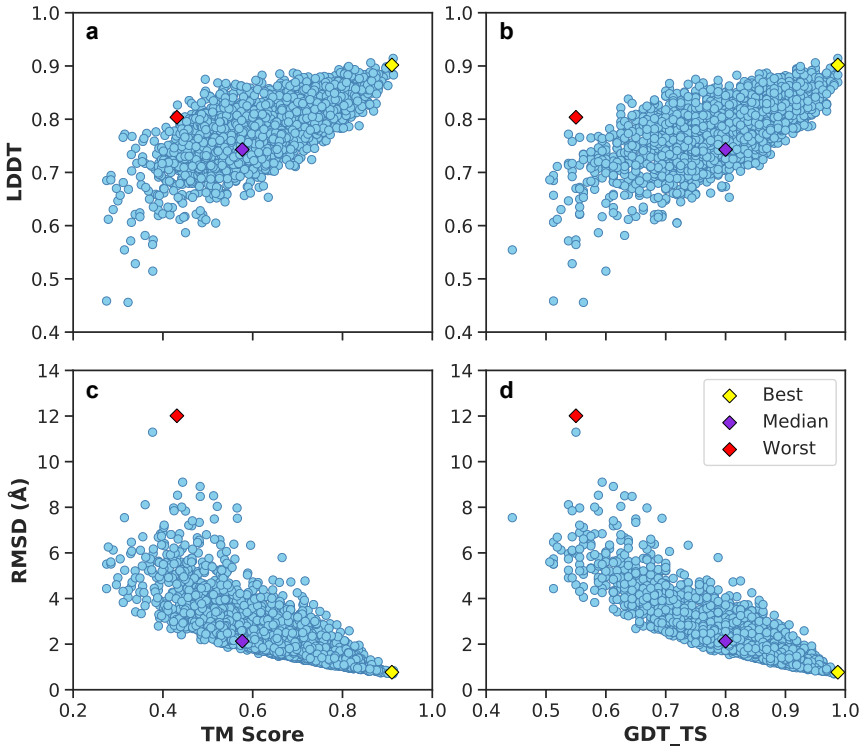


Figure 6.7: **Juxtaposition of 3D structural metrics of the combined validation-test set: TM score, LDDT, GDT_TS, and RMSD.** In addition, best, median, and worst predictions are shown based on PIDs. A) LDDT vs. TM score metrics of the validation-test set. B) LDDT vs. GDT_TS score metrics of the validation-test set. C) RMSD vs. TM score metrics of the validation-test set. D) RMSD vs. GTD_TS score metrics of the validation-test set.

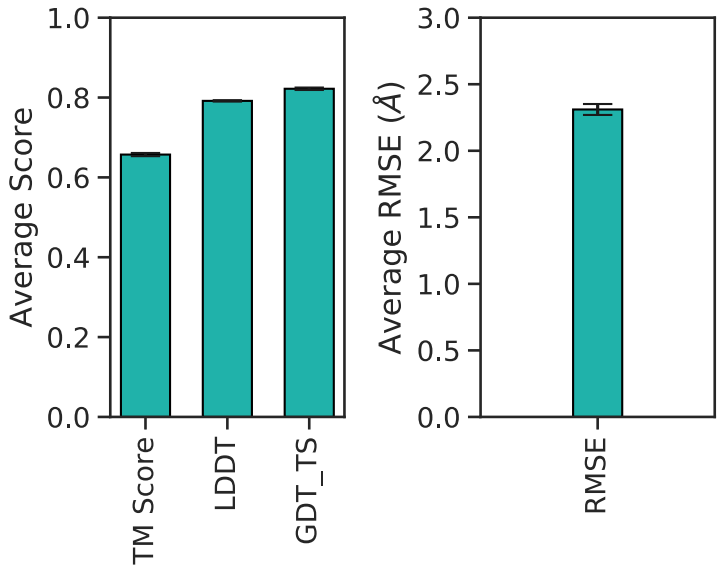


Figure 6.8: Barplot of average 3D accuracy metrics and corresponding 95% confidence intervals per score type.

6.5 Discussion

In this work, we have implemented a custom MLP neural network model approach to reconstruct atom-level representations of amyloid-β from coarse PSNs. Although this neural upscaling method is specific to amyloid-β, the MLP neural network model can be retrained to other biomolecular systems from a variety of different sources (e.g. MD simulation, NMR ensemble, etc.), and thus can be generalizable and adaptable. For any given biomolecular coordinate structure, input (contact adjacency matrices) and output (PIDs) data for neural network retraining can be extracted.

Although previous reverse mapping methods (e.g. random placement, geometric-based, etc.) are able to reconstruct atomistic models, they do so typically from coarse grain force field models (e.g. MARTINI [180]). The advantage of a MLP neural network is the ability to learn and fine-tune parameters specific to the system under investigation from minimal information (contact adjacency matrices) in comparison to coarse grain force fields. The MLP neural

network can thus familiarize itself with a specific target system of interest and coarse grain network simulations [96] can be used to explore these biomolecules.

In the literature, another example of neural networks specifically, variational autoencorders (VAE), have been used primarily on single small molecules and bulk-phase simulations as test cases for reverse mapping [312]. This VAE methodology, although not tested on proteins, could possibly be adapted for such systems, however we are able to demonstrate backmapping with a MLP neural network architecture. To better generalize our neural upscaling technique to protein systems of different sizes, convolutional neural network architectures similar to AlphaFold [251] could be also be incorporated and trained to predict regions (e.g. N x N residue regions).

6.6 Conclusion

Direct predictions of PID metrics demonstrate the predictive capabilities of the MLP neural network to reconstruct all-atom representations of proteins from binary contact adjacency matrices. Example conformations of the best, median and worst PID-based predictions in the test set illustrate the MLP performance. In the worst prediction (frame 7560), the RMSD between the N-terminal halves of the original vs. predicted is quite favorable (0.98 Å). Chirality corrections and conjugate gradient minimization were vital post-prediction processing steps in generating stereochemically reasonable 3D structures. Three-dimensional accuracy metrics, in particular GDT_TS – the main assessment metric in the CASP competition – suggests the neural network performed well given the average values and 95% confidence intervals. In totality, we're able to illustrate the viability of the MLP neural network architecture in this transformation experiment. This work exemplifies neural network-based techniques capable of extracting useful, meaningful data from coarse grained models.

Bibliography

[1] Autonomio talos. `http://github.com/autonomio/talos`, 2019.

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[3] S. Abdelkafi, H. Ogata, N. Barouh, B. Fouquet, and Lebrun. Identification and biochemical characterization of a GDSL-motif carboxylester hydrolase from *Carica papaya* latex. *Biochimica et Biophysica Acta*, 1791:1048–1056, 2009.

[4] J. Abendroth, D. Dranow, D. Lorimer, and T. Edwards. Crystal structure of probable uroporphyrinogen decarboxylase (UPD) (URO-D) from *Pseudomonas aeruginosa*. *To be published*, 2014.

[5] J. Abendroth, J. Fairman, D. Lorimer, and T. Edwards. Structure of uroporphyrinogen decarboxylase from Acinetobacter baumannii. *To be published*, 2015.

[6] M. Abu-Odeh, T. Bar-Mag, H. Huang, T. Kim, Z. Salah, S. K. Abdeen, M. Sudol, D. Reichmann, S. Sidhu, P. M. Kim, and R. I. Aqeilan. Characterizing ww domain interactions of tumor suppressor wwox reveals its association with multiprotein networks. *J Biol Chem*, 289(13):8865–80, 2014.

[7] L. Adamec. Leaf absorption of mineral nutrients in carnivorous plants stimulates root nutrient uptake. *New Phytolologist*, 2155:89–100, 2002.

[8] C. Akoh, G. Lee, Y. Liaw, T. Huang, and J. Shaw. GDSL family of serine esterases/lipases. *Progress in Lipid Research*, 43(6):534–552, 2004.

[9] M. Andersen, A. Jensen, J. Robertus, R. Leah, and K. Skriver. Heterologous expression and characterization of wild-type and mutant forms of a 26 kda endochitinase from barley (hordeum vulgare l.). *Biochemical Journal*, 822:815–822, 1997.

[10] E. Aragon, N. Goerner, Q. Xi, T. Gomes, S. Gao, J. Massague, and M. J. Macias. Structural basis for the versatile interactions of smad7 with regulator ww domains in tgf-beta pathways. *Structure*, 20(10):1726–36, 2012.

[11] S. C. Atkinson, M. D. Audsley, K. G. Lieu, G. A. Marsh, D. R. Thomas, S. M. Heaton, J. J. Paxman, K. M. Wagstaff, A. M. Buckle, G. W. Moseley, et al. Recognition by host nuclear transport proteins drives disorder-to-order transition in hendra virus v. *Scientific reports*, 8(1):1–17, 2018.

[12] A. Babtie, N. Tokuriki, and F. Hollfelder. What makes an enzyme promiscuous? *Current Opinion in Chemical Biology*, 14(2):200–207, 2010.

[13] M. M. Babu, R. van der Lee, N. S. de Groot, and J. Gsponer. Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol*, 21(3):432–40, 2011.

[14] B. Bakan and D. Marion. Assembly of the cutin polyester: From cells to extracellular cell walls. *Plants*, 6(57):doi:10.3390/plants6040057, 2017.

[15] E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol*, 44(2):97–179, 1984.

[16] D. J. Barlow and J. M. Thornton. Ion-pairs in proteins. *J Mol Biol*, 168(4):867–85, 1983.

[17] J. L. Battiste. Structure determination of an hiv-1 rre rnarev peptide complex by nmr spectroscopy. 1996.

[18] J. L. Battiste, H. Mao, N. S. Rao, R. Tan, D. R. Muhandiram, L. E. Kay, A. D. Frankel, and J. R. Williamson. Alpha helix-rna major groove recognition in an hiv-1 rev peptide-rre rna complex. *Science*, 273(5281):1547–51, 1996.

[19] J. J. Beintema. Structural features of plant chitinases and chitin-binding proteins. *FEBS Letters*, 350(2):159–163, 1994.

[20] K. K. Bejagam, S. Singh, Y. An, and S. A. Deshmukh. Machine-learned coarse-grained models. *The journal of physical chemistry letters*, 9(16):4667–4672, 2018.

[21] N. C. Benson and V. Daggett. A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations. *Journal of Bioinformatics and Computational Biology*, 10(04):1250008, 2012.

[22] R. B. Best, N. V. Buchete, and G. Hummer. Are current molecular dynamics force fields too helical? *Biophys J*, 95(1):L07–9, 2008.

[23] R. B. Best, W. Zheng, and J. Mittal. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J Chem Theory Comput*, 10(11):5113–5124, 2014.

[24] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and J. Mackerell, A. D. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput*, 8(9):3257–3273, 2012.

[25] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. Mackerell, Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput*, 8(9):3257–3273, Sep 2012.

[26] E. Bokma, H. Rozeboom, M. Sibbald, B. Dijkstra, and J. Beintema. Expression and characterization of active site mutants of hevamine, a chitinase from the rubber tree *Hevea brasiliensis*. *European Journal of Biochemistry*, 269:893–901, 2002.

[27] L. Boninsegna, G. Gobbo, F. Noé, and C. Clementi. Investigating molecular kinetics by variationally optimized diffusion maps. *Journal of chemical theory and computation*, 11(12):5947–5960, 2015.

[28] P. Bork and M. Sudol. The ww domain: a signalling site in dystrophin? *Trends Biochem Sci*, 19(12):531–3, 1994.

[29] W. M. Botello-Smith and R. Luo. Applications of mmpbsa to membrane proteins i: Efficient numerical solutions of periodic poisson-boltzmann equation. *J Chem Inf Model*, 55(10):2187–99, 2015.

[30] P. Brocos, P. Mendoza-Espinosa, R. Castillo, J. Mas-Oliva, and Á. Pineiro. Multiscale molecular dynamics simulations of micelles: coarse-grain for self-assembly and atomic resolution for finer details. *Soft Matter*, 8(34):9005–9014, 2012.

[31] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a. Swaminathan, and M. Karplus. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217, 1983.

[32] F. Buch, M. Rott, S. Rottloff, C. Paetz, I. Hilke, M. Raessler, and A. Mithöfer. Secreted pitfall-trap fluid of carnivorous nepenthes plants is unsuitable for microbial growth. *Annals of Botany*, 111(375–383), 2013.

[33] G. Busam, H.-H. Kassemeyer, and U. Matern. Differential expression of chitinases in vitis vinifera l. responding to systemic acquired resistance activators or fungal challenge. *Plant Physiol.*, 15:1029–1038, 1997.

[34] C. T. Butts. Network: a Package for Managing Relational Data in R. *Journal of Statistical Software*, 24(2):1–36, 2008.

[35] C. T. Butts. Social Network Analysis with sna. *Journal of Statistical Software*, 24(6):1–51, 2008.

[36] C. T. Butts, J. C. Bierma, and R. W. Martin. Novel proteases from the genome of the carnivorous plant *Drosera capensis*: structural prediction and comparative analysis. *Proteins: Structure, Function, and Bioinformatics*, 84(10):1517–1533, 2016.

[37] C. T. Butts and K. M. Carley. Some Simple Algorithms for Structural Comparison. *Computational and Mathematical Organization Theory*, 11(4):291–305, 2005.

[38] C. T. Butts, X. Zhang, J. E. Kelly, K. W. Roskamp, M. H. Unhelkar, J. A. Freites, S. Tahir, and R. W. Martin. Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Computational and Structural Biotechnology Journal*, 14:271–282, 2016.

[39] Q. Cai, M. J. Hsieh, J. Wang, and R. Luo. Performance of nonlinear finite-difference poisson-boltzmann solvers. *J Chem Theory Comput*, 6(1):203–211, 2010.

[40] M. Campbell, M. Law, C. Holt, J. Stein, G. Moghe, D. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C. J. Lawrence, D. Ware, S. H. Shiu, K. L. Childs, Y. Sun, N. Jiang, , and M. Yandell. MAKER-P: A Tool-kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiology*, 164:513–524, 2013.

[41] M. S. Campbell, C. Holt, B. Moore, and M. Yandell. Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, 48:4.11.1–4.11.39, 2014.

[42] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research*, 37:D233–D238, 2009.

[43] D. Cao, H. Cheng, W. Wu, H. Soo, and J. Peng. Gibberellin mobilizes distinct DELLA-dependent transcriptomes to regulate seed germination and floral development in *Arabidopsis*. *Plant Physiology*, 142:509–525, 2006.

[44] J. J. Cao, X. M. Zhao, D. L. Wang, K. H. Chen, X. Sheng, W. B. Li, M. C. Li, W. J. Liu, and J. He. Yap is overexpressed in clear cell renal cell carcinoma and its knockdown reduces cell proliferation and induces cell cycle arrest and apoptosis. *Oncol Rep*, 32(4):1594–600, 2014.

[45] D. Case, T. Darden, T. Cheatham III, C. Simmerling, J. Wang, R. Duke, R. Luo, M. Crowley, R. Walker, W. Zhang, and K. Merz. Amber 2017 reference manual. *University of California, San Francisco*, 2017.

[46] D. A. Case, r. Cheatham, T. E., T. Darden, H. Gohlke, R. Luo, J. Merz, K. M., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–88, 2005.

[47] D. A. Case and M. Karplus. Dynamics of ligand binding to heme proteins. *Journal of molecular biology*, 132(3):343–368, 1979.

[48] F. Casu, B. M. Duggan, and M. Hennig. The arginine-rich rna-binding motif of hiv-1 rev is intrinsically disordered and folds upon rre binding. *Biophysical journal*, 105(4):1004–1017, 2013.

[49] B. S. Cavada, F. B. B. Moreno, B. A. M. da Rocha, W. F. de Azevedo Jr., R. E. R. Castellón, G. V. Goersch, C. S. Nagano, E. P. de Souza, K. S. Nascimento, G. Radis-Baptista, P. Delatorre, Y. Leroy, M. H. Toyama, V. P. T. Pinto, A. H. Sampaio,

D. Barettino, H. Debray, J. J. Calvete, and L. Sanz. cDNA cloning and 1.75 åcrystal structure determination of PPL2, an endochitinase and N-acetylglucosamine-binding hemagglutinin from *Parkia platycephala* seeds. *FEBS Journal*, 273(17):3962–3974, 2006.

[50] M. Chakraborty, C. Xu, and A. D. White. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. *The Journal of Chemical Physics*, 149(13):134106, 2018.

[51] S. W. Chan, C. J. Lim, Y. F. Chong, A. V. Pobbati, C. Huang, and W. Hong. Hippo pathway-independent restriction of taz and yap by angiomotin. *J Biol Chem*, 286(9):7018–26, 2011.

[52] C. Chang, G. Chhor, J. Bearden, and A. Joachimiak. Crystal structure of lipolytic protein G-D-S-L family from *Alicyclobacillus acidocaldarius* subsp. acidocaldarius DSM 446. *To be published*, 2011.

[53] N. S. Chang, L. J. Hsu, Y. S. Lin, F. J. Lai, and H. M. Sheu. Ww domain-containing oxidoreductase: a candidate tumor suppressor. *Trends Mol Med*, 13(1):12–22, 2007.

[54] G. Charras and A. S. Yap. Tensile forces and mechanotransduction at cell-cell junctions. *Curr Biol*, 28(8):R445–R457, 2018.

[55] S. Chatterjee, A. J. Matas, T. Isaacson, C. Kehlet, J. K. Rose, and R. E. Stark. Solid-state ^{13}C NMR delineates the architectural design of biopolymers in native and genetically altered tomato fruit cuticles. *Biomacromolecules*, 17(1):215–224, 2016.

[56] M. M. Chaudet, T. A. Naumann, N. P. Price, and D. R. Rose. Crystallographic structure of ChitA, a glycoside hydrolase family 19, plant class IV chitinase from *Zea mays*. *Protein Science*, 23(5):586–593, 2014.

[57] H. Chepyshko, C.-P. Lai, L.-M. Huang, J.-H. Liu, and J.-F. Shaw. Multifunctionality and diversity of GDSL esterase/lipase gene family in rice (*Oryza sativa* L. *japonica*) genome: new insights from bioinformatics analysis. *BMC Genomics*, 13:309, 2012.

[58] F. Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[59] K. Clauß, A. Baumert, M. Nimtz, C. Milkowski, and D. Strack. Role of a GDSL lipase-like protein as sinapine esterase in Brassicaceae. *The Plant Journal*, 53(5):802–813, 2008.

[60] G. B. Cohen, R. Ren, and D. Baltimore. Modular binding domains in signal transduction proteins. *Cell*, 80(2):237–48, 1995.

[61] J. L. Cole, J. D. Gehman, J. A. Shafer, and L. C. Kuo. Solution oligomerization of the rev protein of hiv-1: implications for function. *Biochemistry*, 32(44):11769–75, 1993.

[62] F. Colonna-Cesari, D. Perahia, M. Karplus, H. Eklund, C. Brädén, and O. Tapia. Interdomain motion in liver alcohol dehydrogenase. structural and energetic analysis of the hinge bending mode. *Journal of Biological Chemistry*, 261(32):15273–15280, 1986.

[63] A. L. Couzens, J. D. Knight, M. J. Kean, G. Teo, A. Weiss, W. H. Dunham, Z. Y. Lin, R. D. Bagshaw, F. Sicheri, T. Pawson, J. L. Wrana, H. Choi, and A. C. Gingras. Protein interaction network of the mammalian hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci Signal*, 6(302):rs15, 2013.

[64] S. A. Dames, R. Aregger, N. Vajpai, P. Bernado, M. Blackledge, and S. Grzesiek. Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. *J Am Chem Soc*, 128(41):13508–14, 2006.

[65] H. Darabi, J. Beesley, A. Droit, S. Kar, S. Nord, M. Moradi Marjaneh, P. Soucy, K. Michailidou, M. Ghoussaini, H. Fues Wahl, M. K. Bolla, Q. Wang, J. Dennis, M. R. Alonso, I. L. Andrulis, H. Anton-Culver, V. Arndt, M. W. Beckmann, J. Benitez, N. V. Bogdanova, S. E. Bojesen, H. Brauch, H. Brenner, A. Broeks, T. Bruning, B. Burwinkel, J. Chang-Claude, J. Y. Choi, D. M. Conroy, F. J. Couch, A. Cox, S. S. Cross, K. Czene, P. Devilee, T. Dork, D. F. Easton, P. A. Fasching, J. Figueroa, O. Fletcher, H. Flyger, E. Galle, M. Garcia-Closas, G. G. Giles, M. S. Goldberg, A. Gonzalez-Neira, P. Guenel, C. A. Haiman, E. Hallberg, U. Hamann, M. Hartman, A. Hollestelle, J. L. Hopper, H. Ito, A. Jakubowska, N. Johnson, D. Kang, S. Khan, V. M. Kosma, M. Kriege, V. Kristensen, D. Lambrechts, L. Le Marchand, S. C. Lee, A. Lindblom, A. Lophatananon, J. Lubinski, A. Mannermaa, S. Manoukian, S. Margolin, K. Matsuo, R. Mayes, J. McKay, A. Meindl, R. L. Milne, K. Muir, S. L. Neuhausen, H. Nevanlinna, C. Olswold, N. Orr, P. Peterlongo, G. Pita, K. Pylkas, A. Rudolph, S. Sangrajrang, E. J. Sawyer, M. K. Schmidt, R. K. Schmutzler, C. Seynaeve, M. Shah, C. Y. Shen, X. O. Shu, M. C. Southey, D. O. Stram, H. Surowy, A. Swerdlow, S. H. Teo, D. C. Tessier, I. Tomlinson, D. Torres, T. Truong, et al. Fine scale mapping of the 17q22 breast cancer locus using dense snps, genotyped within the collaborative oncological gene-environment study (cogs). *Sci Rep*, 6:32512, 2016.

[66] S. Das and T. F. Smith. Identifying nature's protein lego set. *Adv Protein Chem*, 54:159–83, 2000.

[67] M. D. Daugherty, D. S. Booth, B. Jayaraman, Y. Cheng, and A. D. Frankel. Hiv rev response element (rre) directs assembly of the rev homooligomer into discrete asymmetric complexes. *Proc Natl Acad Sci U S A*, 107(28):12481–6, 2010.

[68] M. D. Daugherty, I. D'Orso, and A. D. Frankel. A solution to limited genomic capacity: using adaptable binding surfaces to assemble the functional hiv rev oligomer on rna. *Mol Cell*, 31(6):824–34, 2008.

[69] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[70] C. Dominguez, R. Boelens, and A. M. Bonvin. Haddock: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7):1731–7, 2003.

[71] Dong, Xiangshu and Yi, Hankuil and Han, Ching Tack and Nou, Ill Sup and Hur, Yoonkang. GDSL esterase/lipase genes in *Brassica rapa* L.: Genome-wide identification and expression analysis. *Molecular Genetics and Genomics*, 291:531–542, 2016.

[72] R. O. Duda and P. E. Hart. Pattern classification and scene analysis. *A Wiley-Interscience Publication, New York: Wiley, 1973*, 1973.

[73] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002.

[74] A. K. Dunker, C. J. Brown, and Z. Obradovic. Identification and functions of usefully disordered proteins. *Advances in protein chemistry*, 62:25–49, 2002.

[75] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59, 2001.

[76] A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner, and C. J. Brown. Intrinsic protein disorder in complete genomes. *Genome informatics*, 11:161–171, 2000.

[77] S. Dutta, S. Mana-Capelli, M. Paramasivam, I. Dasgupta, H. Cirka, K. Billiar, and D. McCollum. Trip6 inhibits hippo signaling in response to tension at adherens junctions. *EMBO Rep*, 19(2):337–350, 2018.

[78] H. Ebata, K. Toshima, and S. Matsumura. Lipase-catalyzed synthesis and curing of high-molecular-weight polyricinoleate. *Macromolecular Bioscience*, 7:798–803, 2007.

[79] H. Eilenberg, S. Pnini-Cohen, S. Schuster, A. Movtchan, and A. Zilberstein. Isolation and characterization of chitinase genes from pitchers of the carnivorous plant *Nepenthes khasiana*. *Journal of Experimental Botany*, 57:2775–2784, 2006.

[80] A. El Moussaoui, M. Nijs, R. Paul, C.and Wintjens, J. Vincentelli, M. Azarkan, and Y. Looze. Revisiting the enzymes stored in the laticifers of *Carica papaya* in the context of their possible participation in the plant defence mechanism. *Cellular and Molecular Life Sciences*, 58:556–570, 2001.

[81] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–14, 2007.

[82] Epanechnikov and V. A. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14(1):153–158, 1969.

[83] J. M. Ervasti. Dystrophin, its interactions with other proteins, and implications for muscular dystrophy. *Biochim Biophys Acta*, 1772(2):108–17, 2007.

[84] X. Espanel and M. Sudol. Yes-associated protein and p53-binding protein-2 interact through their ww and sh3 domains. *J Biol Chem*, 276(17):14514–23, 2001.

[85] P. W. Faber, G. T. Barnes, J. Srinidhi, J. Chen, J. F. Gusella, and M. E. MacDonald. Huntingtin interacts with a family of ww domain proteins. *Hum Mol Genet*, 7(9):1463–74, 1998.

[86] X. Feng, P. Liu, X. Zhou, M.-T. Li, F.-L. Li, Z. Wang, Z. Meng, Y.-P. Sun, Y. Yu, Y. Xiong, H.-X. Yuan, and K.-L. Guan. Thromboxane a2 activates yap/taz protein to induce vascular smooth muscle cell proliferation and migration. *Journal of Biological Chemistry*, 291(36):18947–18958, 2016.

[87] O. Ferrigno, F. Lallemand, F. Verrecchia, S. L'Hoste, J. Camonis, A. Atfi, and A. Mauviel. Yes-associated protein (yap65) interacts with smad7 and potentiates its inhibitory activity against tgf-beta/smad signaling. *Oncogene*, 21(32):4879–84, 2002.

[88] A. Fiser, R. K. Do, and A. Sali. Modeling of loops in protein structures. *Protein Sci*, 9(9):1753–73, 2000.

[89] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophys J*, 94(10):L75–7, 2008.

[90] C. A. Galea, Y. Wang, S. G. Sivakolundu, and R. W. Kriwacki. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry*, 47(29):7598–609, 2008.

[91] A. E. Garcia and K. Y. Sanbonmatsu. Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc Natl Acad Sci U S A*, 99(5):2782–7, 2002.

[92] G. Garcia-Casado, C. Carmen, I. Allona, R. Casado, L. Pacios, C. Aragoncillo, and L. Gomez. Site-directed mutagenesis of active site residues in a class i endochitinase from chestnut seeds. *Glycobiology*, 8(10):1021–1028, 1998.

[93] A.-L. Girard, F. Mounet, M. Lemaire-Chamley, C. Gaillard, K. Elmorjani, J. Vivancos, J.-L. Runavot, B. Quemener, J. Petit, V. Germain, C. Rothan, D. Marion, and B. Bakana. Tomato GDSL1 is required for cutin deposition in the fruit cuticle. *Plant Cell*, 24(7):3119–3134, 2012.

[94] J. Godlewski, J. Kiezun, B. E. Krazinski, Z. Kozielec, P. M. Wierzbicki, and Z. Kmiec. The immunoexpression of yap1 and lats1 proteins in clear cell renal cell carcinoma: Impact on patients' survival. *Biomed Res Int*, 2018:2653623, 2018.

[95] S. M. Gopal, S. Mukherjee, Y.-M. Cheng, and M. Feig. Primo/primona: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins: Structure, Function, and Bioinformatics*, 78(5):1266–1281, 2010.

[96] G. Grazioli, Y. Yu, M. H. Unhelkar, R. W. Martin, and C. T. Butts. Network-based classification and modeling of amyloid fibrils. *The Journal of Physical Chemistry B*, 123(26):5452–5462, 2019.

[97] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*, 322(5906):1365–8, 2008.

[98] Y. Gu, D.-W. Li, and R. Brüschweiler. Decoding the mobility and time scales of protein loops. *Journal of Chemical Theory and Computation*, 11(3):1308–1314, 2015.

[99] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, and T. Schwede. The Protein Model Portal - a comprehensive resource for protein structure and model information. Database (PMID: 23624946), 2013.

[100] M. Hahn, M. Hennig, B. Schlesier, and W. Höhne. Structure of jack bean chitinase. *Acta Crystallographica D: Biological Crystallography*, 56(9):1096–1099, 2000.

[101] G. Halder and R. L. Johnson. Hippo signaling: growth control and beyond. *Development*, 138(1):9–22, 2011.

[102] M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris. statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software*, 24(1):1–11, 2008.

[103] J. H. Hansson, L. Schild, Y. Lu, T. A. Wilson, I. Gautschi, R. Shimkets, C. Nelson-Williams, B. C. Rossier, and R. P. Lifton. A de novo missense mutation of the beta subunit of the epithelial sodium channel causes hypertension and liddle syndrome, identifying a proline-rich segment critical for regulation of channel activity. *Proc Natl Acad Sci U S A*, 92(25):11495–9, 1995.

[104] Y. Hao, A. Chun, K. Cheung, B. Rashidi, and X. Yang. Tumor suppressor lats1 is a negative regulator of oncogene yap. *J Biol Chem*, 283(9):5496–509, 2008.

[105] P. J. Hart, H. D. Pfluger, A. F. Monzingo, T. Hollis, and J. D. Robertus. The refined crystal structure of an endochitinase from *Hordeum vulgare L.* seeds at 1.8åresolution-tructure of an endochitinase from *Hordeum vulgare L.* seeds at 1.8åresolution. *Journal of Molecular Biology*, 248:402, 1995.

[106] S. C. Harvey, M. Prabhakaran, B. Mao, and J. A. McCammon. Phenylalanine transfer rna: molecular dynamics simulation. *Science*, 223(4641):1189–1191, 1984.

[107] J. W. Haskins, D. X. Nguyen, and D. F. Stern. Neuregulin 1-activated erbb4 interacts with yap to induce hippo pathway target genes and promote cell migration. *Sci Signal*, 7(355):ra116, 2014.

[108] N. Hatano and T. Hamada. Proteomic analysis of secreted protein induced by a component of prey in pitcher fluid of the carnivorous plant *Nepenthes alata*. *Journal of Proteomics*, 75:4844–4852, 2012.

[109] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. A. Kern. Hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, 2007.

[110] B. Hess, S. León, N. Van Der Vegt, and K. Kremer. Long time atomistic polymer trajectories from coarse grained simulations: bisphenol-a polycarbonate. *Soft Matter*, 2(5):409–414, 2006.

[111] J. Hong, H. Choi, I. Hwang, D. Kim, N. Kim, d. Choi, Y. Kim, and B. Hwang. Function of a novel GDSL-type pepper lipase gene, CaGLIP1, in disease susceptibility and abiotic stress tolerance. *Planta*, 227:539–558, 2008.

[112] S. J. Horn, P. Sikorski, J. B. Cederkvist, G. Vaaje-Kolstad, M. Sórlie, B. Synstad, G. Vriend, K. M. Vårum, and V. G. H. Eijsink. Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides. *Proceedings of the National Academy of Sciences of the United States of America*, 103(48):18089–18094, 2006.

[113] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–25, 2006.

[114] H. Hu, J. Columbus, Y. Zhang, D. Wu, L. Lian, S. Yang, J. Goodwin, C. Luczak, M. Carter, L. Chen, M. James, R. Davis, M. Sudol, J. Rodwell, and J. J. Herrero. A map of ww domain family interactions. *Proteomics*, 4(3):643–55, 2004.

[115] H. Hu and H. Du. α-to-β structural transformation of ovalbumin: heat and ph effects. *Journal of Protein Chemistry*, 19(3):177–183, 2000.

[116] Z. Hua, C. Zou, S. H. Shiu, and R. Vierstra. Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. *PLoS One*, 6:e16219, 2011.

[117] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmuller, and J. MacKerell, A. D. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*, 14(1):71–73, 2017.

[118] J. Huet, P. Rucktooa, B. Clantin, M. Azarkan, Y. Looze, V. Villeret, and R. Wintjens. X-ray structure of papaya chitinase reveals the substrate binding mode of glycosyl hydrolase family 19 chitinases. *Biochemistry*, 47:8283–8291, 2008.

[119] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 27–28, Feb. 1996.

[120] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, Feb 1996.

[121] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, 9(3):90–95, 2007.

[122] E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi, and J. W. Harper. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 2017.

[123] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, and A. K. Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*, 323(3):573–84, 2002.

[124] T. Isaacson, D. K. Kosma, A. J. Matas, G. J. Buda, Y. He, B. Yu, A. Pravitasari, J. D. Batteas, R. E. Stark, M. A. Jenks, and J. K. C. Rose. Cutin deficiency in the tomato fruit cuticle consistently affects resistance to microbial infection and biomechanical properties, but not transpirational water loss. *The Plant Journal*, 60:363–377, 2009.

[125] B. Iseli, S. Armand, T. Boller, J. Neuhaus, and B. Henrissat. Plant chitinases use two different hydrolytic mechanisms. *FEBS Letters*, 382:186–188, 1996.

[126] K. Ishisaki, Y. Honda, H. Taniguchi, N. Hatano, and T. Hamada. Heterogonous expression and characterization of a plant class IV chitinase from the pitcher of the carnivorous plant *Nepenthes alata*. *Glycobiology*, 22(3):345–351, 2012.

[127] L. Jiang, N. Kon, T. Li, S.-J. Wang, T. Su, H. Hibshoosh, R. Baer, and W. Gu. Ferroptosis as a p53-mediated activity during tumour suppression. *Nature*, 520(7545):57–62, 2015.

[128] D. T. Jones and D. Cozzetto. Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31(6):857–863, 2015.

[129] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[130] B. Juniper, R. Robins, and D. Joel. *The Carnivorous Plants*. Academic Press, London, UK, 1989.

[131] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[132] S. Karmakar, K. Molla, P. Chanda, S. Sarkar, S. Datta, and K. Datta. Green tissue-specific co-expression of chitinase and oxalate oxidase 4 genes in rice for enhanced resistance against sheath blight. *Planta*, 243(1):115–130, 2016.

[133] F. Kaschani, M. Shabab, T. Bozkurt, T. Shindo, S. Schornack, C. Gu, M. Ilyas, J. Win, S. Kamoun, and R. van der Hoorn. An effector-targeted protease contributes to defense against *Phytophthora infestans* and is under diversifying selection in natural hosts. *Plant Physiology*, 154(4):1794–1804, 2010.

[134] P. Kesari, D. N. Patil, P. Kumar, S. Tomar, A. K. Sharma, and P. Kumar. Structural and functional evolution of chitinase-like proteins from plants. *PROTEOMICS*, 15(10):1693–1705, 2015.

[135] Y. Kezuka, M. Kojima, R. Mizuno, K. Suzuki, T. Watanabe, and N. T. Structure of full-length class i chitinase from rice revealed by x-ray crystallography and small-angle x-ray scattering. *Proteins*, 78(10):2295–2305, 2010.

[136] F. Khoushab and M. Yamabhai. Chitin research revisited. *Marine Drugs*, 8(7):1988–2012, 2010.

[137] Y. Kikkawa, M. Fukuda, A. Kashiwada, K. Matsuda, M. Kanesato, M. Wada, T. Imanaka, and T. Tanaka. Binding ability of chitinase onto cellulose: an atomic force microscopy study. *Polymer Journal*, 43:742–744, 2011.

[138] Y. Kikuta, H. Ueda, M. Takahashi, T. Mitsumori, G. Yamada, K. Sakamori, K. Takeda, S. Furutani, K. Nakayama, and Y. Katsuda. Identification and characterization of a gdsl lipase-like protein that catalyzes the ester-forming reaction for pyrethrin biosynthesis in tanacetum cinerariifolium–a new target for plant protection. *The Plant Journal*, 71(2):183–193, 2012.

[139] D. Kim, D. Chivian, and D. Baker. Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Research*, 32(Supplement 2):W526–31, 2004.

[140] Y. Kitaoku, N. Umemoto, T. Ohnuma, T. Numata, T. Taira, S. Sakuda, and T. Fukamizo. A class III chitinase without disulfide bonds from the fern, *Pteris ryukyuensis*: crystal structure and ligand-binding studies. *Planta*, 242:895–907, 2015.

[141] S. Kobayashi and A. Makino. Enzymatic polymer synthesis: An opportunity for green polymer chemistry. *Chemical Reviews*, 109(11):5288–5353, 2009.

[142] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020, 2019.

[143] S. Kumar, N. Singh, B. Mishra, D. Dube, M. Sinha, S. B. Singh, S. Dey, P. Kaur, S. Sharma, and T. P. Singh. Modulation of inhibitory activity of xylanase -α-amylase inhibitor protein (XAIP): binding studies and crystal structure determination of XAIP-II from *Scadoxus multiflorus* at 1.2 åresolution. *BMC Structural Biology*, 10:41, 2010.

[144] S. Kumar, N. Singh, M. Sinha, S. Singh, A. Bhushan, P. Kaur, A. Srinivasan, S. Sharma, and T. Singh. Crystal structure of haementhin from *Haemanthus multiflorus* at 2.0 åresolution: Formation of a novel loop on a TIM barrel fold and its functional significance. *To be published*, 2009.

[145] C. P. Lai, L. M. Huang, L. F. O. Chen, M. T. Chan, and J. F. Shaw. Genome-wide analysis of GDSL-type esterases/lipases in *Arabidopsis*. *Plant Molecular Biology*, 95:181–197, 2017.

[146] S. Lansky, O. Alalouf, V. Solomon, A. Alhassid, H. Belrahli, L. Govada, N. Chayan, Y. Shoham, and G. Shoham. An Axe2 mutant (W190I), an acetyl-xylooligosaccharide esterase from *Geobacillus stearothermophilus*. *To be published*, 2014.

[147] S. Lansky, O. Alalouf, V. Solomon, A. Alhassid, H. Belrahli, L. Govada, N. Chayan, Y. Shoham, and G. Shoham. Crystal structure of a catalytic mutant of Axe2 (Axe2-H194A), an acetylxylan esterase from *Geobacillus stearothermophilus*. *To be published*, 2014.

[148] S. Lansky, O. Alalouf, V. Solomon, A. Alhassid, H. Belrahli, L. Govada, N. Chayan, Y. Shoham, and G. Shoham. Crystal Structure of Axe2, an Acetylxylan Esterase from *Geobacillus stearothermophilus*. *Acta Crystallographica, Section D: Biological Crystallography*, 70:261–278, 2014.

[149] Y.-L. Lee, J. C. Chen, and J.-F. Shaw. The thioesterase I of *Escherichia coli* has arylesterase activity and shows stereospecificity for protease substrates. *Biochemical and Biophysical Research Communications*, 231(2):452–456, 1997.

[150] R. Leinonen, F. G. Diez, D. Binns, W. Fleischmann, R. Lopez, and R. Apweiler. Uniprot archive. *Bioinformatics*, 20(17):3236–3237, 2004.

[151] T. Lemke and C. Peter. Neural network based prediction of conformational free energies-a new route toward coarse-grained simulation models. *Journal of chemical theory and computation*, 13(12):6213–6221, 2017.

[152] H. T. Leung, O. Bignucolo, R. Aregger, S. A. Dames, A. Mazur, S. Berneche, and S. Grzesiek. A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. *J Chem Theory Comput*, 12(1):383–94, 2016.

[153] X. Li, K. M. Tran, K. E. Aziz, A. V. Sorokin, J. Chen, and W. Wang. Defining the protein-protein interaction network of the human protein tyrosine phosphatase family. *Mol Cell Proteomics*, 15(9):3030–44, 2016.

[154] Z. Lin, Z. Yang, R. Xie, Z. Ji, K. Guan, and M. Zhang. Decoding ww domain tandem-mediated target recognitions in tissue growth and cell polarity. *Elife*, 8, 2019.

[155] F. Liu, B. Li, E. J. Tung, I. Grundke-Iqbal, K. Iqbal, and C. X. Gong. Site-specific effects of tau phosphorylation on its microtubule assembly activity and self-aggregation. *Eur J Neurosci*, 26(12):3429–36, 2007.

[156] H. Liu, X. Dai, X. Cao, H. Yan, X. Ji, H. Zhang, S. Shen, Y. Si, H. Zhang, J. Chen, L. Li, J. C. Zhao, J. Yu, X. H. Feng, and B. Zhao. Prdm4 mediates yap-induced cell invasion by activating leukocyte-specific integrin beta2 expression. *EMBO Rep*, 19(6), 2018.

[157] H. Liu, D. Song, H. Lu, R. Luo, and H. F. Chen. Intrinsically disordered protein-specific force field charmm36idpsff. *Chem Biol Drug Des*, 2018.

[158] J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, and A. K. Dunker. Intrinsic disorder in transcription factors. *Biochemistry*, 45(22):6873–88, 2006.

[159] X. Liu, N. Yang, S. A. Figel, K. E. Wilson, C. D. Morrison, I. H. Gelman, and J. Zhang. Ptpn14 interacts with and negatively regulates the oncogenic function of yap. *Oncogene*, 32(10):1266–73, 2013.

[160] Y.-C. Lo, S.-C. Lin, J.-F. Shaw, and Y.-C. Liaw. Crystal structure of *Escherichia coli* Thioesterase I/Protease I/Lysophospholipase L1: Consensus sequence blocks constitute the catalytic center of SGNH-hydrolases through a conserved hydrogen bond network. *Journal of Molecular Biology*, 330:539–551, 2003.

[161] D. E. Lockhart, A. Schuettelkopf, D. E. Blair, and v. D. M.F. Screening-based discovery of *Aspergillus fumigatus* plant-type chitinase inhibitors. *FEBS Letters*, 588(17):3282–3290, 2014.

[162] V. Lombard, H. G. Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, 42:D490–D495, 2014.

[163] M. Louhivuori, K. Fredriksson, K. Paakkonen, P. Permi, and A. Annila. Alignment of chain-like molecules. *J Biomol NMR*, 29(4):517–24, 2004.

[164] B. C. Low, C. Q. Pan, G. V. Shivashankar, A. Bershadsky, M. Sudol, and M. Sheetz. Yap/taz as mechanosensors and mechanotransducers in regulating organ size and tumor growth. *FEBS Lett*, 588(16):2663–70, 2014.

[165] H. Lubs, F. E. Abidi, R. Echeverri, L. Holloway, A. Meindl, R. E. Stevenson, and C. E. Schwartz. Golabi-ito-hall syndrome results from a missense mutation in the ww domain of the pqbp1 gene. *J Med Genet*, 43(6):e30, 2006.

[166] S. Z. Lukas Hartl and V. Seidl-Seiboth. Fungal chitinases: diversity, mechanistic properties and biotechnological potential. *Appl Microbiol Biotechnol*, pages 533 – 543, 2011.

[167] J. Ma, Q. Lu, Y. Yuan, H. Ge, K. Li, W. Zhao, Y. Gao, L. Niu, and M. Teng. Crystal structure of isoamyl acetate-hydrolyzing esterase from *Saccharomyces cerevisiae* reveals a novel active site architecture and the basis of substrate specificity. *Proteins*, 79:662–668, 2011.

[168] M. R. Machado and S. Pantano. Sirah tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics*, 32(10):1568–1570, 2016.

[169] I. Mack, A. Hector, M. Ballbach, J. Kohlhäufl, K. J. Fuchs, A. Weber, M. A. Mall, and D. Hartl. The role of chitin, chitinases, and chitinase-like proteins in pediatric lung diseases. *Molecular and Cellular Pediatrics*, 2(3):DOI 10.1186/s40348–015–0014–6, 2015.

[170] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102(18):3586–616, 1998.

[171] J. Mackerell, A. D., M. Feig, and r. Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*, 25(11):1400–15, 2004.

[172] J. MacKerell, A. D., M. Feig, and r. Brooks, C. L. Improved treatment of the protein backbone in empirical force fields. *J Am Chem Soc*, 126(3):698–9, 2004.

[173] J. Madhuprakash, A. Singh, S. Kumar, M. Sinha, P. Kaur, S. Sharma, A. R. Podile, and T. P. Singh. Structure of chitinase d from serratia proteamaculans reveals the structural basis of its dual action of hydrolysis and transglycosylation. *International journal of biochemistry and molecular biology*, 4(4):166, 2013.

[174] A. Mahapatro, A. Kumar, and R. Gross. Mild, solvent-free omega-hydroxy acid polycondensations catalyzed by *Candida antarctica* lipase B. *Biomacromolecules*, 5:62–68, 2004.

[175] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J Chem Theory Comput*, 11(8):3696–713, 2015.

[176] M. H. Malim and B. R. Cullen. Hiv-1 structural gene expression requires the binding of multiple rev monomers to the viral rre: implications for hiv-1 latency. *Cell*, 65(2):241–8, 1991.

[177] E. M. Mandelkow and E. Mandelkow. Tau in alzheimer's disease. *Trends Cell Biol*, 8(11):425–7, 1998.

[178] D. A. Mann, I. Mikaelian, R. W. Zemmel, S. M. Green, A. D. Lowe, T. Kimura, M. Singh, P. J. Butler, M. J. Gait, and J. Karn. A molecular rheostat. co-operative rev binding to stem i of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. *J Mol Biol*, 241(2):193–207, 1994.

[179] V. Mariani, M. Biasini, A. Barbato, and T. Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[180] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries. The martini force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry B*, 111(27):7812–7824, 2007.

[181] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

[182] S. Martinez-Caballero, P. Cano-Sanchez, I. Mares-Mejia, A. Diaz-Sanchez, M. Macias-Rubalcava, J. Hermoso, and A. Rodriguez-Romero. Comparative study of two GH19 chitinase-like proteins from *Hevea brasiliensis*, one exhibiting a novel carbohydrate-binding domain. *FEBS Journal*, 281:4535–4554, 2014.

[183] T. Masuda, G. Zhao, and B. Mikami. Crystal structure of class III chitinase from pomegranate provides the insight into its metal storage capacity. *Bioscience, Biotechnology, and Biochemistry*, 79(1):45–50, 2015.

[184] I. Mathews, M. Soltis, M. Saldajeno, G. Ganshaw, R. Sala, W. Weyler, M. A. Cervin, G. Whited, and R. Bott. Structure of a novel enzyme that catalyzes acyl transfer to alcohols in aqueous conditions. *Biochemistry*, 46(31):8969–8979, 2007.

[185] I. Matusíková, J. Salaj, J. Moravcíková, L. Mlynárová, J. Nap, and J. Libantová. Tentacles of in vitro-grown round-leaf sundew (*Drosera rotundifolia* L.) show induction of chitinase activity upon mimicking the presence of prey. *Planta*, 222:1020–1027, 2005.

[186] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[187] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande. Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal*, 109(8):1528–1532, 2015.

[188] F. Meins, B. Fritig, H. Linthorst, J. Mikkelsen, J. Neuhaus, and J. Ryals. Plant chitinase genes. *Plant Molecular Biology Reporter*, 12(2):S22–S28, 1994.

[189] H. Merzendorfer. The cellular basis of chitin synthesis in fungi and insects: Common principles and differences. *European Journal of Cell Biology*, 90(9):759 – 769, 2011.

[190] C. Michaloglou, W. Lehmann, T. Martin, C. Delaunay, A. Hueber, L. Barys, H. Niu, E. Billy, M. Wartmann, M. Ito, C. J. Wilson, M. E. Digan, A. Bauer, H. Voshol, G. Christofori, W. R. Sellers, F. Hofmann, and T. Schmelzle. The tyrosine phosphatase ptpn14 is a negative regulator of yap activity. *PLoS One*, 8(4):e61916, 2013.

[191] r. Miller, B.R., J. McGee, T. D., J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. Mmpbsa.py: An efficient program for end-state free energy calculations. *J Chem Theory Comput*, 8(9):3314–21, 2012.

[192] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky. Analysis of molecular recognition features (morfs). *Journal of molecular biology*, 362(5):1043–1059, 2006.

[193] R. A. Mosher, W. E. Durrant, D. Wang, J. Song, and X. Dong. A comprehensive structure-function analysis of *Arabidopsis* SNI1 defines essential regions and transcriptional repressor activity. *The Plant Cell*, 18:1750–1765, 2006.

[194] K. Murayama, K. Kano, Y. Matsumoto, and D. Sugimori. Crystal structure of phospholipase A1 from *Streptomyces albidoflavus* NA297. *Journal of Structural Biology*, 182:192–196, 2013.

[195] I. Na, D. Redmon, M. Kopa, Y. Qin, B. Xue, and V. N. Uversky. Ordered disorder of the astrocytic dystrophin-associated protein complex in the norm and pathology. *PloS one*, 8(8), 2013.

[196] M. Naranjo, J. Forment, M. Roldan, R. Serrano, and O. Vicente. Overexpression of *Arabidopsis thaliana* LTL1, a salt-induced gene encoding a GDSL-motif lipase, increases salt tolerance in yeast and transgenic plants. *Plant, Cell & Environment*, 29:1890–1900, 2006.

[197] A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, 4(10):1419–40, 2005.

[198] J. Neuhaus, B. Fritig, H. Linthorst, F. Meins, J. Mikkelsen, and J. Ryals. A revised nomenclature for chitinase genes. *Plant Mol Biol Rep*, 14:102–104, 1996.

[199] T. Ohnuma, T. Numata, T. Osawa, H. Inanaga, Y. Okazaki, S. Shinya, K. Kondo, T. Fukuda, and T. Fukamizo. Crystal structure and chitin oligosaccharide-binding mode of a 'loopful' family GH19 chitinase from rye, *Secale cereale*, seeds. *FEBS Journal*, 279:3639–3651, 2012.

[200] M. H. Olsson, C. R. Sondergaard, M. Rostkowski, and J. H. Jensen. PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, 2011.

[201] Y. Otaka, S. Rokudai, K. Kaira, M. Fujieda, I. Horikoshi, R. Iwakawa-Kawabata, S. Yoshiyama, T. Yokobori, Y. Ohtaki, K. Shimizu, T. Oyama, J. Tamura, C. Prives, and M. Nishiyama. Stxbp4 drives tumor growth and is associated with poor prognosis through pdgf receptor signaling in lung squamous cell carcinoma. *Clin Cancer Res*, 23(13):3442–3452, 2017.

[202] D. Pan. The hippo signaling pathway in development and cancer. *Dev Cell*, 19(4):491–505, 2010.

[203] D. Panikashvili, J. Shi, S. Bocobza, R. Franke, L. Schreiber, and A. Aharoni. The *Arabidopsis* DSO/ABCG11 transporter affects cutin metabolism in reproductive organs and suberin in roots. *Molecular Plant*, 3:563–575, 2010.

[204] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio, and R. Nussinov. The role of protein loops and linkers in conformational dynamics and allostery. *Chemical Reviews*, 116:6391–6423, 2016.

[205] L. A. Passani, M. T. Bedford, P. W. Faber, K. M. McGinnis, A. H. Sharp, J. F. Gusella, J. P. Vonsattel, and M. E. MacDonald. Huntingtin's ww domain partners in huntington's disease post-mortem brain fulfill genetic criteria for direct involvement in huntington's disease pathogenesis. *Hum Mol Genet*, 9(14):2175–82, 2000.

[206] P. Paszota, M. Escalante-Perez, L. R. Thomsen, M. W. Risør, A. Dembski, L. Sanglas, T. A. Nielsen, H. Karring, I. B. Thøgersen, R. Hedrich, J. J. Enghild, I. Kreuzer, and K. W. Sanggaard. Secreted major Venus flytrap chitinase enables digestion of arthropod prey. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(2):374 – 383, 2014.

[207] A. Pavlovic, M. Krausko, and L. Adamec. A carnivorous sundew plant prefers protein over chitin as a source of nitrogen from its traps. *Plant Physiology and Biochemistry*, 104:11–16, 2016.

[208] A. Pavlovic, M. Krausko, M. Libiakova, and L. Adamec. Feeding on prey increases photosynthetic efficiency in the carnivorous sundew *Drosera capensis*. *Ann. Bot.*, 113:69–78, 2014.

[209] T. Pawson and J. D. Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–80, 1997.

[210] C. M. Payne, J. Baban, S. J. Horn, P. H. Backe, A. S. Arvai, B. Dalhus, M. Bjørås, V. G. H. Eijsink, M. Sørlie, G. T. Beckham, and G. Vaaje-Kolstad. Hallmarks of processivity in glycoside hydrolases from crystallographic and computational studies of the serratia marcescens chitinases. *Journal of Biological Chemistry*, 287(43):36322–36330, 2012.

[211] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[212] Y. Y. Peng, V. Glattauer, J. A. Ramshaw, and J. A. Werkmeister. Evaluation of the immunogenicity and cell compatibility of avian collagen for biomedical applications. *Journal of Biomedical Materials Research Part A: An Official Journal of The Society for Biomaterials, The Japanese Society for Biomaterials, and The Australian Society for Biomaterials and the Korean Society for Biomaterials*, 93(4):1235–1244, 2010.

[213] Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky, and L. Kurgan. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cellular and Molecular Life Sciences*, 72(1):137–151, 2015.

[214] H. Pereira, P. Castro-Landin, J. Brandao-Neto, and T. Grangeiro. Crystal structure of chitinase (GH19) from *Vigna unguiculata*. *To be published*, 2015.

[215] C. Peter and K. Kremer. Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back. *Soft Matter*, 5(22):4357–4366, 2009.

[216] T. Petersen, S. Brunak, G. von Heijne, and H. Henrik Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8:785–786, 2011.

[217] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[218] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802, 2005.

[219] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, Dec. 2005.

[220] M. Pollard, B. F., Y. Li, and J. Ohlrogge. Building lipid barriers: biosynthesis of cutin and suberin. *Trends in Plant Science*, 13:236–246, 2008.

[221] S. J. Pond, W. K. Ridgeway, R. Robertson, J. Wang, and D. P. Millar. Hiv-1 rev protein assembles on viral rna one molecule at a time. *Proc Natl Acad Sci U S A*, 106(5):1404–8, 2009.

[222] P. Purushotham, P. V. P. S. Arun, J. S. S. Prakash, and A. R. Podile. Chitin binding proteins act synergistically with chitinases in *Serratia proteamaculans* 568. *PLoS ONE*, 7(5):e36714, 2012.

[223] Y. Qiao, J. Chen, Y. B. Lim, M. L. Finch-Edmondson, V. P. Seshachalam, L. Qin, T. Jiang, B. C. Low, H. Singh, C. T. Lim, and M. Sudol. Yap regulates actin dynamics through arhgap29 and promotes metastasis. *Cell Rep*, 19(8):1495–1502, 2017.

[224] Y. Qiao, S. J. Lin, Y. Chen, D. C. Voon, F. Zhu, L. S. Chuang, T. Wang, P. Tan, S. C. Lee, K. G. Yeoh, M. Sudol, and Y. Ito. Runx3 is a novel negative regulator of oncogenic tead-yap complex in gastric cancer. *Oncogene*, 35(20):2664–74, 2016.

[225] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. InterProScan: Protein Domains Identifier. *Nucleic Acids Research*, 33:W116–W120, 2005.

[226] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[227] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[228] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin, and D. Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, 77(Suppl 9):89–99, 2009.

[229] A. Rathore and R. Gupta. Chitinases from bacteria to human: Properties, applications, and future perspectives. *Enzyme Research*, 2015(8):1 – 8, 2015.

[230] C. Rauskolb, S. Sun, G. Sun, Y. Pan, and K. D. Irvine. Cytoskeletal tension inhibits hippo signaling through an ajuba-warts complex. *Cell*, 158(1):143–156, 2014.

[231] A. Rauwerdink and R. J. Kazlauskas. How the same core catalytic machinery catalyzes 17 different reactions: the serine-histidine-aspartate catalytic triad of α/β-hydrolase fold enzymes. *ACS Catalysis*, 5(10):6153–6176, 2015.

[232] T. Renner and C. D. Specht. Molecular and functional evolution of Class I chitinases for plant carnivory in the caryophyllales. *Molecular Biology and Evolution*, 29(10):2971–2985, 2012.

[233] S. Rentschler, H. Linn, K. Deininger, M. T. Bedford, X. Espanel, and M. Sudol. The ww domain of dystrophin requires ef-hands region to interact with beta-dystroglycan. *Biol Chem*, 380(4):431–42, 1999.

[234] M. W. Ris/or, L. R. Thomsen, K. W. Sanggaard, T. A. Nielsen, I. B. Th/ogersen, M. V. Lukassen, L. Rossen, I. Garcia-Ferrer, T. Guevara, C. Scavenius, E. Meinjohanns, F. X. Gomis-Rüth, and J. J. Enghild. Enzymatic and structural characterization of the major endopeptidase in the Venus flytrap digestion fluid. *The Journal of Biological Chemistry*, 291(5):2271–2287, 2016.

[235] D. R. Roe and r. Cheatham, T. E. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput*, 9(7):3084–95, 2013.

[236] S. Rokudai, Y. Li, Y. Otaka, M. Fujieda, D. M. Owens, A. M. Christiano, M. Nishiyama, and C. Prives. Stxbp4 regulates apc/c-mediated p63 turnover and drives squamous cell carcinogenesis. *Proc Natl Acad Sci U S A*, 115(21):E4806–E4814, 2018.

[237] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins*, 42(1):38–48, 2001.

[238] S. Rottloff, R. Stieber, H. Maischak, F. G. Turini, G. Heubl, and A. Mithöfer. Functional characterization of a class iii acid endochitinase from the traps of the carnivorous pitcher plant genus, *Nepenthes*. *Journal of Experimental Botany*, 62:4639–4647, 2011.

[239] A. J. Rzepiela, L. V. Schäfer, N. Goga, H. J. Risselada, A. H. De Vries, and S. J. Marrink. Reconstruction of atomistic details from coarse-grained structures. *Journal of computational chemistry*, 31(6):1333–1343, 2010.

[240] Z. Salah and R. I. Aqeilan. Ww domain interactions regulate the hippo tumor suppressor pathway. *Cell Death Dis*, 2:e172, 2011.

[241] Z. Salah, S. Cohen, E. Itzhaki, and R. I. Aqeilan. Nedd4 e3 ligase inhibits the activity of the hippo pathway by targeting lats1 for degradation. *Cell Cycle*, 12(24):3817–23, 2013.

[242] Z. Salah, G. Melino, and R. I. Aqeilan. Negative regulation of the hippo pathway by e3 ubiquitin ligase itch is sufficient to promote tumorigenicity. *Cancer Res*, 71(5):2010–20, 2011.

[243] R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand, and R. C. Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *J Chem Theory Comput*, 9(9):3878–88, 2013.

[244] M. J. Scanlon, D. P. Fairlie, D. J. Craik, D. R. Englebretsen, and M. L. West. Nmr solution structure of the rna-binding peptide from human immunodeficiency virus (type 1) rev. *Biochemistry*, 34(26):8242–9, 1995.

[245] K. Schlegelmilch, M. Mohseni, O. Kirak, J. Pruszak, J. R. Rodriguez, D. Zhou, B. T. Kreger, V. Vasioukhin, J. Avruch, T. R. Brummelkamp, and F. D. Camargo. Yap1 acts downstream of alpha-catenin to control epidermal proliferation. *Cell*, 144(5):782–95, 2011.

[246] U. Schutte, S. Bisht, L. C. Heukamp, M. Kebschull, A. Florin, J. Haarmann, P. Hoffmann, G. Bendas, R. Buettner, P. Brossart, and G. Feldmann. Hippo signaling mediates proliferation, invasiveness, and metastatic potential of clear cell renal cell carcinoma. *Transl Oncol*, 7(2):309–21, 2014.

[247] G. Schwarz. Estimating dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[248] W. R. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren. The gromos biomolecular simulation program package. *The Journal of Physical Chemistry A*, 103(19):3596–3607, 1999.

[249] A. Sebe-Pedros, Y. Zheng, I. Ruiz-Trillo, and D. Pan. Premetazoan origin of the hippo signaling pathway. *Cell reports*, 1(1):13–20, 2012.

[250] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.

[251] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, pages 1–5, 2020.

[252] O. Serra, S. Chatterjee, W. Huang, and R. E. Stark. Review: What Nuclear Magnetic Resonance can tell us about protective tissues. *Plant Science*, 195:120–124, 2012.

[253] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten. Dynamical networks in tRNA:protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6620–6625, 2009.

[254] R. Sharma, Y. Chisti, and U. C. Banerjee. Production, purification, characterization, and applications of lipases. *Biotechnology Advances*, 19(8):627–662, 2001.

[255] Y. Shen and A. Bax. Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *J Biomol NMR*, 48(1):13–22, 2010.

[256] A. Shevchenko, M. Wilm, O. Vorm, and M. Mann. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem*, 68(5):850–8, 1996.

[257] J. Si, R. Yan, C. Wang, Z. Zhang, , and X. Su. TIM-Finder: A new method for identifying TIM-barrel proteins. *BMC Structural Biology*, 9(73):doi:10.1186/1472–6807–9–73, 2009.

[258] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7:539–539, 2011.

[259] A. Singh, S. Kumar, M. Sinha, S. Sharma, and T. Singh. Crystal structure of a new form of xylanase-A-amylase inhibitor protein (XAIP-III) at 2.4 åresolution. *To be published*, 2010.

[260] C. A. Smith, V. Calabro, and A. D. Frankel. An rna-binding chameleon. *Mol Cell*, 6(5):1067–76, 2000.

[261] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, and C. M. Dobson. Analysis of main chain torsion angles in proteins: prediction of nmr coupling constants for native and random coil conformations. *J Mol Biol*, 255(3):494–506, 1996.

[262] D. Song, R. Luo, and H. F. Chen. The idp-specific force field ff14idpsff improves the conformer sampling of intrinsically disordered proteins. *J Chem Inf Model*, 57(5):1166–1178, 2017.

[263] D. Song, W. Wang, W. Ye, D. Ji, R. Luo, and H. F. Chen. ff14idps force field improving the conformation sampling of intrinsically disordered proteins. *Chem Biol Drug Des*, 89(1):5–15, 2017.

[264] H. K. Song and S. W. Suh. Refined structure of the chitinase from barley seeds at 2.0 åresolution. *Acta Crystallographica D: Biological Crystallography*, 52(2):289–298, 1996.

[265] S. Spera and A. Bax. Empirical correlation between protein backbone conformation and c. alpha. and c. beta. *Journal of the American Chemical Society*, 113(14):5490–5492, 1991.

[266] R. S. Spolar and J. Record, M. T. Coupling of local folding to site-specific binding of proteins to dna. *Science*, 263(5148):777–84, 1994.

[267] S. Strano, O. Monti, N. Pediconi, A. Baccarini, G. Fontemaggi, E. Lapi, F. Mantovani, A. Damalas, G. Citro, A. Sacchi, G. Del Sal, M. Levrero, and G. Blandino. The transcriptional coactivator yes-associated protein drives p73 gene-target specificity in response to dna damage. *Mol Cell*, 18(4):447–59, 2005.

[268] S. Strano, E. Munarriz, M. Rossi, L. Castagnoli, Y. Shaul, A. Sacchi, M. Oren, M. Sudol, G. Cesareni, and G. Blandino. Physical interaction with yes-associated protein enhances p73 transcriptional activity. *J Biol Chem*, 276(18):15164–73, 2001.

[269] M. Sudol. Newcomers to the ww domain-mediated network of the hippo tumor suppressor pathway. *Genes Cancer*, 1(11):1115–8, 2010.

[270] M. Sudol, P. Bork, A. Einbond, K. Kastury, T. Druck, M. Negrini, K. Huebner, and D. Lehman. Characterization of the mammalian yap (yes-associated protein) gene and its role in defining a novel protein module, the ww domain. *J Biol Chem*, 270(24):14733–41, 1995.

[271] M. Sudol, H. I. Chen, C. Bougeret, A. Einbond, and P. Bork. Characterization of a novel protein-binding module–the ww domain. *FEBS Lett*, 369(1):67–71, 1995.

[272] M. Sudol and K. F. Harvey. Modularity in the hippo signaling pathway. *Trends Biochem Sci*, 35(11):627–33, 2010.

[273] G. J. Székely and M. L. Rizzo. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, 22(2):151–183, 2005.

[274] F. T. Chitinolytic enzymes: catalysis, substrate binding, and their application. *Current Protein and Peptide Science*, 1:105–124, 2000.

[275] K. Takahashi, T. Shimada, M. Kondo, A. Tamai, M. Mori, M. Nishimura, and I. Hara-Nishimura. Ectopic expression of an esterase, which is a candidate for the unidentified plant cutinase, causes cuticular defects in *Arabidopsis thaliana*. *Plant Cell Physiology*, 5:123–131, 2010.

[276] D. Talamantes, N. Biabini, H. Dang, K. Abdoun, and R. Berlemont. Natural diversity of cellulases, xylanases, and chitinases in bacteria. *Biotechnology for Biofuels*, 9(133):DOI: 10.1186/s13068–016–0538–6, 2016.

[277] R. Tan, L. Chen, J. A. Buettner, D. Hudson, and A. D. Frankel. Rna recognition by an isolated alpha helix. *Cell*, 73(5):1031–40, 1993.

[278] R. Tan and A. D. Frankel. Structural variety of arginine-rich rna-binding peptides. *Proc Natl Acad Sci U S A*, 92(12):5282–6, 1995.

[279] T. Tanaka, S. Fujiwara, S. Nishikori, T. Fukui, M. Takagi, and T. Imanaka. A unique chitinase with dual active sites and triple substrate binding sites from the hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. *Applied and Environmental Microbiology*, 65(12):5338–5344, 1999.

[280] C. Tang, M. Chye, S. Ramalingam, S. Ouyang, K. Zhao, W. Ubhayasekera, and S. Mowbray. Functional analyses of the chitin-binding domains and the catalytic domain of brassica juncea chitinase bjchi1. *Plant Molecular Biology*, 56:285–298, 2004.

[281] A. Tanwar, D. Sindhikara, F. Hirata, and R. Anand. Determination of the formylglycinamide ribonucleotide amidotransferase ammonia pathway by combining 3D-RISM theory with experiment. *ACS Chemical Biology*, 10(3):698–704, 2015.

[282] V. E. Tapia, E. Nicolaescu, C. B. McDonald, V. Musi, T. Oka, Y. Inayoshi, A. C. Satteson, V. Mazack, J. Humbert, C. J. Gaffney, M. Beullens, C. E. Schwartz, C. Landgraf, R. Volkmer, A. Pastore, A. Farooq, M. Bollen, and M. Sudol. Y65c missense mutation in the ww domain of the golabi-ito-hall syndrome protein pqbp1 affects its binding activity and deregulates pre-mrna splicing. *J Biol Chem*, 285(25):19391–401, 2010.

[283] A. Tarakhovsky and R. K. Prinjha. Drawing on disorder: How viruses use histone mimicry to their advantage. *Journal of Experimental Medicine*, 215(7):1777–1787, 2018.

[284] O. Tavana, D. Li, C. Dai, G. Lopez, D. Banerjee, N. Kon, C. Chen, A. Califano, D. J. Yamashiro, H. Sun, and W. Gu. Hausp deubiquitinates and stabilizes n-myc in neuroblastoma. *Nature Medicine*, 22(10):1180–1186, 2016.

[285] A. Terwisscha van Scheltinga, K. Kalk, J. Beintema, and D. B.W. Crystal structures of hevamine, a plant defence protein with chitinase and lysozyme activity, and its complex with an inhibitor. *Structure*, 2:1181–1189, 1994.

[286] A. C. Terwisscha van Scheltinga, M. Hennig, and B. W. Dijkstra. The 1.8 åresolution structure of hevamine, a plant chitinase/lysozyme, and analysis of the conserved sequence and structure motifs of glycosyl hydrolase family 18. *Journal of Molecular Biology*, 262(2):243–257, 1996.

[287] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[288] W. Ubhayasekera, R. Rawat, S. W. T. Ho, M. Wiweger, S. Von Arnold, M.-L. Chye, and S. L. Mowbray. The first crystal structures of a family 19 class iv chitinase: the enzyme from norway spruce. *Plant Molecular Biology*, 71(3):277–289, 2009.

[289] W. Ubhayasekera, C. Tang, S. Ho, G. Berglund, T. Bergfors, M.-L. Chye, and S. Mowbray. Crystal structures of a family 19 chitinase from *Brassica juncea* show flexibility of binding cleft loops. *FEBS Journal*, 274:3695–3703, 2007.

[290] A. Ulbricht, F. J. Eppler, V. E. Tapia, P. F. van der Ven, N. Hampe, N. Hersch, P. Vakeel, D. Stadel, A. Haas, P. Saftig, C. Behrends, D. O. Furst, R. Volkmer, B. Hoffmann, W. Kolanus, and J. Hohfeld. Cellular mechanotransduction relies on tension-induced and chaperone-assisted autophagy. *Curr Biol*, 23(5):430–5, 2013.

[291] M. H. Unhelkar, V. T. Duong, K. N. Enendu, J. E. Kelly, S. Tahir, C. T. Butts, and R. W. Martin. Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*. *Biochimica et Biophysica Acta*, 1861:636–643, 2017.

[292] C. Upton and J. Buckley. A new family of lipolytic enzymes? *Trends in Biochemical Science*, 20:178–179, 1995.

[293] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys.*, 37:215–246, 2008.

[294] D. Van Aalten, D. Komander, B. Synstad, S. Gåseidnes, M. Peter, and V. Eijsink. Structural insights into the catalytic mechanism of a family 18 exo-chitinase. *Proceedings of the National Academy of Sciences*, 98(16):8979–8984, 2001.

[295] B. van den Berg. Crystal structure of a full-length autotransporter. *Journal of Molecular Biology*, 396(3):627–633, 2010.

[296] X. Varelas, B. W. Miller, R. Sopko, S. Song, A. Gregorieff, F. A. Fellouse, R. Sakuma, T. Pawson, W. Hunziker, H. McNeill, J. L. Wrana, and L. Attisano. The hippo pathway regulates wnt/beta-catenin signaling. *Dev Cell*, 18(4):579–91, 2010.

[297] T. Vavouri, J. I. Semple, R. Garcia-Verdugo, and B. Lehner. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 138(1):198–208, 2009.

[298] K. Vega and M. Kalkum. Chitin, chitinase responses, and invasive fungal infections. *International Journal of Microbiology*, 2012:Article ID 920459, 2012.

[299] A. Verma, F. Jing-Song, M. L. Finch-Edmondson, A. Velazquez-Campoy, S. Balasegaran, M. Sudol, and J. Sivaraman. Biophysical studies and nmr structure of yap2 ww domain - lats1 ppxy motif complexes reveal the basis of their interaction. *Oncotarget*, 9(8):8068–8080, 2018.

[300] C. Vilela, A. F. Sousa, A. C. Fonseca, A. C. Serra, J. F. Coelho, C. S. R. Freire, and A. J. Silvestre. The quest for sustainable polyesters – insights into the future. *Polymer Chemistry*, 5:3119–3141, 2014.

[301] A. Vite, C. Zhang, R. Yi, S. Emms, and G. L. Radice. alpha-catenin-dependent cytoskeletal tension controls yap activity in the heart. *Development*, 145(5), 2018.

[302] J. A. Vizcaino, R. G. Cote, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger, D. Ovelleiro, Y. Perez-Riverol, F. Reisinger, D. Rios, R. Wang, and H. Hermjakob. The proteomics identifications (pride) database and associated tools: status in 2013. *Nucleic Acids Res*, 41(Database issue):D1063–9, 2013.

[303] B. Vogeli, J. Ying, A. Grishaev, and A. Bax. Limits on variations in protein backbone dynamics from precise measurements of scalar couplings. *J Am Chem Soc*, 129(30):9377–85, 2007.

[304] M. Volokita, T. Rosilio-Brami, N. Rivkin, and M. Zik. Combining comparative sequence and genomic data to ascertain phylogenetic relationships and explore the evolution of the large GDSL-lipase family in land plants. *Molecular Biology and Evolution*, 28(1):551–565, 2011.

[305] I. von Ossowski, S. J., A. Koivula, K. Piens, D. Becker, H. Boer, R. Harle, M. Harris, C. Divne, S. Mahdi, Y. Zhao, D. H., M. Claeyssens, M. Sinnott, and T. Teeri. Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. a comparison with *Phanerochaete chrysosporium* Cel7D. *Journal of Molecular Biology*, 333(4):817–829, 2003.

[306] I. Vujaklija, A. Bielen, T. Paradžik, S. Bidin, P. Goldstein, and D. Vujaklija. An effective approach for annotation of protein families with low sequence similarity and conserved motifs: identifying GDSL hydrolases across the plant kingdom. *BMC Bioinformatics*, 17:91, 2016.

[307] C. Wang, P. H. Nguyen, K. Pham, D. Huynh, T. B. Le, H. Wang, P. Ren, and R. Luo. Calculating protein-ligand binding affinities with mmpbsa: Method and error analysis. *J Comput Chem*, 37(27):2436–46, 2016.

[308] C. Wang, W. Zhang, M. X. Yin, L. Hu, P. Li, J. Xu, H. Huang, S. Wang, Y. Lu, W. Wu, M. S. Ho, L. Li, Y. Zhao, and L. Zhang. Suppressor of deltex mediates pez degradation and modulates drosophila midgut homeostasis. *Nat Commun*, 6:6607, 2015.

[309] J. Wang, Q. Cai, Y. Xiang, and R. Luo. Reducing grid-dependence in finite-difference poisson-boltzmann calculations. *J Chem Theory Comput*, 8(8):2741–2751, 2012.

[310] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS central science*, 5(5):755–767, 2019.

[311] W. Wang, L. Chen, Y. Ding, J. Jin, and K. Liao. Centrosome separation driven by actin-microfilaments during mitosis is mediated by centrosome-associated tyrosine-phosphorylated cortactin. *J Cell Sci*, 121(Pt 8):1334–43, 2008.

[312] W. Wang and R. Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials*, 5(1):1–9, 2019.

[313] W. Wang, J. Huang, and J. Chen. Angiomotin-like proteins associate with and negatively regulate yap1. *J Biol Chem*, 286(6):4364–70, 2011.

[314] W. Wang, J. Huang, X. Wang, J. Yuan, X. Li, L. Feng, J. I. Park, and J. Chen. Ptpn14 is required for the density-dependent control of yap1. *Genes Dev*, 26(17):1959–71, 2012.

[315] W. Wang, X. Li, J. Huang, L. Feng, K. G. Dolinta, and J. Chen. Defining the protein-protein interaction network of the human hippo pathway. *Mol Cell Proteomics*, 13(1):119–31, 2014.

[316] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, 337(3):635–645, 2004.

[317] T. A. Wassenaar, K. Pluhackova, R. A. Böckmann, S. J. Marrink, and D. P. Tieleman. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of chemical theory and computation*, 10(2):676–690, 2014.

[318] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

[319] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994.

[320] M. A. Webb, J.-Y. Delannoy, and J. J. de Pablo. Graph-based approach to systematic molecular coarse-graining. *Journal of chemical theory and computation*, 15(2):1199–1208, 2018.

[321] Y. Wei, J. Schottel, U. Derewenda, L. Swenson, S. Patkar, and Z. Derewenda. A novel variant of the catalytic triad in the Streptomyces scabies esterase. *Nature Structural Biology*, 2:218–223, 1995.

[322] M. A. Weiss, T. Ellenberger, C. R. Wobbe, J. P. Lee, S. C. Harrison, and K. Struhl. Folding transition in the dna-binding domain of gcn4 on specific binding to dna. *Nature*, 347(6293):575–8, 1990.

[323] D. B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ, 1996.

[324] R. M. Williams, Z. Obradovi, V. Mathura, W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown, and A. K. Dunker. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput*, pages 89–100, 2001.

[325] J. R. Williamson. Induced fit in rna-protein recognition. *Nat Struct Biol*, 7(10):834–7, 2000.

[326] K. E. Wilson, Y. W. Li, N. Yang, H. Shen, A. R. Orillion, and J. Zhang. Ptpn14 forms a complex with kibra and lats1 proteins and negatively regulates the yap oncogenic function. *J Biol Chem*, 289(34):23693–700, 2014.

[327] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*, 293(2):321–31, 1999.

[328] B. Xue, A. K. Dunker, and V. N. Uversky. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics*, 30(2):137–149, 2012.

[329] B. Xue, R. W. Williams, C. J. Oldfield, A. K. Dunker, and V. N. Uversky. Archaic chaos: intrinsically disordered proteins in archaea. *BMC Systems Biology*, 4(1):S1, 2010.

[330] S. Yan, W. Wang, J. Marqués, R. Mohan, A. Saleh, W. E. Durrant, J. Song, and X. Dong. Salicylic acid activates DNA damage responses to potentiate plant immunity. *Molecular Cell*, 52(4):602–610, 2013.

[331] W. Ye, D. Ji, W. Wang, R. Luo, and H. F. Chen. Test and evaluation of ff99idps force field for intrinsically disordered proteins. *J Chem Inf Model*, 55(5):1021–9, 2015.

[332] T. H. Yeats, L. B. B. Martin, H. M. Viart, T. Isaacson, Y. He, L. Zhao, A. J. Matas, G. J. Buda, D. S. Domozych, M. H. Clausen, and J. K. C. Rose. The identification of cutin synthase: formation of the plant polyester cutin. *Nature Chemical Biology*, 8(7):609–611, 2012.

[333] B. Yeung, K. C. Ho, and X. Yang. Wwp1 e3 ligase targets lats1 for ubiquitin-mediated degradation in breast cancer cells. *PLoS One*, 8(4):e61027, 2013.

[334] S. Yonemura, Y. Wada, T. Watanabe, A. Nagafuchi, and M. Shibata. alpha-catenin as a tension transducer that induces adherens junction development. *Nat Cell Biol*, 12(6):533–42, 2010.

[335] J. Yu, Y. Zheng, J. Dong, S. Klusza, W. M. Deng, and D. Pan. Kibra functions as a tumor suppressor protein that regulates hippo signaling in conjunction with merlin and expanded. *Dev Cell*, 18(2):288–99, 2010.

[336] A. Zemla. Lga: a method for finding 3d similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374, 2003.

[337] B. Zhang, L. Zhang, F. Li, D. Zhang, X. Liu, H. Wang, Z. Xu, C. Chu, and Y. Zhou. Control of secondary cell wall patterning involves xylan deacetylation by a GDSL esterase. *Nature Plants*, 3:17017, 2017.

[338] L. Zhang, J. Han, H. Wang, R. Car, and W. E. Deepcg: Constructing coarse-grained models via deep neural networks. *The Journal of chemical physics*, 149(3):034101, 2018.

[339] Y. Zhang. I-tasser server for protein 3d structure prediction. *BMC Bioinformatics*, 9:40, 2008.

[340] Y. Zhang, B. Bai, M. Lee, Y. Alfiko, A. Suwanto, and G. H. Yue. Cloning and characterization of *Eg*GDSL, a gene associated with oil content in oil palm. *Scientific Reports*, 8:11406, 2018.

[341] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Appendix A

Supplement: Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*.

Sequence Alignments

The catalytic action of family 18 chitinases, which retains the β-anomeric carbon stereochemistry from the substrate to the product, is based on substrate-assisted hydrolysis of the glycosidic bond [294, 210, 173]. Catalysis is initiated by distorting the -1 sugar ring subsite adjacent to the glycosidic bond. Next, Asp 123 rotates to form hydrogen bonds with both Glu 127 and the N-acetyl group of the +1 sugar. This step protonates Glu 127. Then, the anomeric carbon is subjected to a nucleophilic attack by the oxygen from the N-acetyl group, forming an oxazolinium ion as an intermediate, followed by cleavage of the glycosidic bond by hydrolysis to generate smaller fragments. The DXDXE motif is essential for activity, hence fragments that were lacking this sequence due to truncation were excluded from our

protein set.

A sequence alignment for Family 18 chitinases from Caryophylalles carnivorous plants is shown in Supplementary Figure A.1. The figure is annotated to highlight specific amino acid properties and important sequence features. The chemical properties of amino acids are color-coded as follows: cysteines are yellow, positively charged residues are blue, negatively charged residues are red, hydrophobic residues are green, and all others are black. Highly conserved residues are indicated with a dot above the sequence position. Cysteine residues involved in structure-stabilizing disulfide bonds are indicated with yellow asterisks, while the active amino acid residues are marked with colored arrows. SignalP 4.1 is used to predict the signal peptide cleavage site, which is specified by underlining the residues on either of the cleavage point. The signal peptide itself is highlighted in light orange. Strikethrough text indicates sequence regions that are absent in the active enzyme, in this case the N-terminal signal peptide that is expressed but removed during maturation. Annotations were performed by homology to a well-characterized acidic endochitinase from *Vitis vinifera* (CHIT3_VITVI, Uniprot ID-P51614).

Family 19 contains Class I, II, and IV chitinases, all of which are characterized by an anomeric inverting mechanism [274, 125]. The N-terminal chitin-binding domain is present in Class I and absent in Class II, which are otherwise similar in sequence. Family 19 chitinases from plants have in common a catalytic domain with an active glutamic acid residue. The active site motif surrounding the active E is either HETT (type I and II) or HETG (type IV) [232], both of which are observed in this set of proteins. Annotations for the Family 19 chitinases are shown in Supplementary Figure A.2. Amino acid and sequence features are indicated as in Supplementary Figure A.1, with the following additions, when present: the C-rich domain is highlighted in light green, the P-rich hinge in light blue, and the C-terminal extension (CTE) in light gray. Both the C-rich domain and the P-rich hinge are highly variable in length and are absent in some sequences. Only three chitinases in this

set contain the CTE, which targets those sequences to the vacuole. The reference sequences for this cluster are CHI3_CASSA (*Castanea sativa*), CHI2_BRANA (*Brassica napus*), and HORV2 (*Hordeum vulgare*).

Figure A.1: Sequence alignment for Family 18 chitinases, annotated by homology to the reference sequence CHIT3_VITVI. The "DXDXE" motif, in which the acidic residues are marked with red arrows, is imperative for the enzyme activity. Orange arrows indicate residues implicated in substrate binding.

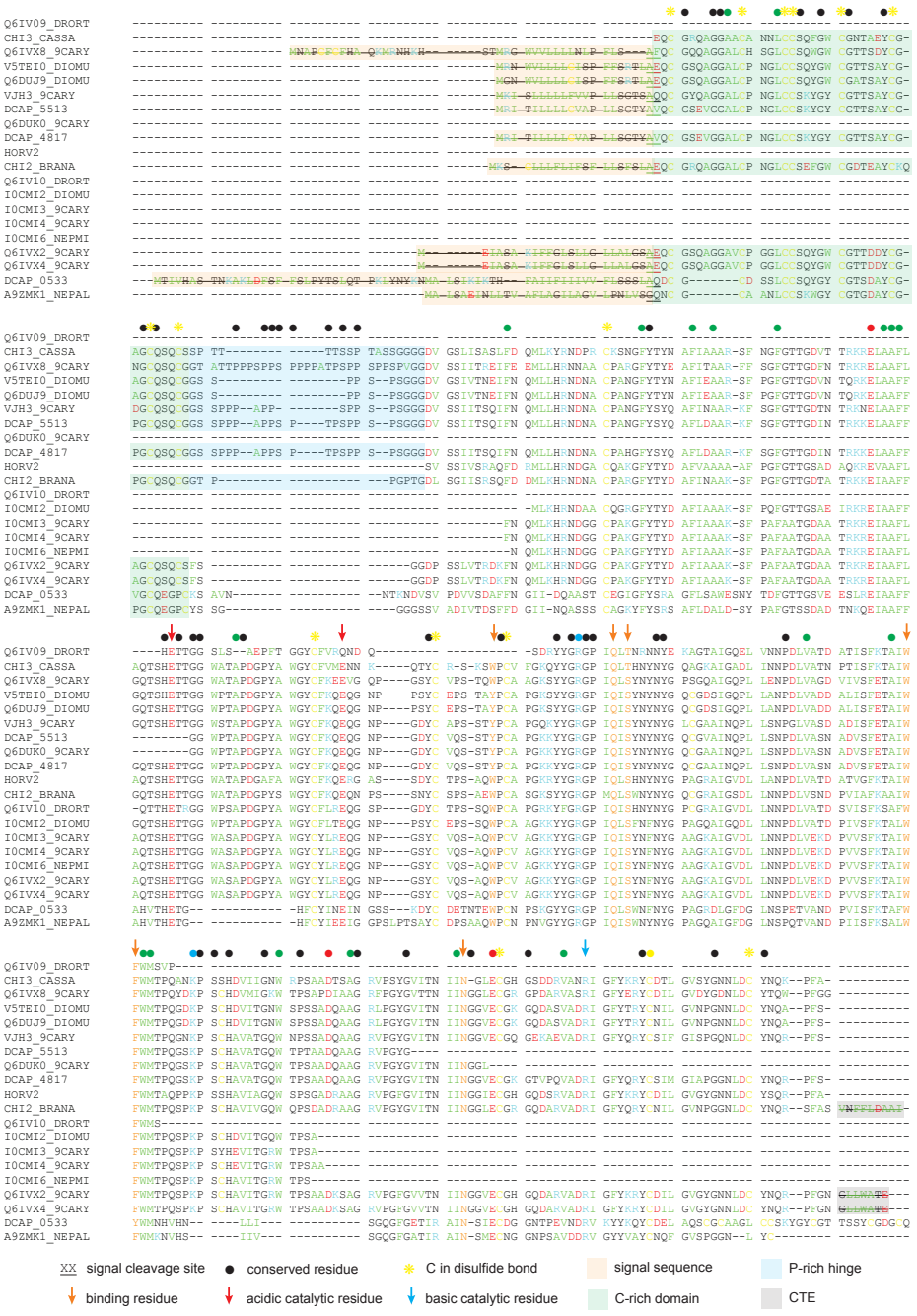


Figure A.2: Sequence alignment and annotation for Family 19 chitinases. Many sequences in this cluster contain a chitin-binding C-rich domain (light green) that is connected to the active region by a P-rich hinge (light blue). Three sequences in this cluster contain a C-terminal extension (CTE) that causes the proteins to be targeted to the vacuole.

Four Family 19 chitinase fragments were identified from the *D. capensis* genome by performing a BLAST search for DcChit_1, a chitinase fragment previously identified from genomic DNA of the same organism [232]. Their sequences range from 41%-100% identity to DcChit1_1. These fragments contain part of the N-terminal region, including the C-rich domain and the P-rich hinge, neither of which was observed in the original fragment, along with part of the catalytic domain (Supplementary Figure A.3). However, these sequences are all truncated before the catalytic residues. Sequencing of the *D. capensis* transcriptome will clarify whether these are fragments of active genes containing one or more introns, or inactive pseudogenes, which are relatively common in gene families undergoing rapid evolution [116] (as is the case for many proteins associated with pathogen defense [133]).

```
DcChitI_2       ---------- ---------- ---------- -MRITILLLL CVAPLLSGTY AVQCGSEVGG ALCPNGLCCS KYGYCGTTSA YCGPGCQSQC GGSSPPPAPP
I0CMI1_DROCA    ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
DcChitI_1       ---------- ---------- ---------- -MRITILLLL CVAPLLSGTY AVQCGSEVGG ALCPNGLCCS KYGYCGTTSA YCGPGCQSQC GGSSPPPAPP
DcChitI_3       MKTRSIPEIS STAPIISFTL DHTIQTRKIM SPPMKSIHMI CLVAAVIIFL TMPRHLAAQS CGCAAGLCCS KYGYCGTTSD YCGDGCQAGP CSSTPA----
DcChitI_4       ---------- ---------- ---------M SPPMRNYHMT CLVTAVIIFF TMPGHLAAQS CGCAAGLCCS KYGYCGTTSS YCGDGCQAGP CSSTPT----

DcChitI_2       SPTPSPPSPS GGGDVSSIIT SQIFNQMLLH RNDNACPANG FYSYQAFLDA ARKFSGFGTT GDINTRKKEL AAFFGQTSHE TTG------- ----------
I0CMI1_DROCA    ------PSPS GGGDVSSIIT SQIFNQMLLH RNDNACPAHG FYSYQAFLDA ARKFSGFGTT GDINTRKKEL AAFFGQTSHE TT-------- ----------
DcChitI_1       SPTPSPPSPS GGGDVSSIIT SQIFNQMLLH RNDNACPAHG FYSYQAFLDA ARKFSGFGTT GDINTRKKEL AAFFGQTSHE TT-------- ----------
DcChitI_3       ---------G SGVSVPAVVT VAFF-NGIIN KAGSGCPGTG FYSRSAFLSA IGSYPSFGTT GTSDAAKREI AAFFAHVTHE TGCKHIHFFL SKFYAVLYRV
DcChitI_4       ---------S SGVSVPAVVT DAFF-NGIIN QAGSGCPGKG FYSRSAFLSA IGSYPSFGTT GTTDASKQEI AAFFAHVTHE T--------- ----------

DcChitI_2       ---------- ---
I0CMI1_DROCA    ---------- ---
DcChitI_1       ---------- ---
DcChitI_3       IILYAWIKDE AID
DcChitI_4       ---------- ---
```

Figure A.3: Chitinase 1 fragments discovered using a BLAST search of the *D. capensis* genome against the DcChitI_1 fragment previously identified by Renner and Specht from *D. capensis* genomic DNA.

Preliminary Structural Models and *In silico* Maturation

Preliminary models for both Family 18 and Family 19 chitinases were produced using Rosetta [139], implemented in the online Robetta server [228]. The Rosetta structures contain the full sequences, including the N-terminal signal peptides, and in some cases, C-terminal targeting peptides that are also cleaved during maturation. The *in silico* maturation process, which we have previously described for cysteine proteases [38], is illustrated in Supplementary Fig-

ure A.4 for a representative family 18 chitinase, DCAP_2209. The initial Rosetta sequence, including the signal peptide and lacking post-translational modifications, is shown in Supplementary Figure A.4. In order to generate the equilibrated structure Supplementary Figure A.4b, which more closely approximates the active form of the enzyme in solution, the signal sequence is removed, disulfide bonds are added using homology to a reference sequence (in this case CHIT3_VITVI), and the structure is equilibrated in explicit solvent. Many Family 18 chitinases from plants contain three disulfide bonds [19, 134], although examples without any disulfide bonds also exist [140]. Three are found in all the Family 18 chitinases in this set, as in CHIT3_VITVI [33], and hevamine from *Hevea brasiliensis* (PDB ID: 2HVM) [285]. The functionally important cis peptide bonds are captured by the molecular models for all the Family 18 chitinases examined here except for DCAP_7323, which unlikely to be active in any case because it is truncated at the N-terminal end.

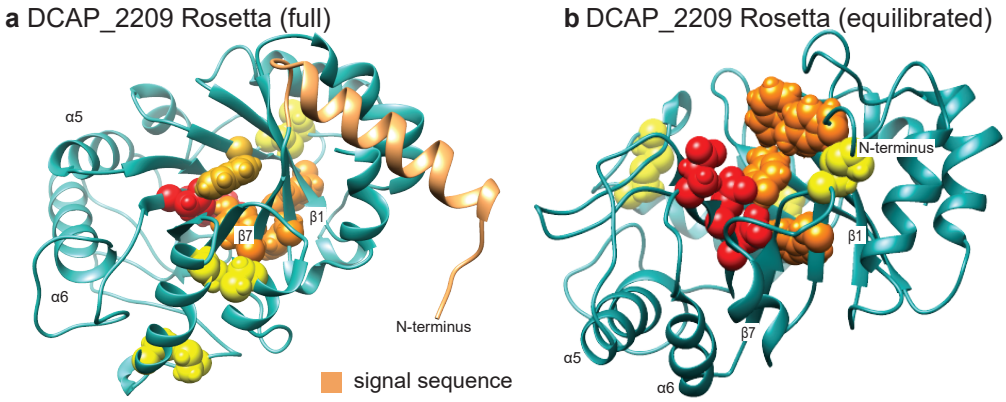


Figure A.4: DCAP_2209 (a) before and (b) after *in silico* maturation. The light orange helix in part a is the N-terminal signal sequence. Important residues are color-coded as follows: Red: catalytically active residues of the "DXDXE" motif. Orange: aromatic substrate-binding residues. Yellow: Cysteines in disulfide bonds.

Supplementary Figure A.5 shows full-length structures for Q6IVX8_9CARY and Q6IVX2_9CARY from *Drosera spatulata*. The N-terminal and C-terminal targeting sequences are exposed on the surface of the protein, as expected. The P-rich hinge in these proteins is variable in length, and highly flexible, as illustrated by the different relative conformations of of the

Supplementary Table A.1: Structures used in DCAP_0106 structure prediction (Ginzu).

PDBID	protein name	organism	% identity to target	citation
4TX6 (B)	AfChiA1	*Aspergillus fumigatus*	26.13	[161]
3MU7 (A)	XAIP-II	*Scadoxus multiflorus*	41.64	[143]
3D5H (A)	haementhin	*Haemanthus multiflorus*	42.14	[144]
2GSJ (A)	PPL2	*Parkia platycephala*	55.64	[49]
4TOQ (A)	Class II chitinase	*Punica granatum*	51.26	[183]
1HVQ (A)	hevamine	*Hevea brasiliensis*	55.27	[285]
2HVM (A)	hevamine	*Hevea brasiliensis*	55.64	[286]
1KR0 (A)	hevamine variant D125A/Y183F	*Hevea brasiliensis*	54.18	[26]
1KR1 (A)	hevamine variant D125A/E127A	*Hevea brasiliensis*	54.18	[26]
1KQY (A)	hevamine variant D125A/E127A/Y183F	*Hevea brasiliensis*	53.82	[26]
3O9N (A)	XAIP-III	*Scadoxus multiflorus*	42.86	[259]

catalytic and C-rich chitin binding domains observed here.

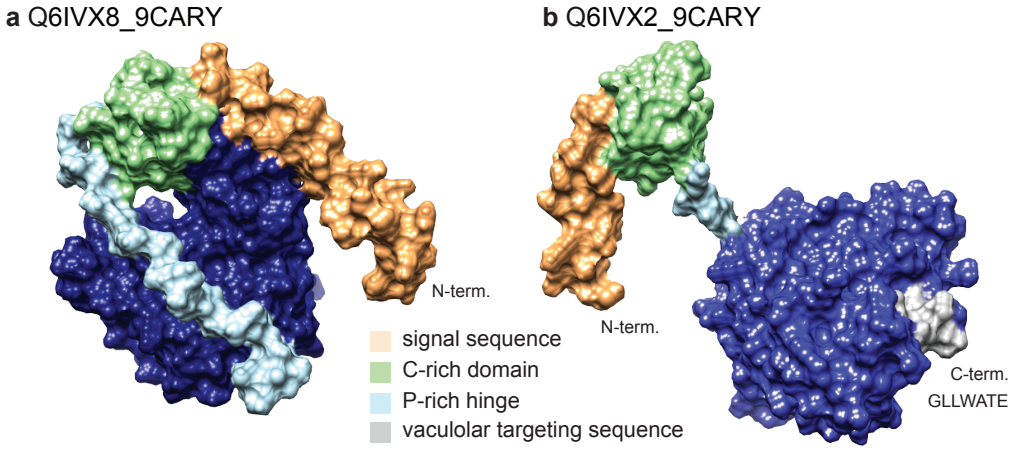


Figure A.5: Initial Rosetta structures for two class I chitinases from *Drosera spatulata*, Q6IVX8_9CARY and Q6IVX2_9CARY, illustrating positioning of the N-terminal and C-terminal targeting sequences and the variability in length and conformation for the P-rich hinge.

Description of a Novel Two-Domain Class IV Chitinase

Class IV chitinases exhibit an amino acid substitution in the first active site region relative to Class I chitinases, resulting in a HETG/I motif instead of the HETT motif [232]. A deletion of four amino acids in the Cys-rich binding domain is also observed in class IV

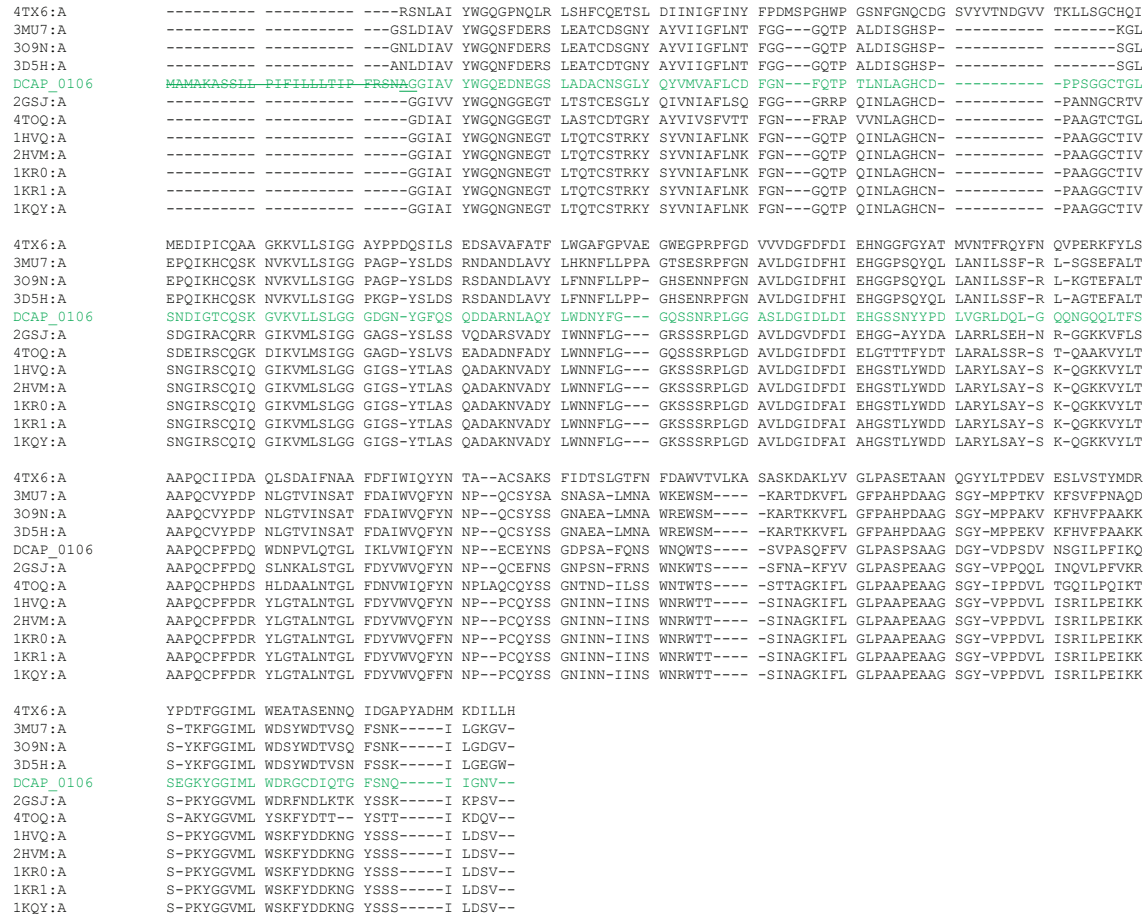


Figure A.6: Sequences used for domain prediction of DCAP_0106, designated by PDBID. The target sequence is colored green. Strikethrough text indicates the N-terminal signal sequence, which is removed during maturation.

Supplementary Table A.2: Structures used in DCAP_5513 structure prediction (Ginzu).

PDBID	protein name	organism	% identity to target	citation
2Z39 (A)	Bjchi3-E234A	*Brassica juncea*	40.65	[289]
2Z38 (A)	Bjchi3	*Brassica juncea*	40.89	[289]
1DXJ (A)	jack bean chitinase	*Canavalia ensiformis*	45.08	[100]
2DKV (A)	OsChia1b	*Oryza sativa* L. japonica	47.21	[135]
4DWX (A)	GH-19 chitinase	*Secale cereale*	45.87	[199]
4J0L (A)	GH-19 chitinase W72A/E67Q	*Secale cereale*	45.68	[199]
1CNS (A)	GH-19 chitinase	*Hordeum vulgare*	46.91	[264]
2BAA (A)	GH-19 chitinase	*Hordeum vulgare*	46.50	[105]
4TX7 (A)	GH-19 chitinase	*Vigna unguiculata*	47.15	[214]
3CQL (A)	GH-19 chitinase	*Carica papaya*	45.68	[118]
4MST (A)	HbCLP1	*Hevea brasiliensis*	48.13	[182]

```
DCAP_5513       MRITILLLLG VAPLLSGTVA VQCGSEVGGA LCPNGLCCSK YGYCGTTSAY CGPGCQSQCG GSSPPPAPPS PTPSPPSPSG GGDVSSIITS QIFNQMLLHR
2Z39:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- FGDLSGIISR DQFYKMLKHM
2Z38:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------E FGDLSGIISR DQFYKMLKHM
1DXJ:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- --DVGSVIDA SLFDQLLKHR
2DKV:A          ---------- ---------M EQCGAQAGGA RCPNCLCCSR WGWCGTTSDF CGDGCQSQCS GCGPTPTPT- ------PPSP SDGVGSIVPR DLFERLLLHR
4DWX:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -MSVSSIISH AQFDRMLLHR
4J0L:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -MSVSSIISH AQFDRMLLHR
1CNS:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- --SVSSIVSR AQFDRMLLHR
2BAA:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- --SVSSIVSR AQFDRMLLHR
4TX7:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------A PSDLSALIPR ATFDQMLKHR
3CQL:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- --GIEKIISR SMFDQMLKHR
4MST:A          ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -----SIISR STFEEMLKHR

DCAP_5513       NDNACPANGF YSYQAFLDAA RKFSGFGTTG DINTRKKELA AFF------- -GGWPTAPDG PYAWGYCFKQ EQGNPGDYCV Q-SSTYPCAP GKKYYGRGPI
2Z39:A          NDNDCHAVGF FTYDAFITAA KSFPSFGNTG DLAMRKKEIA AFFGQTSHET TGGWSGAPDG ANTWGYCYKE AIDKSDPHCD SNNLEWPCAP GKFYYGRGPM
2Z38:A          NDNDCHAVGF FTYDAFITAA KSFPSFGNTG DLAMRKKEIA AFFGQTSHET TGGWSGAPDG ANTWGYCYKE EIDKSDPHCD SNNLEWPCAP GKFYYGRGPM
1DXJ:A          NDPACEGKGF YSYNAFVTAA RSFGGFGTTG DTNTRKREVA AFLAQTSHET TGGAAGSPDG PYAWGYCFVT ERDKSNKYCD P-G--TPCPA GKSYYGRGPI
2DKV:A          NDGACPARGF YTYEAFLAAA AAFPAFGGTG NTETRKREVA AFLGQTSHET TGGWPTAPDG PFSWGYCFKQ EQNPPSDYCQ P-SPEWPCAP GRKYYGRGPI
4DWX:A          NDGACQAKGF YTYDAFVAAA NAFPGFGATG STDARKRDVA AFLAQTSHET TGGWATAPDG AFAWGYCFKQ ERGAAADYCT P-SAQWPCAP GKRYYGRGPI
4J0L:A          NDGACQAKGF YTYDAFVAAA NAFPGFGATG STDARKRDVA AFLAQTSHQT TGGAATAPDG AFAWGYCFKQ ERGAAADYCT P-SAQWPCAP GKRYYGRGPI
1CNS:A          NDGACQAKGF YTYDAFVAAA AAFSGFGTTG SADVQKREVA AFLAQTSHET TGGWATAPDG AFAWGYCFKQ ERGASSDYCT P-SAQWPCAP GKRYYGRGPI
2BAA:A          NDGACQAKGF YTYDAFVAAA AAFPGFGTTG SADAQKREVA AFLAQTSHET TGGWATAPDG AFAWGYCFKQ ERGASSDYCT P-SAQWPCAP GKRYYGRGPI
4TX7:A          NDGACPARGF YTYDAFIAAA RAFPSFGNTG DTATRKREIA AFLGQTSHET TGGWPSAPDG PYAWGYCFVR EQNP-SAYCS P-TPQFPCAS GQQYYGRGPI
3CQL:A          NNPACPAKGF YTYDAFIAAA KSFPSFGTTG STDVRKREIA AFLGQTSHET TGGWPSAPDG PYAWGYCFLK ERNPSSNYCA P-SPRYPCAP GKSYYGRGPI
4MST:A          NDAACPAKGF YTYDAFISAA KAFPAFGTTG DVDTCKREIA AFFGQTSHAT TGGWPTAPDG PYAWGYCYKE ELNQASSYCS P-SPAYPCAP GKKYYGRGPI

DCAP_5513       QISYNYNYGQ CGVAINQPLL SNPDLVASNA DVSFETAIWF WMTPQGSKPS CHAVATGQWT PTAADQAAGR VPGYGVITNI INGGVECGKG TVPQVADRIG
2Z39:A          MLSWNYNYGP CGRDLGLELL KNPDVASSDP VIAFKTAIWF WMTPQAPKPS CHDVITDQWE PSAADISAGR LPGYGVITNI INGGLECAGR DVAKVQDRIS
2Z38:A          MLSWNYNYGP CGRDLGLELL KNPDVASSDP VIAFKTAIWF WMTPQAPKPS CHDVITDQWE PSAADISAGR LPGYGVITNI INGGLECAGR DVAKVQDRIS
1DXJ:A          QLTHNYNYAQ AGRALGVDLI NNPDLVARDA VISFKTAIWF WMTPQGNKPS CHDVITNRWT PSAADVAANR TPGFGVITNI INGGIECGRG PSPASGDRIG
2DKV:A          QLSFNFNYGP AGRAIGVDLL SNPDLVATDA TVSFKTALWF WMTPQGNKPS SHDVITGRWA PSPADAAAGR APGYGVITNI VNGGLECGHG PDDRVANRIG
4DWX:A          QLSHNYNYGP AGRAIGVDLL RNPDLVATDP TVSFKTALWF WMTAQAPKPS SHAVITGKWS PSGADRAAGR APGFGVITNI INGGLECGHG QDSRVADRIG
4J0L:A          QLSHNYNYGP AGRAIGVDLL RNPDLVATDP TVSFKTALWF WMTAQAPKPS SHAVITGKWS PSGADRAAGR APGFGVITNI INGGLECGHG QDSRVADRIG
1CNS:A          QLSHNYNYGP AGRAIGVDLL ANPDLVATDA TVSFKTAMWF WMTAQPPKPS SHAVIVGQWS PSGADRAAGR VPGFGVITNI INGGIECGHG QDSRVADRIG
2BAA:A          QLSHNYNYGP AGRAIGVDLL ANPDLVATDA TVGFKTAIWF WMTAQPPKPS SHAVIAGQWS PSGADRAAGR VPGFGVITNI INGGIECGHG QDSRVADRIG
4TX7:A          QISWNYNYGQ CGNAIGVDLI NNPDLVATDP VVSFKSAIWF WMTPQSPKPS SHDVITSQWT PSAADVAAGK LPGYGTVTNI INGGLECGRG QDSRVEDRIG
3CQL:A          QLSWNYNYGP CGEALRVNLL GNPDLVATDR VISFKTALWF WMTPQAPKPS CHDVITGRWQ PSAADTAAGR LPGYGVITNI INGGLECGKG PNPQVADRIG
4MST:A          QLSWNYNYGQ CGQALGLDLL NNPDLVATDR VISFKAAIWF WMTPQFPKPS CHDVITGQWS PTGHDISAGR APGYGVITNI INGGLECGRG WDARVEDRIG

DCAP_5513       FYQRYCSIMG ISPGGNLDCY NQRPFS---- ------
2Z39:A          FYTRYCGMFG VDPGSNIDCD NQRPFN---- ------
2Z38:A          FYTRYCGMFG VDPGSNIDCD NQRPFN---- ------
1DXJ:A          FYKRYCDVLH LSYGPNLNCR DQRPFGG--- ------
2DKV:A          FYQRYCGAFG IGTGGNLDCY NQRPFNSGSS VGLAEQ
4DWX:A          FYKRYCDILG VGYGDNLDCY NQRPFA---- ------
4J0L:A          FYKRYCDILG VGYGDNLDCY NQRPFA---- ------
1CNS:A          FYKRYCDILG VGYGNNLDCY SQRPFA---- ------
2BAA:A          FYKRYCDILG VGYGNNLDCY SQRPFA---- ------
4TX7:A          FFKQYCDLFG VGYGNNLDCY SQAPFG---- ------
3CQL:A          FFRRYCGILG VGTGNNLDCY NQRPFG---- ------
4MST:A          FYKRYCDMFA VGYGSNLDCY NQTPFGLG-- ------
```

Figure A.7: Sequences used for domain prediction of DCAP_5513, designated by PDBID. The target sequence is colored green. Strikethrough text indicates the N-terminal signal sequence, which is removed during maturation.

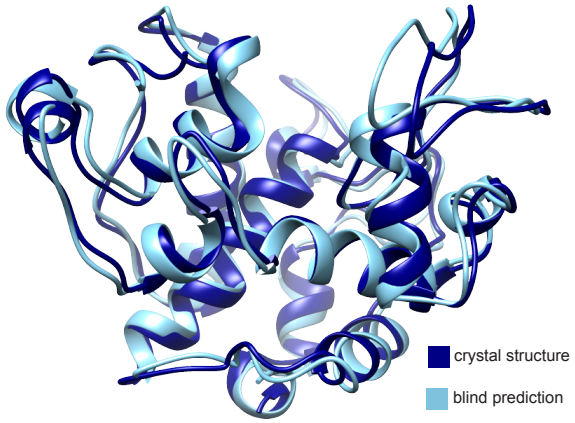


Figure A.8: Blind prediction of the HORV2 structure (PDBID:1CNS chain A, light blue) compared with the experimentally determined crystal structure (dark blue).

chitinases, as shown for a class IV chitinase from *Nepenthes alata* (A9ZMK1_NEPAL) [126] and DCAP_0533 in A.2. Supplementary Figure A.9 shows a sequence alignment of the N- and C-terminal domains of the Class IV chitinase DCAP_0533 with single domain class IV chitinases from *Picea abies* (Q6WSR8_PICAB), *Zea mays* (CHIA_MAIZE), and *Sorghum bicolor* (C5YBE7_SORBI). The two domains of DCAP_0533 were aligned with the most closely related annotated class IV chitinases, those from *Picea abies* (EC: 3.2.1.14, Uniprot: Q6WSR8_PICAB), *Zea mays* (EC: 3.2.1.14, Uniprot: CHIA_MAIZE), and *Sorghum bicolor* (Uniprot: C5YBE7_SORBI) [288, 232, 56] (Supplementary Figure A.10).

Structurally, each domain consists of two lobes with eight helices each, separated by a large active site cleft (Supplementary Figure A.10(a). In Supplementary Figure A.10(b), the two domains of this protein are shown overlaid with the crystal structures of class IV chitinases from *Zea mays* (PDBID: 4MCK, 60% identity with the NTD) and *Picea abies* (PDBID: 3HBE, 64% identity with the CTD). The NTD Supplementary Figure A.10(c) has an N-terminal signal peptide, a conserved C-rich binding domain, and a catalytic domain that appears to be functional. In its homolog CHIA_MAIZE, Chaudet et. al. characterized four

catalytic residues (E62, E71, E165, and R171) [56], all of which have counterparts in the NTD of DCAP_0533 (E173, E182, E278, R290) (Supplementary Figures A.9, A.10. Previous modeling studies of well-characterized class I chitinases from barley, mustard, and chestnut seed homologs (barley: E67, mustard: E212, chestnut: E124) suggest the necessity of E62 in CHIA_MAIZE and E173 in the NTD of DCAP_0533 as a proton donor [9, 92, 280]. Overall, mutagenesis studies highlight the significance of E62 as an essential residue of the catalytic triad (E62, E165, R171 in CHIA_MAIZE) which we use to infer an equivalent catalytic triad in the NTD of DCAP_0533 (E173, E278, and R290). It has also been hypothesized that purpose of the triad is to alter the surrounding environment to induce activation of the glutamic acid in the HETG/I (class IV) or HETT (class I/II) motif by changing its pKa [280].

Linked to the NTD by a cysteine and glycine-rich linker sequence, the CTD of DCAP_0533 (Supplementary Figure A.10(d) potentially houses a second catalytic domain or binding domain whose closest structural homolog is Q6WSR8_PICAB from Norway spruce (*Picea abies*) (Supplementary Figure A.9). Binding site residues and cysteines involved in disulfide bond formation are conserved in both chitinases. Comparing this sequence with the catalytic triad of Q6WSR8_PICAB (E113, R230, E218), we observe a potentially equivalent triad in the CTD (E407, E507, R519) (Supplementary Figure A.10). Ubhayasekera et. al. describe the flexibility of E113 and demonstrate two conformations that it can adopt during catalysis [288]. Although E407 is not located in the equivalent sequence position to E113, the flexibility of this residue in Q6WSR8_PICAB suggests that Glu407 may be at an appropriate distance to function as part of the CTD triad. Alternatively, the CTD may lack catalytic activity and act as a binding domain as in multidomain chitinases from archaea and bacteria.

All initial and equilibrated structures are available for download as PDB files. The available structures for Families 18 and 19 are tabulated in Supplementary Tables 1 and 2, respectively.

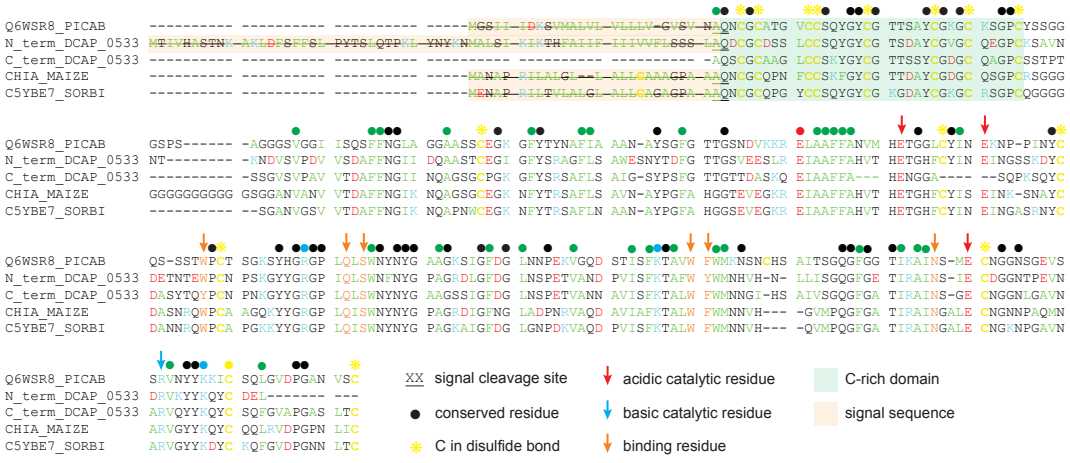


Figure A.9: Sequence alignment and annotation of Q6WSR8_PICAB, CHIA_MAIZE, and the N-terminal domain (NTD) and C-terminal domain (CTD) of DCAP_0533. For the purpose of comparison, the sequence is manually separated above. We observe high sequence conservation regarding: the signal cleavage site, C-rich domain length and location, cysteines composing disulfide bonds, other binding site residues surrounding the main binding site residues (orange arrows), and catalytic residues except Glu407 of the CTD which is unaligned with Glu113 of Q6WSR8_PICAB

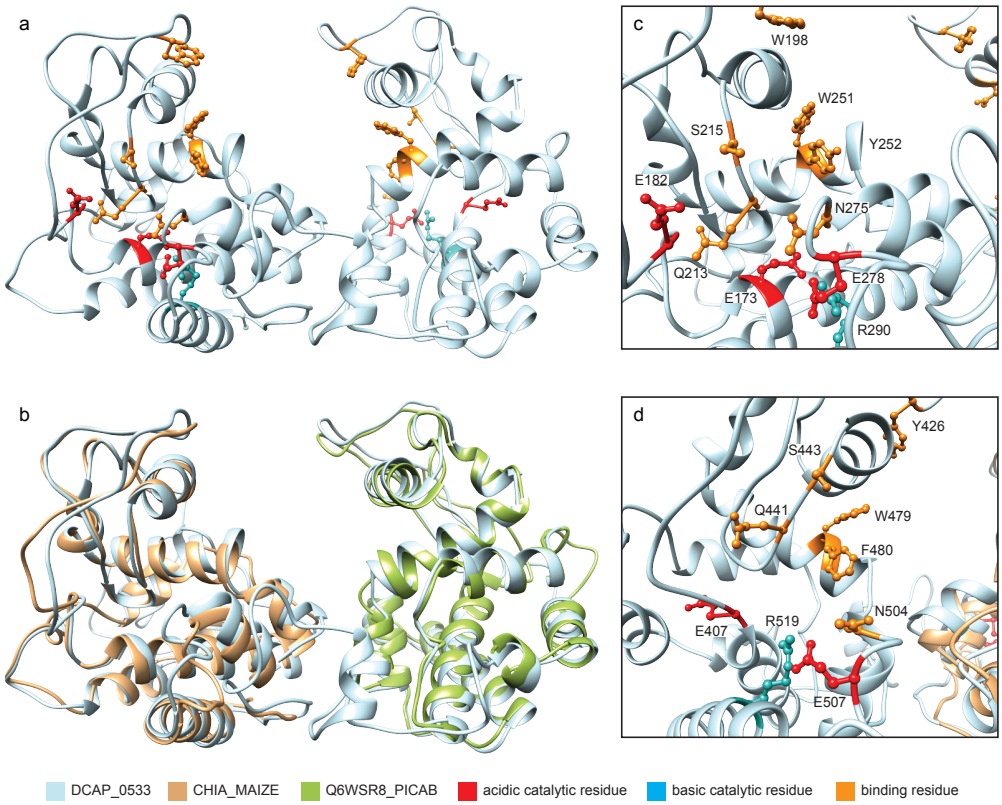


Figure A.10: DCAP_0533 comparison with CHIA_MAIZE (4MCK) and Q6WSR8_PICAB (3HBE) and close up of catalytic residues and binding residues: (a) Robetta generated predicted structure with highlighted catalytic residues and binding residues. (b) Superimposition of CHIA_MAIZE and Q6WSR8_PICAB against DCAP_0533. (c) Catalytic site of NTD with 1-letter residue code and specifier. Catalytic triad consists of E173, E278, R290. (d) Catalytic site of CTD with 1-letter residue code and specifier. Catalytic triad consists of E407, E507, R519.

Supplementary Table A.3: Rosetta and equilibrated structures for Family 18 Chitinases PDB files available for download

Protein	Organism	Sequence Elements included	File Name
CHIT3_VITVI	*Vitis vinifera*	signal, active region	CHIT3_VITVI_m1.pdb
CHIT3_VITVI	*Vitis vinifera*	active region	CHIT3_VITVI_mature_m1.pdb
DCAP_7323	*D. capensis*	active region	DCAP_7323_m1.pdb
DCAP_7323	*D. capensis*	active region	DCAP_7323_mature_m1.pdb
DCAP_0106	*D. capensis*	signal, active region	DCAP_0106_m1.pdb
DCAP_0106	*D. capensis*	active region	DCAP_0106_mature_m1.pdb
DCAP_7544	*D. capensis*	signal, active region	DCAP_7544_m1.pdb
DCAP_7544	*D. capensis*	active region	DCAP_7544_mature_m1.pdb
DCAP_2209	*D. capensis*	signal, active region	DCAP_2209_m1.pdb
DCAP_2209	*D. capensis*	active region	DCAP_2209_mature_m1.pdb
C7F821_NEPMI	*N. mirabilis*	signal, active region	C7F821_NEPMI_m1.pdb
C7F821_NEPMI	*N. mirabilis*	active region	C7F821_NEPMI_mature_m1.pdb
C7F817_9CARY	*D. spatulata*	signal, active region	C7F817_9CARY_m1.pdb
C7F817_9CARY	*D. spatulata*	active region	C7F817_9CARY_mature_m1.pdb
I7HCY7_NEPAL	*N. alata*	signal, active region	I7HCY7_NEPAL_m1.pdb
I7HCY7_NEPAL	*N. alata*	active region	I7HCY7_NEPAL_mature_m1.pdb
C7F818_9CARY	*D. spatulata*	signal, active region	C7F818_9CARY_m1.pdb
C7F818_9CARY	*D. spatulata*	active region	C7F818_9CARY_mature_m1.pdb
Q06SN0_9CARY	*D. spatulata*	signal, active region	Q06SN0_9CARY_m1.pdb
Q06SN0_9CARY	*D. spatulata*	active region	Q06SN0_9CARY_mature_m1.pdb
C7F824_9CARY	*D. spatulata*	signal, active region	C7F824_9CARY_m1.pdb
C7F824_9CARY	*D. spatulata*	active region	C7F824_9CARY_mature_m1.pdb
C7F822_9CARY	*D. spatulata*	signal, active region	C7F822_9CARY_m1.pdb
C7F822_9CARY	*D. spatulata*	active region	C7F822_9CARY_mature_m1.pdb
C7F819_9CARY	*D. spatulata*	signal, active region	C7F819_9CARY_m1.pdb
C7F819_9CARY	*D. spatulata*	active region	C7F819_9CARY_mature_m1.pdb
C7F823_NEPGR	*N. gracilis*	signal, active region	C7F823_NEPGR_m1.pdb
C7F823_NEPGR	*N. gracilis*	active region	C7F823_NEPGR_mature_m1.pdb
DCAP_5455	*D. capensis*	signal, active region	DCAP_5455_m1.pdb
DCAP_5455	*D. capensis*	active region	DCAP_5455_mature_m1.pdb
DCAP_2879	*D. capensis*	signal, active region	DCAP_2879_m1.pdb
DCAP_2879	*D. capensis*	active region	DCAP_2879_mature_m1.pdb
DCAP_4799	*D. capensis*	signal, active region	DCAP_4799_m1.pdb
DCAP_4799	*D. capensis*	active region	DCAP_4799_mature_m1.pdb
DCAP_2737	*D. capensis*	signal, active region	DCAP_2737_m1.pdb
DCAP_2737	*D. capensis*	active region	DCAP_2737_mature_m1.pdb

Supplementary Table A.4: Rosetta and equilibrated structures for Family 19 Chitinases PDB files available for download

Protein	Organism	Sequence Elements included	File Name
HORV2	*H. vulgare*	active region	HORV2 PDBID: 2BAA
HORV2	*H. vulgare*	active region	HORV2_crystal_struc_mature_m1.pdb
Q6IV09_DRORT	*D. rotundifolia*	active region	Q6IV09_DRORT_m1.pdb
Q6IV09_DRORT	*D. rotundifolia*	active region	Q6IV09_DRORT_mature_m1.pdb
CHI3_CASSA	*Castanea sativa*	C-rich domain, P-rich hinge, active region	CHI3_CASSA_m1.pdb
CHI3_CASSA	*Castanea sativa*	C-rich domain, P-rich hinge, active region	CHI3_CASSA_mature_m1.pdb
Q6IVX8_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region	Q6IVX8_9CARY_m1.pdb
Q6IVX8_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	Q6IVX8_9CARY_mature_m1.pdb
V5TEI0_DIOMU	*D. muscipula*	signal, C-rich domain, P-rich hinge, active region	V5TEI0_DIOMU_m1.pdb
V5TEI0_DIOMU	*D. muscipula*	C-rich domain, P-rich hinge, active region	V5TEI0_DIOMU_mature_m1.pdb
Q6DUJ9_DIOMU	*D. muscipula*	signal, C-rich domain, P-rich hinge, active region	Q6DUJ9_DIOMU_m1.pdb
Q6DUJ9_DIOMU	*D. muscipula*	C-rich domain, P-rich hinge, active region	6DUJ9_DIOMU_mature_m1.pdb
VJH3_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region	VJH3_9CARY_m1.pdb
VJH3_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	VJH3_9CARY_mature_m1.pdb
DCAP_5513	*D. capensis*	signal, C-rich domain, P-rich hinge, active region	DCAP_5513_m1.pdb
DCAP_5513	*D. capensis*	C-rich domain, P-rich hinge, active region	DCAP_5513_mature_m1.pdb
Q6DUKO_9CARY	*D. spatulata*	active region	Q6DUKO_9CARY_m1.pdb
Q6DUKO_9CARY	*D. spatulata*	active region	Q6DUKO_9CARY_mature_m1.pdb
DCAP_4817	*D. capensis*	signal, C-rich domain, P-rich hinge, active region	DCAP_4817_m1.pdb
DCAP_4817	*D. capensis*	C-rich domain, P-rich hinge, active region	DCAP_4817_mature_m1.pdb
CHI2_BRANA	*B. napus*	signal, C-rich domain, P-rich hinge, active region, CTE	CHI2_BRANA_m1.pdb
CHI2_BRANA	*B. napus*	C-rich domain, P-rich hinge, active region	CHI2_BRANA_mature_m1.pdb
Q6IV10_DRORT	*D. rotundifolia*	active region	Q6IV10_DRORT_m1.pdb
Q6IV10_DRORT	*D. rotundifolia*	active region	Q6IV10_DRORT_mature_m1.pdb
I0CMI2_DIOMU	*D. muscipula*	active region	I0CMI2_DIOMU_m1.pdb
I0CMI2_DIOMU	*D. muscipula*	active region	I0CMI2_DIOMU_mature_m1.pdb
I0CMI3_9CARY	*D. spatulata*	active region	I0CMI3_9CARY_m1.pdb
I0CMI3_9CARY	*D. spatulata*	active region	I0CMI3_9CARY_mature_m1.pdb
I0CMI4_9CARY	*D. spatulata*	active region	I0CMI4_9CARY_m1.pdb
I0CMI4_9CARY	*D. spatulata*	active region	I0CMI4_9CARY_mature_m1.pdb
I0CMI6_NEPMI	*N. mirabilis*	active region	I0CMI6_NEPMI_m1.pdb
I0CMI6_NEPMI	*N. mirabilis*	active region	I0CMI6_NEPMI_mature_m1.pdb
Q6IVX2_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region, CTE	Q6IVX2_9CARY_m1.pdb
Q6IVX2_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	Q6IVX2_9CARY_mature_m1.pdb
Q6IVX4_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region, CTE	Q6IVX4_9CARY_m1.pdb
Q6IVX4_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	Q6IVX4_9CARY_mature_m1.pdb
DCAP_0533	*D. capensis*	signal, C-rich domain, P-rich hinge, active region, C-terminal domain	DCAP_0533_m1.pdb
DCAP_0533	*D. capensis*	C-rich domain, P-rich hinge, active region, C-terminal domain	DCAP_0533_mature_m1.pdb
A9ZMK1_NEPAL	*N. alata*	signal, C-rich domain, P-rich hinge, active region	A9ZMK1_NEPAL_m1.pdb
A9ZMK1_NEPAL	*N. alata*	C-rich domain, P-rich hinge, active region	A9ZMK1_NEPAL_mature_m1.pdb

Appendix B

Supplement: Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*

Sequence Alignments

Sequence alignments for the esterase/lipases from *D. capensis* are shown along with annotation reference sequences from other plants. Cluster 1 (Figure B.1) contains enzymes with the traditional GDSL motif, including GDL1_CARPA from *Carica papaya*. Cluster 2 (Figure B.2) contains only sequences from *D. capensis*, while Cluster 3 (Figure B.3) contains two reference sequences from *Arabidopsis thaliana*. Cluster 4 is split into two figures for legibility (Figures SB.4 and SB.5). The alignment figures are annotated to highlight chemical properties of the amino acid residues as well as important sequence features. The amino acid

attributes are color-coded as follows: cysteines are yellow, positively charged residues are blue, negatively charged residues are red, hydrophobic residues are green, and all others are black. Highly conserved residues are indicated with a dot above the sequence position. The catalytic triad residues are marked with colored arrows. SignalP 4.1 [216] is used to predict the signal peptide cleavage site, which is specified by underlining the residues on either end of the cleavage point. The signal peptide itself is highlighted in light orange. Strikethrough text indicates sequence regions that are absent in the active enzyme, in this case the N-terminal signal peptide that is expressed but removed during maturation. Functional blocks I-IV are highlighted with colored boxes. Annotations were performed by homology to the annotations reference sequences from *C. papaya* and *A. thaliana* found in the UniProt database and identified by their UniProt IDs.

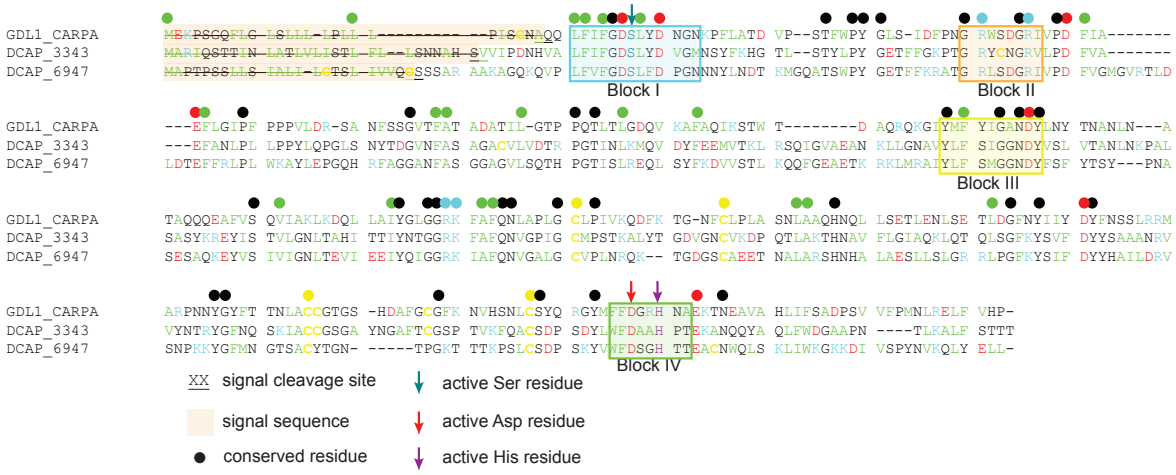


Figure B.1: Sequence alignment for Cluster 1 esterase/lipases, annotated by homology to the reference sequence GDL1_CARPA. The four functional blocks that are critical for enzyme function are highlighted using outlined colored boxes. The N-terminal signal peptide is highlighted in light orange. Colored arrows indicate the catalytic triad residues. Conserved residues are marked using colored dots: acidic (red), basic (blue), hydrophobic (green), and hydrophilic (black) residues.

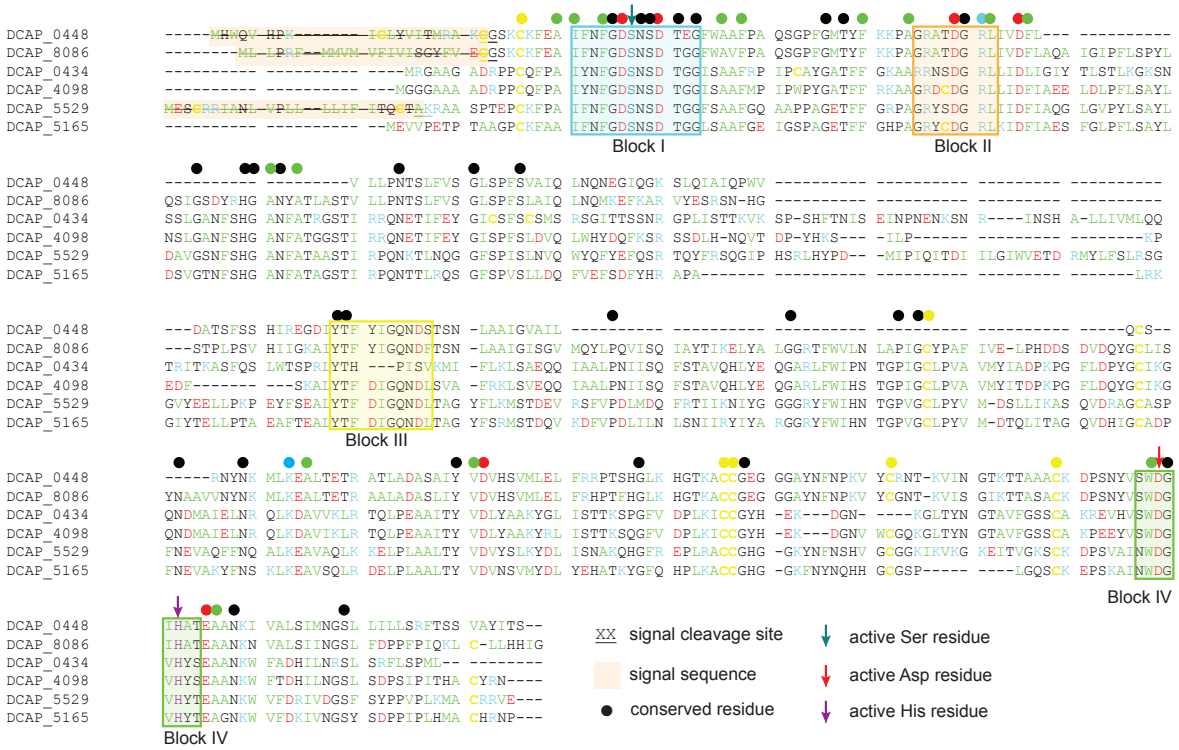


Figure B.2: Sequence alignment and annotation for Cluster 2. The four block regions are determined by sequence conservation and outlined with colored boxes. Three *D. capensis* esterase/lipases contain the N-terminal signal sequence (highlighted in light orange) and three lack it. The catalytic triad is indicated using colored arrows. Colored dots denote conserved residues.

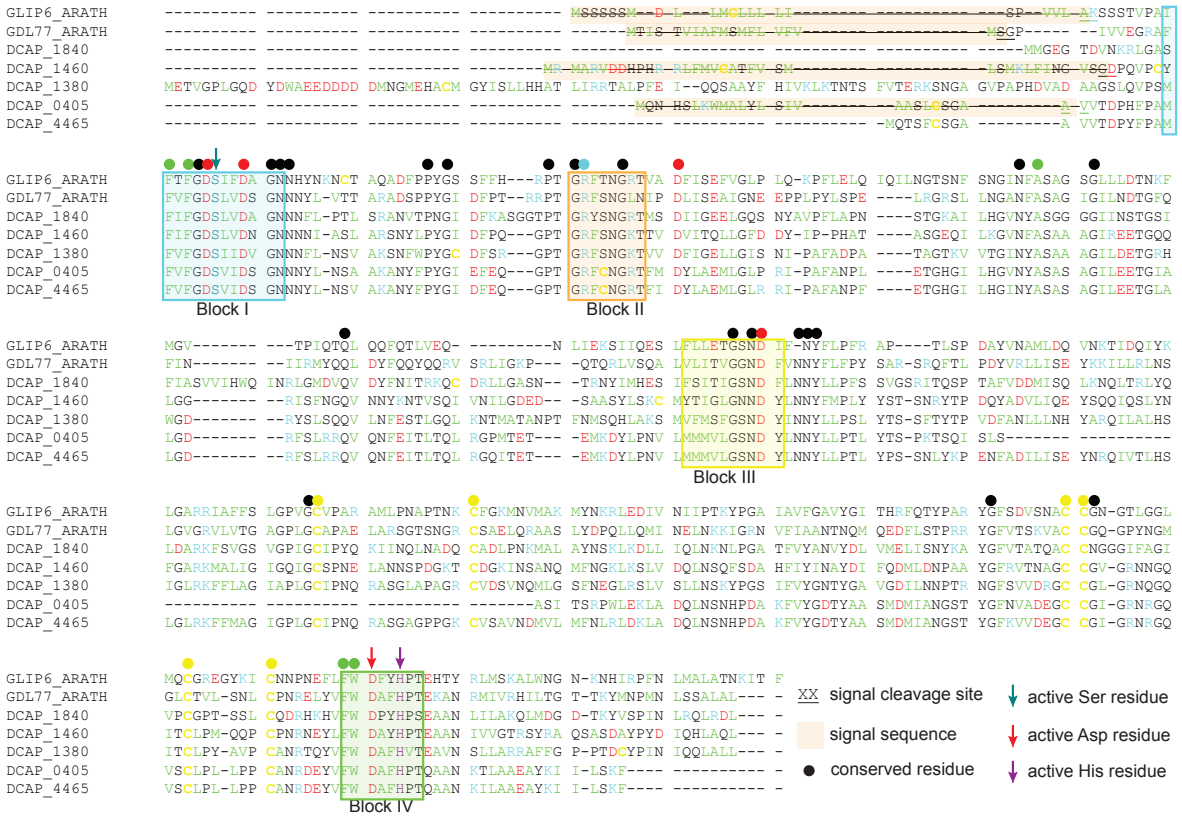


Figure B.3: Sequence alignment and annotation for Cluster 3. Reference sequences are GLIP6_ARATH and GDL77_ARATH. All but three Cluster 3 esterase/lipases contain a N-terminal signal peptide (highlighted in light orange). Functional block regions are outlined using colored boxes. Colored dots indicate conserved residues.

Figure B.4: Sequence alignment and annotation of Cluster 4a (first set), annotated by homology to EXL3_ARATH. Cluster 4 is separated into two parts (4a and 4b) for clarity. Block regions I-IV are shown in colored boxes with active site residues marked by colored arrows. Colored dots indicate conserved residues. When present, the N-terminal signal peptide is highlighted in light orange.

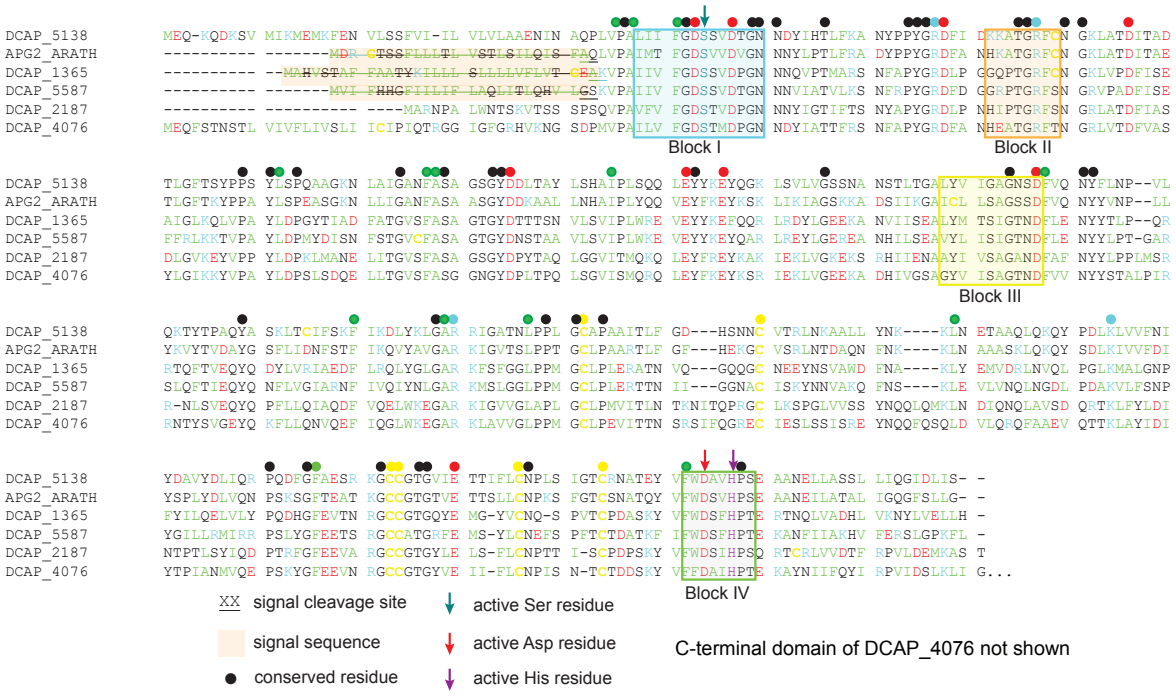


Figure B.5: Sequence alignment and annotation of Cluster 4b (second set), annotated by homology to APG2_ARATH. Cluster 4 is separated into two parts (4a and 4b) for clarity. Block regions I-IV are shown in colored boxes with active site residues marked by colored arrows. Colored dots indicate conserved residues. When present, the N-terminal signal peptide is highlighted in light orange. DCAP_4076 has an additional C-terminal domain (shown in Figure B.8).

Preliminary Structural Models and *In silico* Maturation

Preliminary models for the esterase/lipases were produced using the online Robetta implementation [228] of Rosetta [139]. The Rosetta structures contain the full sequences, including the N-terminal signal peptides that are cleaved during maturation. We performed *in silico* maturation, which we have previously described for cysteine proteases [38], for each protein. The initial Rosetta structure for each enzyme includes the signal peptide and lacks post-translational modifications. During *in silico* maturation, the signal sequence is removed and the structure is equilibrated for 500 ps in explicit TIP3P solvent using NAMD[219]. Figures of predicted structures were generated using Chimera [217]. Figure B.6A shows the workflow

of the overall enzyme discovery process. Panels (B) and (C) show an example of a Cluster 2 esterase/lipase, DCAP_8086, before (B) and after (C) the *in silico* maturation process. Further comparison of a Cluster 3 esterase/lipase (DCAP_1460) to Cluster 4 enzymes and a cutin synthase from *Solanum lycopersicum* (tomato), G1DEX3_SOLLC, is shown in Figure B.7. Functional sequence blocks DCAP_1460 and G1DEX3_SOLLC are highlighted by color (Figure B.7). DCAP_4076, has an additional C-terminal domain. A PSI-BLAST search for the sequence of this domain indicated that it is related to the negative regulator of systemic acquired resistance proteins previously discovered in other plants [330], with approximately 36% sequence identity to the SNI1 proteins from *Arabidopsis thaliana* (Uniprot ID: SNI1_ARATH) and *Glycine max* (Uniprot ID: Q0ZFU8_SOYBN). The *Arabidopsis* protein negatively regulates DNA recombination and gene expression during short-term stress responses. It has been suggested that SNI1_ARATH provides a scaffold for other proteins involved in regulation of transcription to bind; [193] it is possible that this domain is playing a similar role here. DCAP_4076 lacks the N-terminal secretion signal common to many of the esterase/lipases, suggesting an intracellular function (Figure B.8).

The template structures used by Rosetta to calculate the predicted structures for a representative esterase / lipase, DCAP_0434, are tabulated in Supplementary Table B.1.

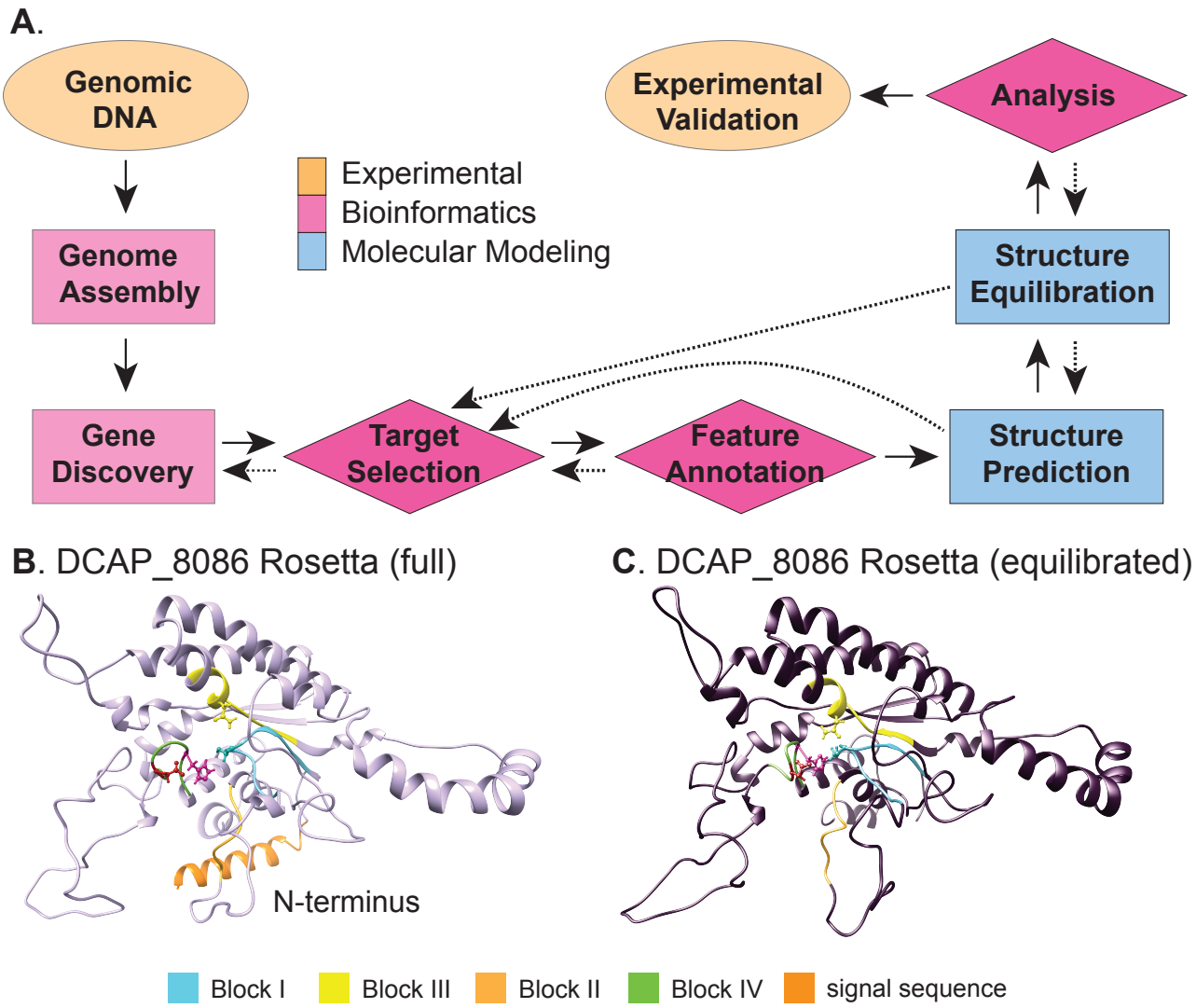


Figure B.6: (A) Flow chart illustrating the overall strategy for identifying enzymatic targets from genomic DNA. The workflow is indicated with solid arrows, while dotted arrows represent steps where information from a later stage of the pipeline enables refinement of earlier stages in an iterative manner. After genome sequencing, assembly, and gene discovery, target proteins are identified based on putative enzymatic activity. Functional sequence features are identified by analogy to annotation reference sequences found in the UniProt database. Structures are predicted using the Rosetta software, and equilibrated in explicit solvent after removal of sequence regions not present in the mature enzyme. Structures are compared using network analytic methods, enabling strategic selection of enzymes for experimental characterization in a future study. (B) DCAP_8086 before and (C) after *in silico* maturation. The light orange helix in part A is the N-terminal signal sequence, which is cleaved upon maturation. Important residues are color-coded as follows: dark cyan (catalytically active serine), red (active site aspartic acid), purple (active site histidine).

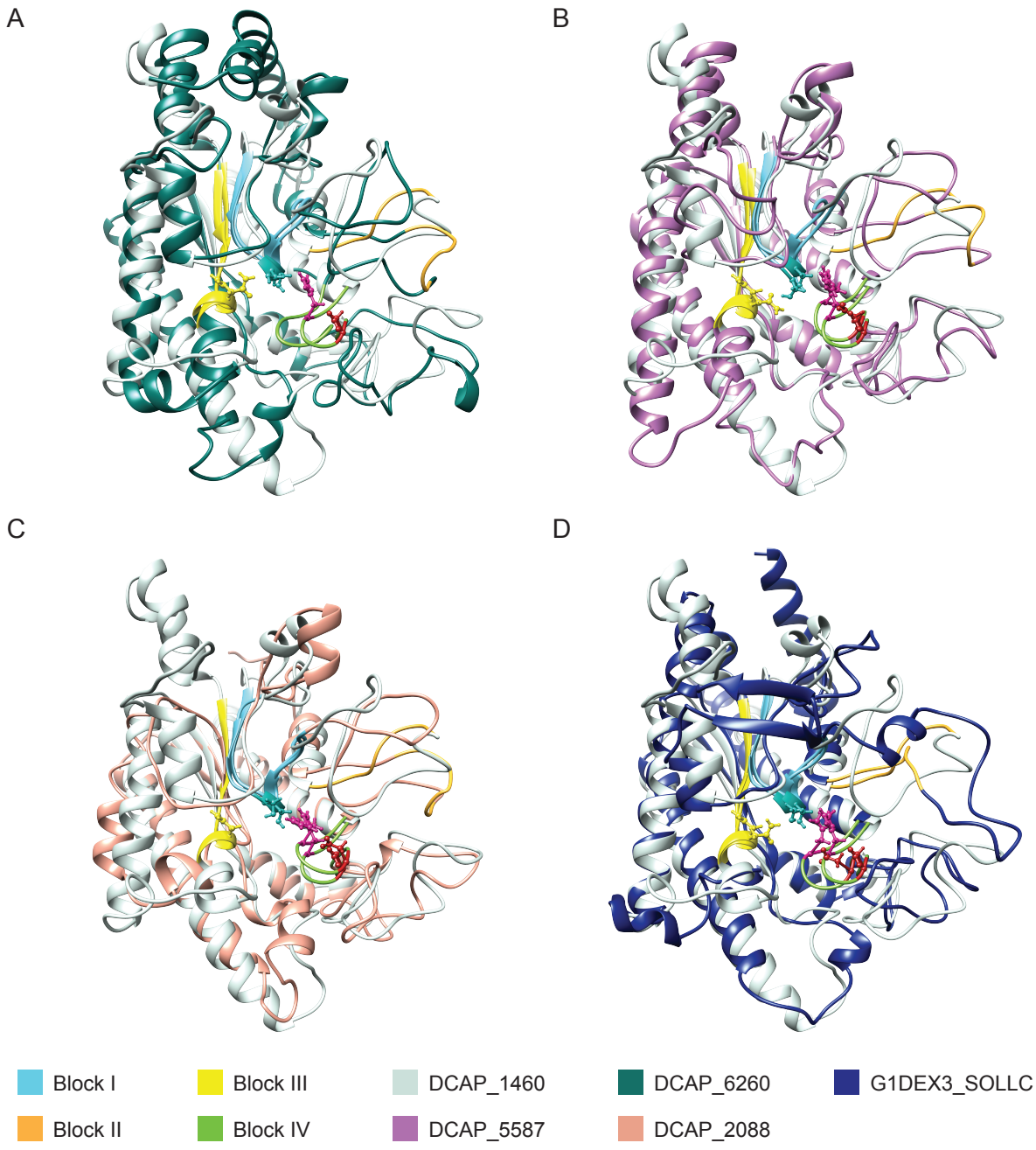


Figure B.7: Comparison of DCAP_1460 (Cluster 3) to *D. capensis* esterase/lipases from each of the other clusters. These pairwise alignments of structural models provide an indication of the type and magnitude of structural differences between clusters: in general, the overall fold and secondary structural elements is conserved, although considerable variation can be observed in their relative positions and the conformations of loops and termini. Alignment was performed using the matchmaker feature of Chimera with default settings [217]. Functional block regions I-IV are colored accordingly while the catalytic triad (Ser-His-Asp) residues are colored dark cyan, red, and purple. Active site residues are located in block I and IV, binding residues in block II-III. A. Comparison of DCAP_1460 to esterase/lipase DCAP_6260 (Cluster 4a). B. Comparison of DCAP_1460 to DCAP_5587 (Cluster 4b). C. Comparison of DCAP_1460 to DCAP_2088 (Cluster 4a). D. Comparison of DCAP_1460 to model esterase/lipase, G1DEX3_SOLLC, from *Solanum lycopersicum* (tomato).

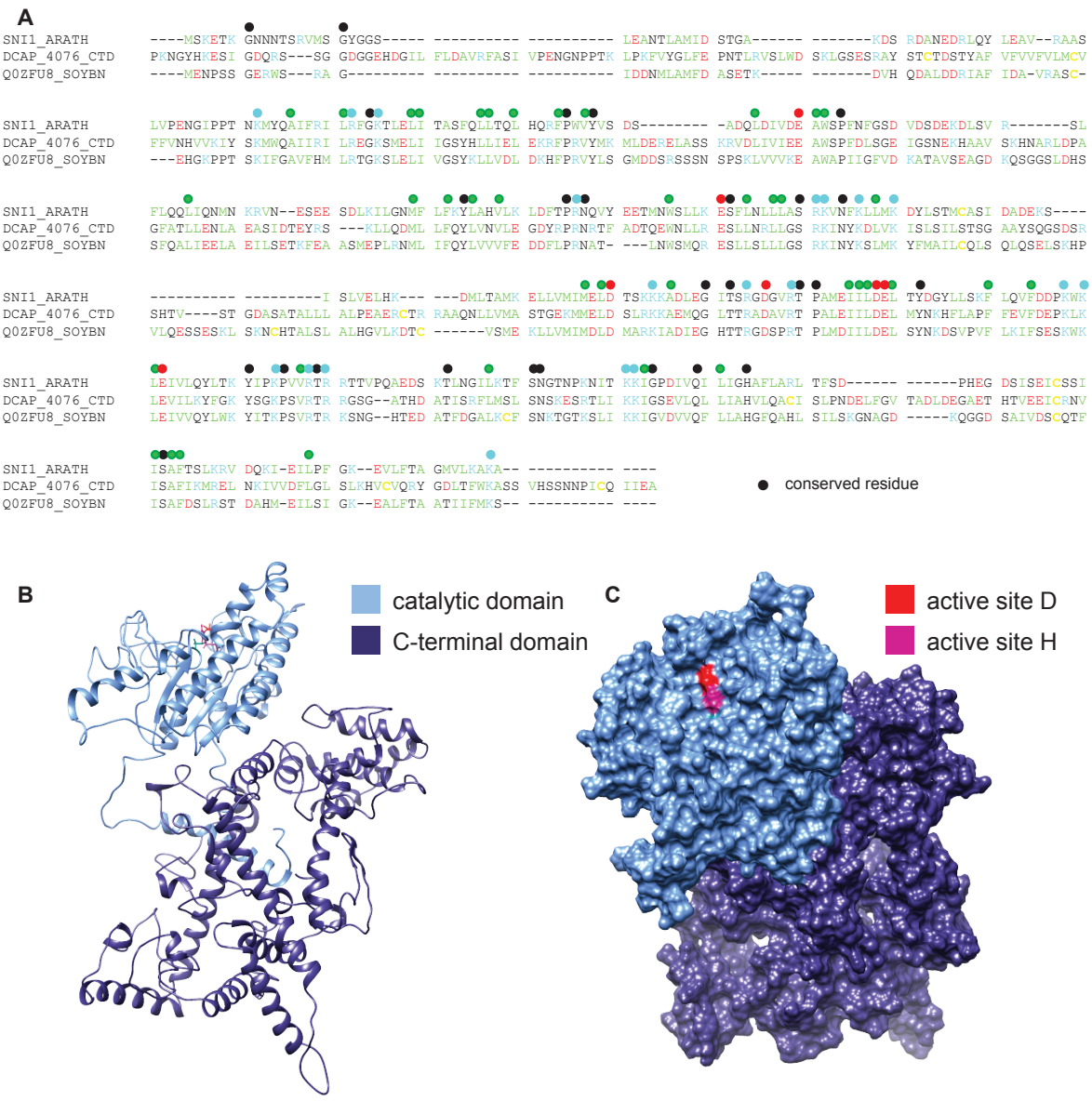


Figure B.8: A. Sequence alignment of the C-terminal domain of DCAP_4076 with the SNI1 proteins from *Arabidopsis thaliana* (Uniprot ID: SNI1_ARATH) and *Glycine max* (Uniprot ID: Q0ZFU8_SOYBN). B. Ribbon structure of DCAP_4076, with the catalytic domain in light blue and the C-terminal domain in dark blue. C. Structural model of DCAP_4076 showing the surface representation. The active site D (red) and H (magenta) residues are visible at the top of the model.

All initial and equilibrated structures available for download as PDB files are tabulated in Supplementary Tables 1 and 2, respectively.

Supplementary Table B.1: Rosetta structures for esterase / lipases (PDB files available for download)

Protein	Organism	Sequence Elements included	File Name
GDL1_CARPA	*Carica papaya*	signal, active region	GDL1_CARPA_m1.pdb
DCAP_3343	*D. capensis*	signal, active region	DCAP_3343_m1.pdb
DCAP_6947	*D. capensis*	signal, active region	DCAP_6947_m1.pdb
DCAP_0448	*D. capensis*	signal, active region	DCAP_0448_m1.pdb
DCAP_8086	*D. capensis*	signal, active region	DCAP_8086_m1.pdb
DCAP_0434	*D. capensis*	active region	DCAP_0434_m1.pdb
DCAP_4098	*D. capensis*	active region	DCAP_4098_m1.pdb
DCAP_5529	*D. capensis*	signal, active region	DCAP_5529_m1.pdb
DCAP_5165	*D. capensis*	active region	DCAP_5165_m1.pdb
GLIP6_ARATH	*A. thaliana*	signal, active region	GLIP6_ARATH_m1.pdb
GDL77_ARATH	*A. thaliana*	signal, active region	GDL77_ARATH_m1.pdb
DCAP_1840	*D. capensis*	active region	DCAP_1840_m1.pdb
DCAP_1460	*D. capensis*	signal, active region	DCAP_1460_m1.pdb
DCAP_1380	*D. capensis*	active region	DCAP_1380_m1.pdb
DCAP_0405	*D. capensis*	signal, active region	DCAP_0405_m1.pdb
DCAP_4465	*D. capensis*	active region	DCAP_4465_m1.pdb
DCAP_6218	*D. capensis*	active region	DCAP_6218_m1.pdb
DCAP_6260	*D. capensis*	active region	DCAP_6260_m1.pdb
EXL3_ARATH	*A. thaliana*	signal, active region	EXL3_ARATH _m1.pdb
DCAP_1761	*D. capensis*	active region	DCAP_1761_m1.pdb
DCAP_6217	*D. capensis*	signal, active region	DCAP_6217_m1.pdb
DCAP_5461	*D. capensis*	signal, active region	DCAP_5461_m1.pdb
DCAP_0158	*D. capensis*	signal, active region	DCAP_0158_m1.pdb
DCAP_2088	*D. capensis*	active region	DCAP_2088_m1.pdb
DCAP_2089	*D. capensis*	active region	DCAP_2089_m1.pdb
DCAP_5138	*D. capensis*	active region	DCAP_5138_m1.pdb
APG2_ARATH	*A. thaliana*	signal, active region	APG2_ARATH _m1.pdb
DCAP_1365	*D. capensis*	signal, active region	DCAP_1365_m1.pdb
DCAP_5587	*D. capensis*	signal, active region	DCAP_5587_m1.pdb
DCAP_2187	*D. capensis*	active region	DCAP_2187_m1.pdb
DCAP_4076	*D. capensis*	active region	DCAP_4076_m1.pdb

Supplementary Table B.2: Mature structures for esterase / lipases (PDB files available for download)

Protein	Organism	Sequence Elements included	File Name
GDL1_CARPA	*Carica papaya*	active region	GDL1_CARPA_mature_m1.pdb
DCAP_3343	*D. capensis*	active region	DCAP_3343_mature_m1.pdb
DCAP_6947	*D. capensis*	active region	DCAP_6947_mature_m1.pdb
DCAP_0448	*D. capensis*	active region	DCAP_0448_mature_m1.pdb
DCAP_8086	*D. capensis*	active region	DCAP_8086_mature_m1.pdb
DCAP_0434	*D. capensis*	active region	DCAP_0434_mature_m1.pdb
DCAP_4098	*D. capensis*	active region	DCAP_4098_mature_m1.pdb
DCAP_5529	*D. capensis*	active region	DCAP_5529_mature_m1.pdb
DCAP_5165	*D. capensis*	active region	DCAP_5165_mature_m1.pdb
GLIP6_ARATH	*A. thaliana*	active region	GLIP6_ARATH_mature_m1.pdb
GDL77_ARATH	*A. thaliana*	active region	GDL77_ARATH_mature_m1.pdb
DCAP_1840	*D. capensis*	active region	DCAP_1840_mature_m1.pdb
DCAP_1460	*D. capensis*	active region	DCAP_1460_mature_m1.pdb
DCAP_1380	*D. capensis*	active region	DCAP_1380_mature_m1.pdb
DCAP_0405	*D. capensis*	active region	DCAP_0405_mature_m1.pdb
DCAP_4465	*D. capensis*	active region	DCAP_4465_mature_m1.pdb
DCAP_6218	*D. capensis*	active region	DCAP_6218_mature_m1.pdb
DCAP_6260	*D. capensis*	active region	DCAP_6260_mature_m1.pdb
EXL3_ARATH	*A. thaliana*	active region	EXL3_ARATH _mature_m1.pdb
DCAP_1761	*D. capensis*	active region	DCAP_1761_mature_m1.pdb
DCAP_6217	*D. capensis*	active region	DCAP_6217_mature_m1.pdb
DCAP_5461	*D. capensis*	active region	DCAP_5461_mature_m1.pdb
DCAP_0158	*D. capensis*	active region	DCAP_0158_mature_m1.pdb
DCAP_2088	*D. capensis*	active region	DCAP_2088_mature_m1.pdb
DCAP_2089	*D. capensis*	active region	DCAP_2089_mature_m1.pdb
DCAP_5138	*D. capensis*	active region	DCAP_5138_mature_m1.pdb
APG2_ARATH	*A. thaliana*	active region	APG2_ARATH _mature_m1.pdb
DCAP_1365	*D. capensis*	active region	DCAP_1365_mature_m1.pdb
DCAP_5587	*D. capensis*	active region	DCAP_5587_mature_m1.pdb
DCAP_2187	*D. capensis*	active region	DCAP_2187_mature_m1.pdb
DCAP_4076	*D. capensis*	active region	DCAP_4076_mature_m1.pdb

Supplementary Table B.3: Templates used for structure prediction of DCAP_0434, designated by PDBID.

PDBID	protein	organism	citation
3KVN (A)	EstA	*Pseudomonas aeruginosa*	[295]
3KVN (X)	EstA	*Pseudomonas aeruginosa*	[295]
1ESC (A)	Streptomyces scabies esterase	*Streptomyces scabiei*	[321]
3RJT (A)	lipolytic protein	*Alicyclobacillus acidocaldarius*	[52]
3MIL (A)	isoamyl acetate- hydrolyzing esterase	*Saccharomyces cerevisiae*	[167]
4JJ6 (A)	Axe2 variant H194A	*Geobacillus stearothermophilus*	[147]
4OAP (A)	Axe2 variant W190I	*Geobacillus stearothermophilus*	[146]
3W7V (A)	Axe2	*Geobacillus stearothermophilus*	[147]
4JKO (A)	Axe2 variant S15A	*Geobacillus stearothermophilus*	[148]
4HYQ (A)	phospholipase A1	*Streptomyces albidoflavus* NA297	[194]
4ZR8 (A)	uroporphyrinogen decarboxylase	*Acinetobacter baumannii*	[5]
4WSH (B)	probable uroporphyrinogen decarboxylase	*Pseudomonas aeruginosa*	[4]
4R7G (A)	Phosphoribosylformylglycinamidine synthase	*Salmonella enterica*	[281]

Active Site Network Constraint Measures

To assess the extent to which each active site was structurally constrained, four base constraint measures and one derived measure (the first principal component of these measures following standardization) were computed as described in the main text. Figure B.9 shows the values of each studied protein on five constraint measures; proteins are ordered vertically by rank on the omnibus site constraint measure.

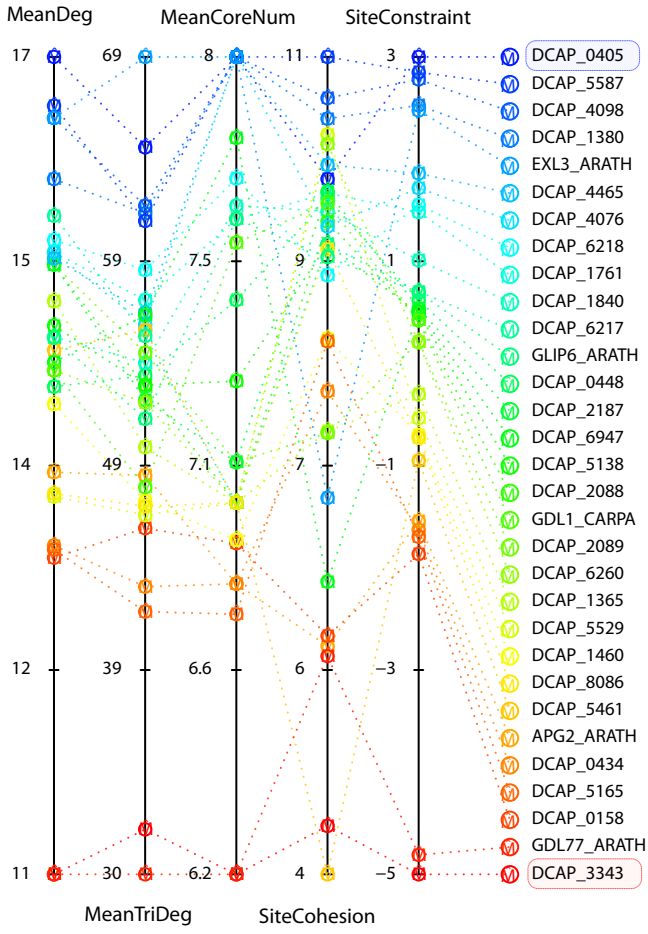


Figure B.9: Constraint measures for the active site networks. Vertical axes indicate values on each of the four base constraint measures and the omnibus derived measure, as described in the main text.

Appendix C

Supplement: Elucidation of WW domain ligand binding specificities in the Hippo pathway reveals STXBP4 as YAP inhibitor

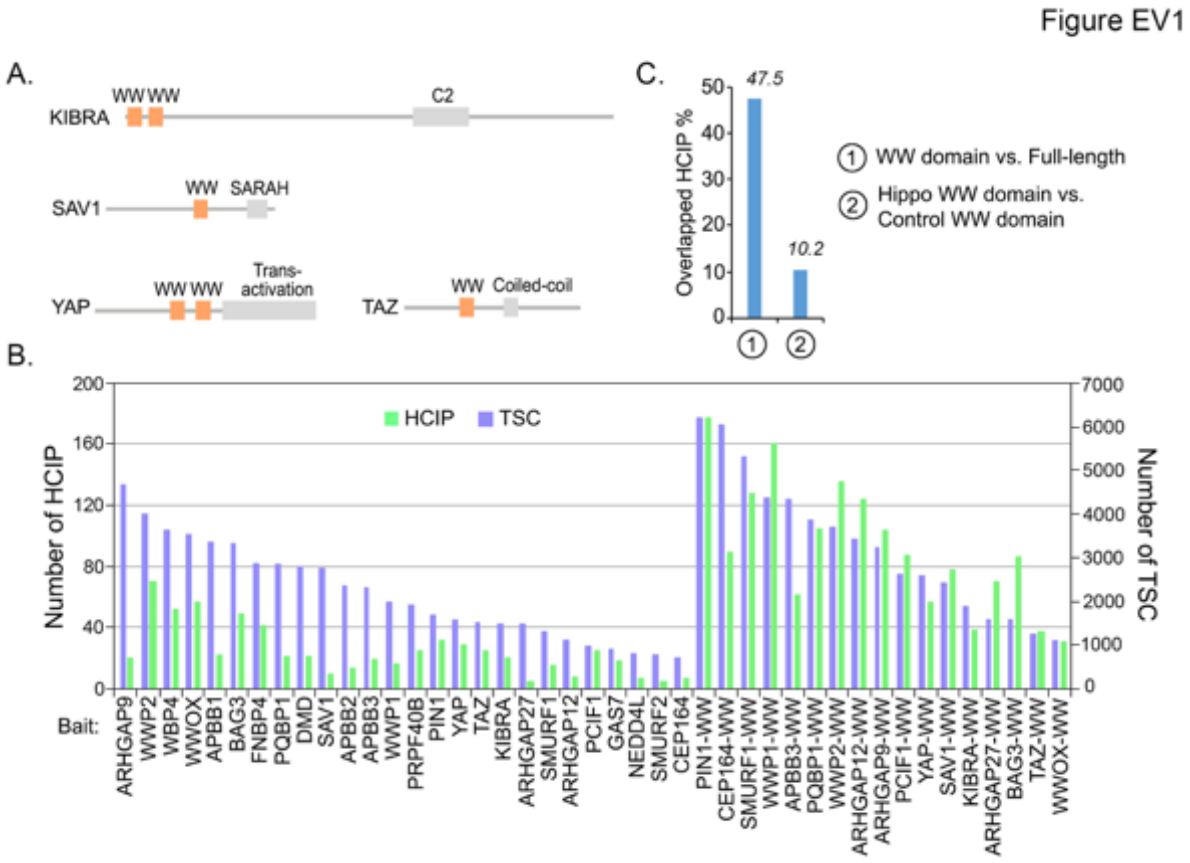


Figure C.1: **Proteomic analysis of the WW-containing proteins. (This figure is related to Figure 4.1).**
(A) Schematic illustration of the Hippo WW domain-containing components.
(B) The total spectral counts (TSCs) and corresponding numbers of HCIPs for the indicated proteomic experiments.
(C) The overlapped HCIP rate was respectively compared for the full-length protein and its WW domain, and Hippo WW domains and control WW domains.

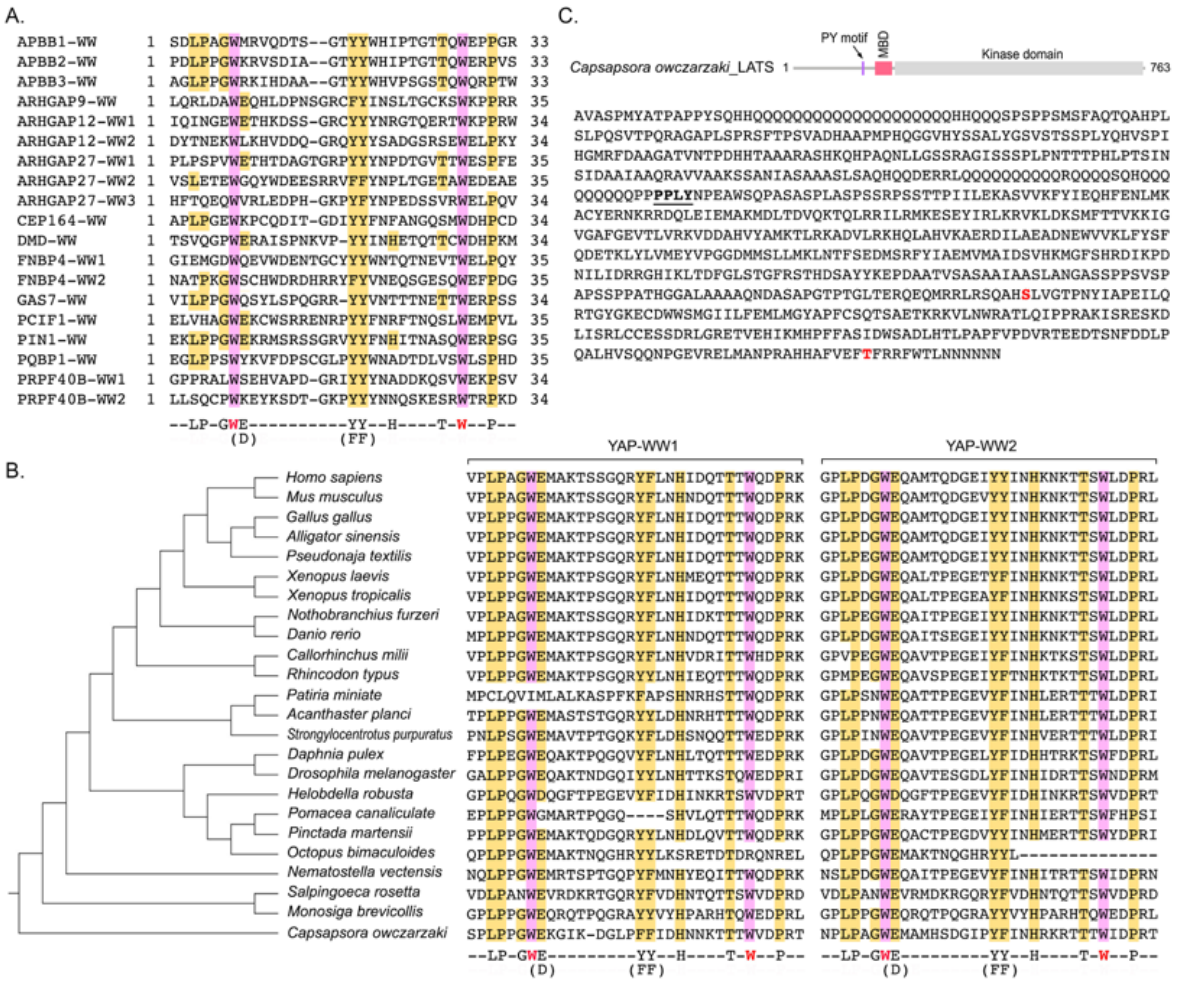


Figure C.2: **Analyses of the identified 9-amino acid sequence in both control WW domains and evolution. (This figure is related to Figure 4.2; Appendix Figure C.6 and C.7; Table S5).**

(A) Sequence alignment of the WW domains derived from the control WW domain-containing proteins that cannot bind the Hippo PY motif-containing proteins. The two conserved tryptophan restudies were highlighted in purple, and the identified 9 amino acid residues were highlighted in yellow.

(B) Evolutionary analysis of the YAP WW domains. The identified 9-amino acid sequence is highlighted in the two YAP WW domains, respectively.

(C) A PY motif is identified in *Capsapsora owczarzaki* LATS. Schematic illustration of the *Capsapsora owczarzaki* LATS protein, where the PY motif is indicated. MBD, MOB1-binding domain. The PY motif is underlined in the *Capsapsora owczarzaki* LATS protein sequence, where the auto-phosphorylation site (S586) and the phosphorylation site (T750) in the hydrophobic motif were shown in red.

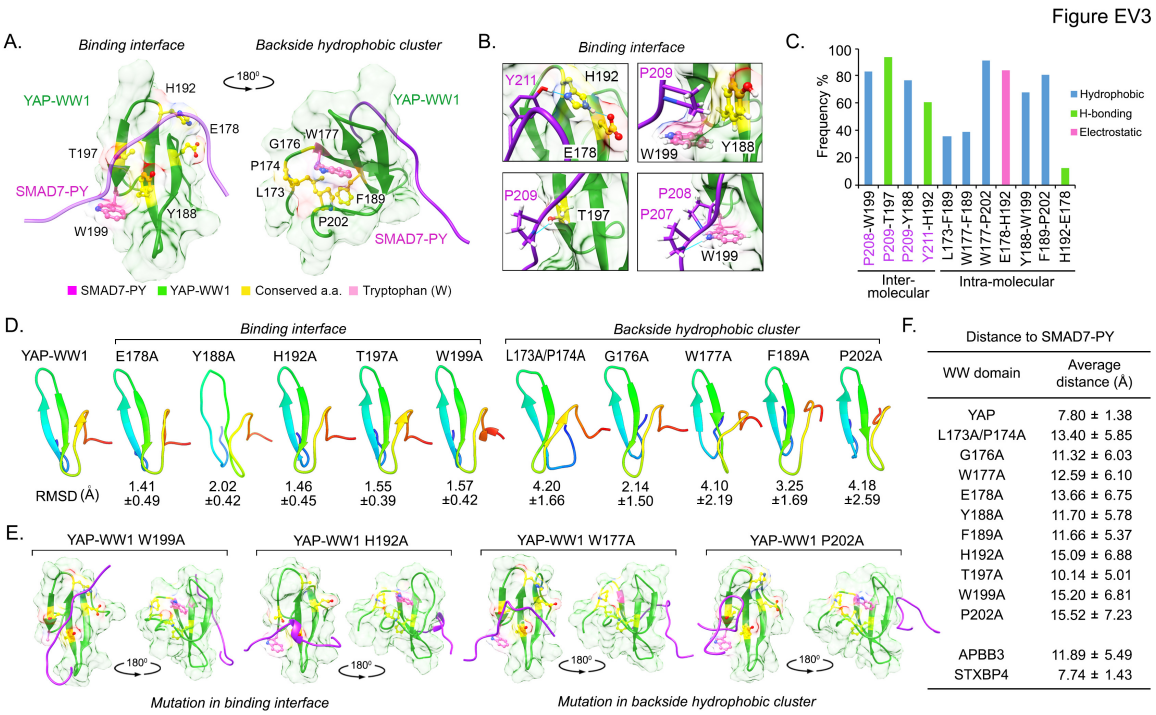


Figure C.3: **Structural analysis of the identified 9-amino acid sequence. (This figure is related to Appendix Figure C.8)**
(A) Illustration of the identified 9-amino acid residues in the average YAP-WW1/SMAD7-PY structure, the initial structure derived from NMR solution structure (2LTW). SMAD7-PY peptide was adjusted to 50% transparence to show the residue details on the binding interface.
(B) Four contact regions within the YAP-WW1/SMAD7-PY complex were shown in details from the representative top cluster structures with key residues indicated. Residues from SMAD7-PY motif peptide were labeled in purple. Hydrogen bond is indicated in blue line.
(C) The binding types and the corresponding frequency rates were shown for the indicated inter- and intra-molecular residue pairs.
(D) Simulation analysis of apo YAP-WW1 domain and its indicated mutants. RMSD value for each mutant simulation (referenced against the average apo YAP-WW1 domain) was shown.
(E) Average structures of the indicated the YAP-WW1 mutant/SMAD7-PY complexes.
(F) The average distance between SMAD7-PY motif peptide and the indicated WW domains was summarized in a table.

Figure EV4

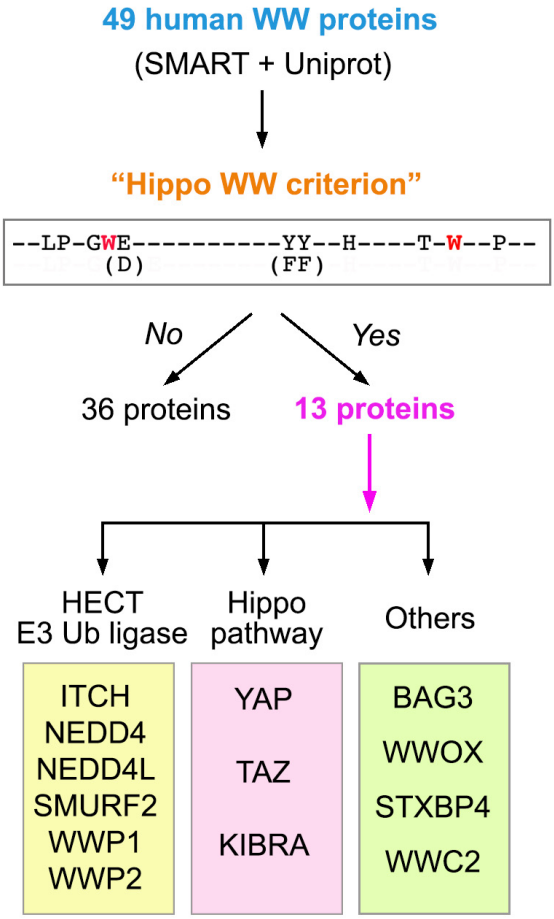


Figure C.4: **Schematic illustration of the human proteome search for the WW domain-containing proteins that fit the Hippo WW domain 9-amino acid sequence criterion. (This figure is related to Figure 4.3; Table S6).**
The identified 9-amino acid sequence was subjected to the 49 WW domain-containing proteins in human proteome. Total 13 WW domain-containing proteins were found to fit the Hippo WW domain criterion.

Figure EV5

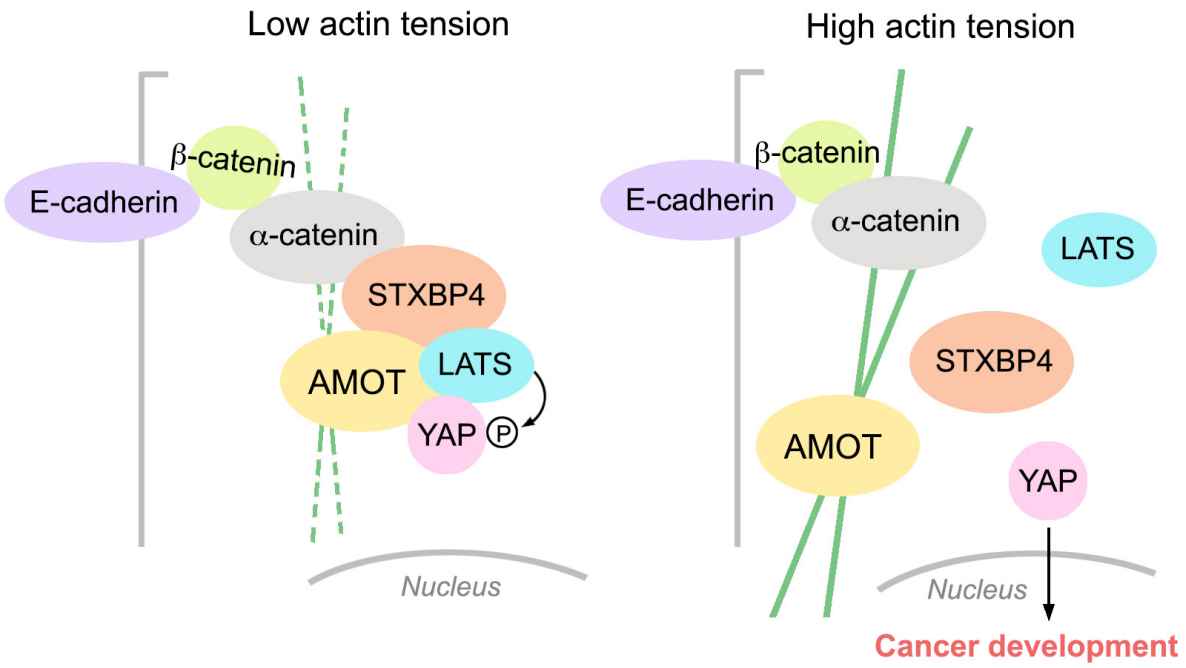


Figure C.5: **A proposed model for STXBP4-mediated Hippo pathway regulation in response to actin cytoskeleton tension change. (This figure is related to Figures 4.4 and 4.5).**
Under low actin tension, STXBP4 assembles a protein complex comprising α-catenin, AMOT, LATS and YAP to promote YAP phosphorylation and cytoplasmic retention. When actin cytoskeleton tension increases, the STXBP4-centered protein complex is dissembled, resulting in YAP dephosphorylation and nuclear translocation as well as the cancer development.

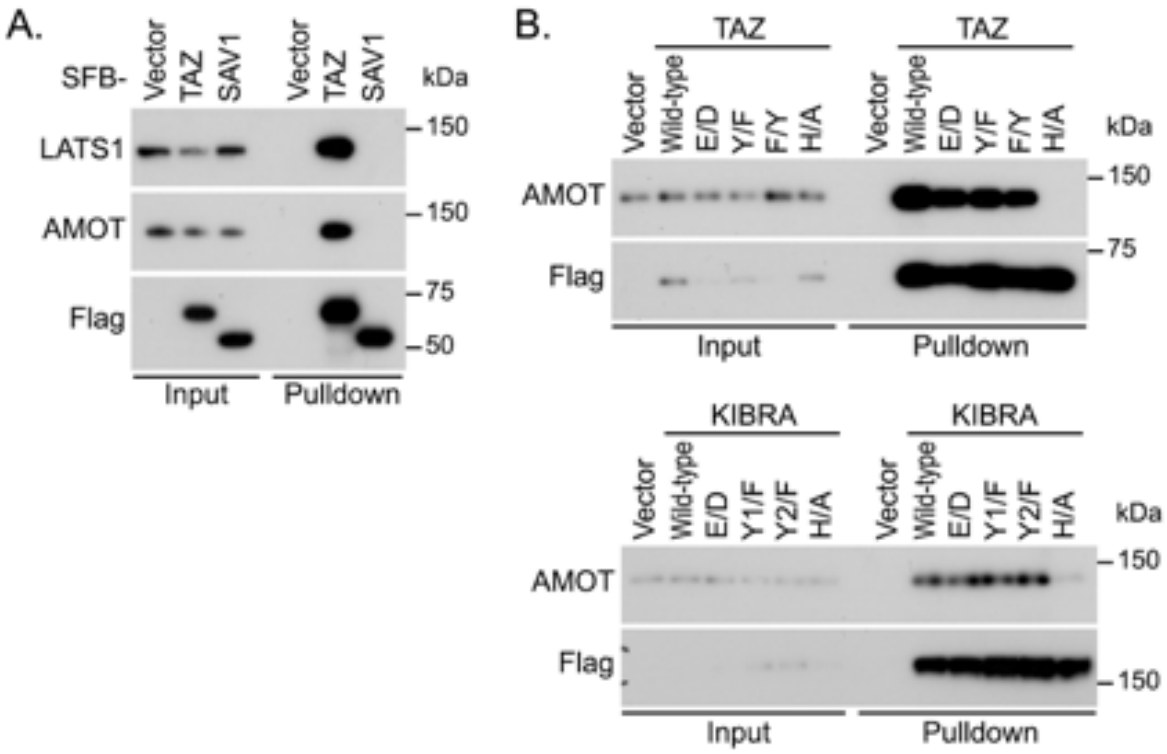


Figure C.6: **Characterization of Hippo WW domain binding specificity. (This figure is related to Figure 4.2).**
(A) Hippo pathway components TAZ but not SAV1 interacts with AMOT and LATS1. HEK293T cells were transfected with the indicated SFB-tagged constructs and subjected to the pulldown assay.
(B) Examination of the conservative substitution mutations for the identified 9-amino acid sequence. HEK293T cells were transfected with the indicated SFB-tagged constructs and subjected to the pulldown assay. The tandem tyrosine residues within the 9-amino acid sequence of KIBRA were indicated as Y1 and Y2, respectively.

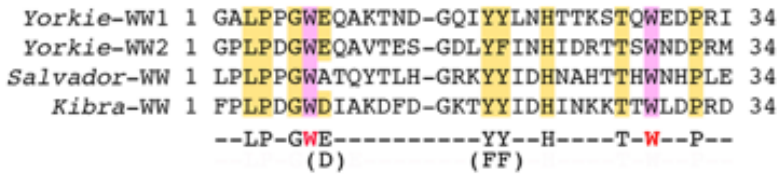


Figure C.7: **Examination of the identified 9-amino acid sequence for the Drosophila Hippo pathway components. (This figure is related to Figure 4.2 and Figure C.2).** Sequence alignment of the WW domains derived from the *Drosophila* Hippo WW domain-containing proteins. The two conserved tryptophan restudies were highlighted in purple. As compared with the 9-amino acid sequence, additional conserved amino acid residues were highlighted in yellow.

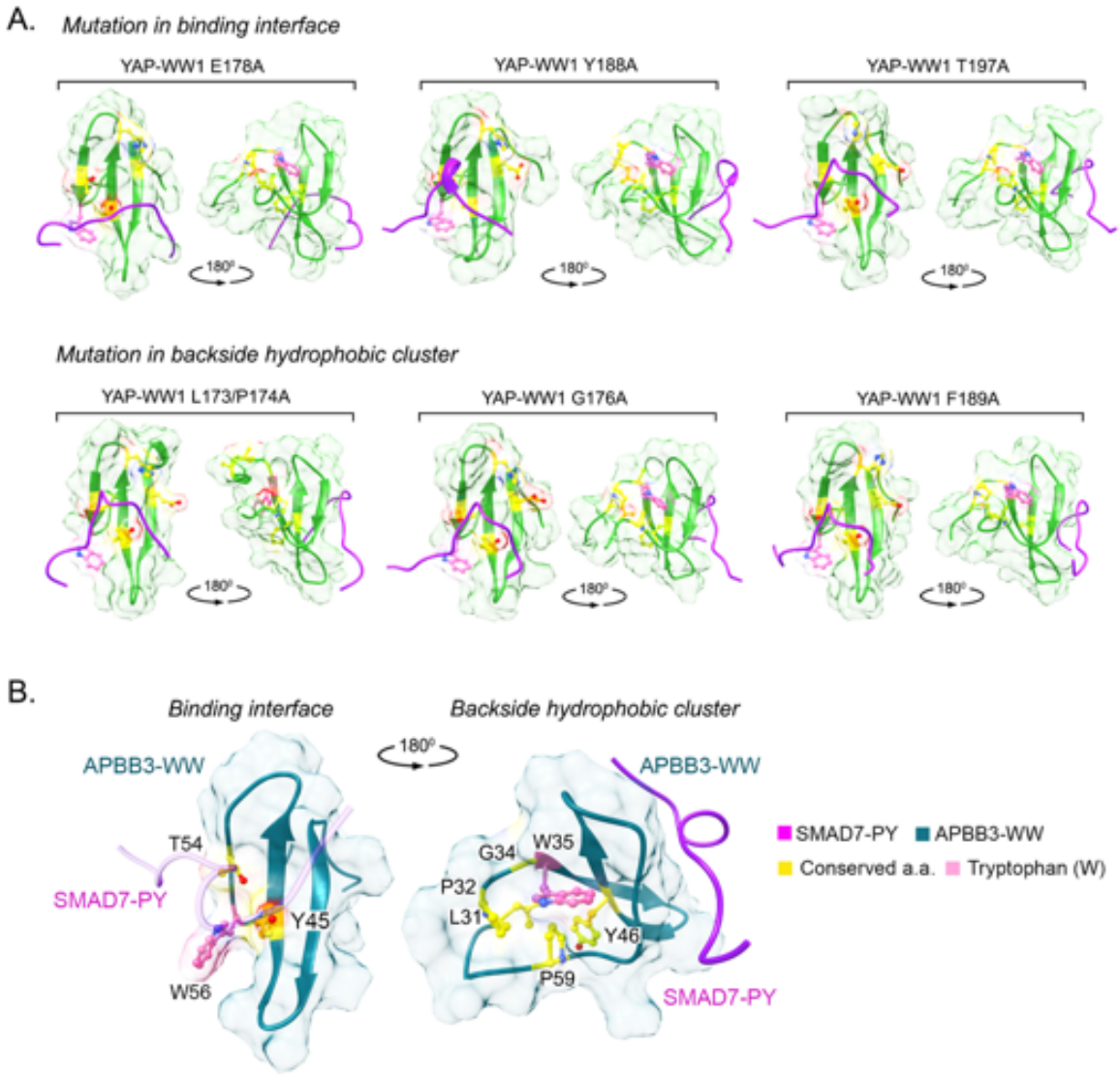


Figure C.8: **Characterization of the identified 9-amino acid sequence through simulation analyses. (This figure is related to Figure C.3).**
(A) Simulation analysis of the indicated YAP-WW1 mutant/SMAD7-PY complexes. (B) Illustration of the identified 9-amino acid sequence in the APBB3-WW/SMAD7-PY complex. The NMR solution structure of the APBB3-WW domain (2YSC) was used for simulation. SMAD7-PY peptide was adjusted to 50% transparence to show the residue details on the binding interface.

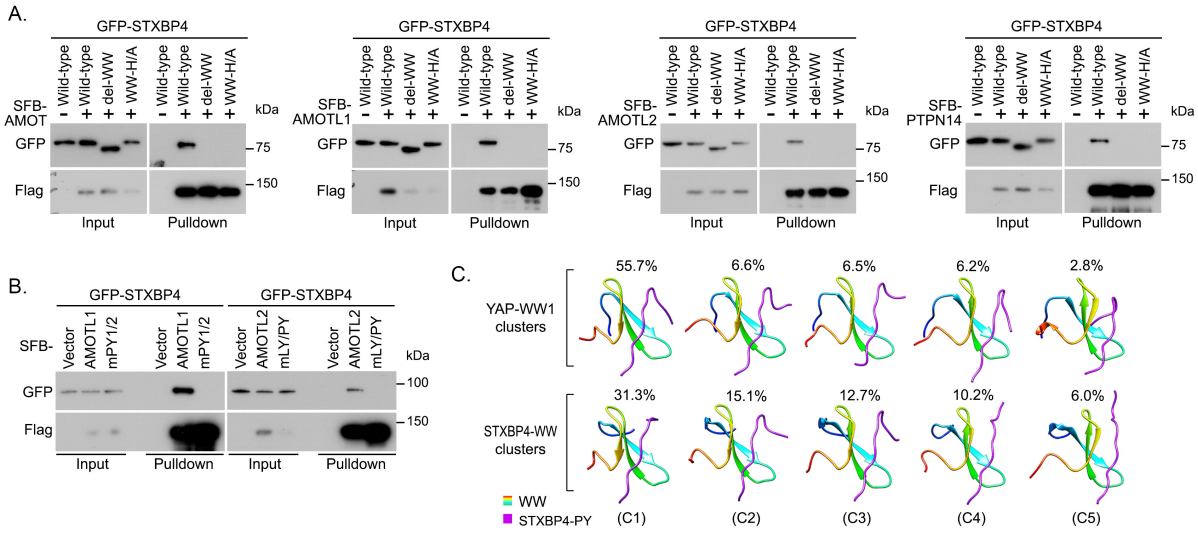


Figure C.9: **STXBP4 associates with the Hippo PY motif-containing proteins. (This figure is related to Figure 4.3).**
(A and B) The association between STXBP4 and the indicated Hippo PY motif-containing proteins is mediated by the WW domain (A) and PY motif (B). HEK293T cells were transfected with the indicated constructs and subjected to the pulldown assay.
(C) Simulation of the STXBP4-WW and SMAD7-PY complex structure. The top five WW-PY structure clusters were shown for both YAP-WW1/SMAD7-PY and STXBP4-WW/SMAD7-PY complexes. The frequency rate was shown for each cluster. C, cluster.

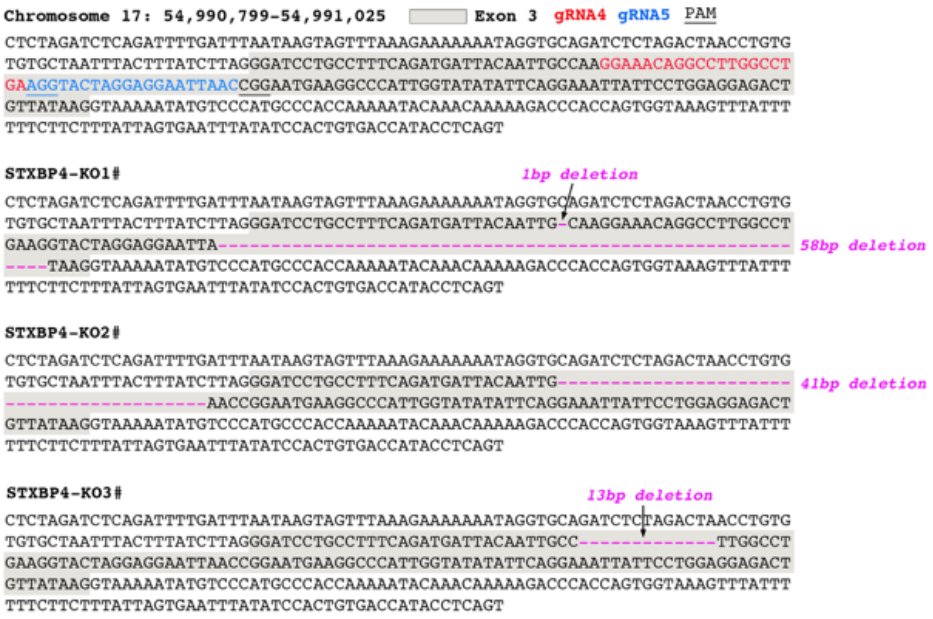


Figure C.10: **Genomic DNA sequencing results for the STXBP4 knockout (KO) cell lines as generated via CRISPR/Cas9. (This figure is related to Figure C.3).** Among the five designed guide RNAs (gRNAs), only the gRNA4 and gRNA5-targeted region shows genomic editing for all the three STXBP4 KO cell lines. The genomic editing details for each STXBP4 KO cell line were indicated.

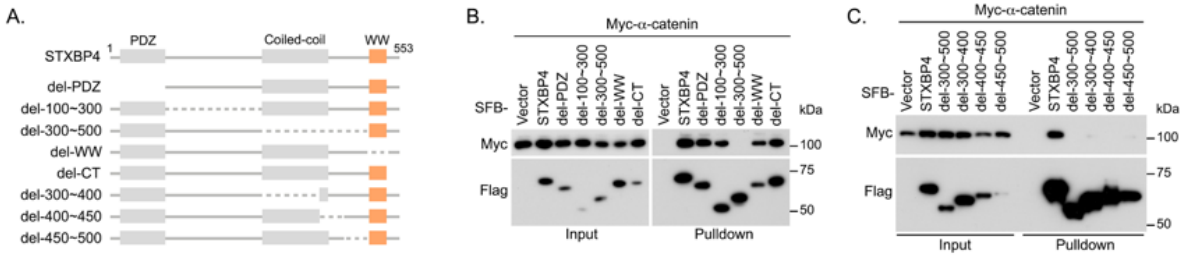


Figure C.11: **STXBP4 interacts with α-catenin. (This figure is related to Figure 4.4).**
(A) Schematic illustration of a series of STXBP4 protein truncation and deletion mutants used in this study.
(B and C) Mapping the α-catenin binding region in STXBP4. An internal region (300 500 residues) of STXBP4 is required to associate with α-catenin (B), and we failed to further narrow down the binding region within the 300$\tilde{5}$00 residues (C).

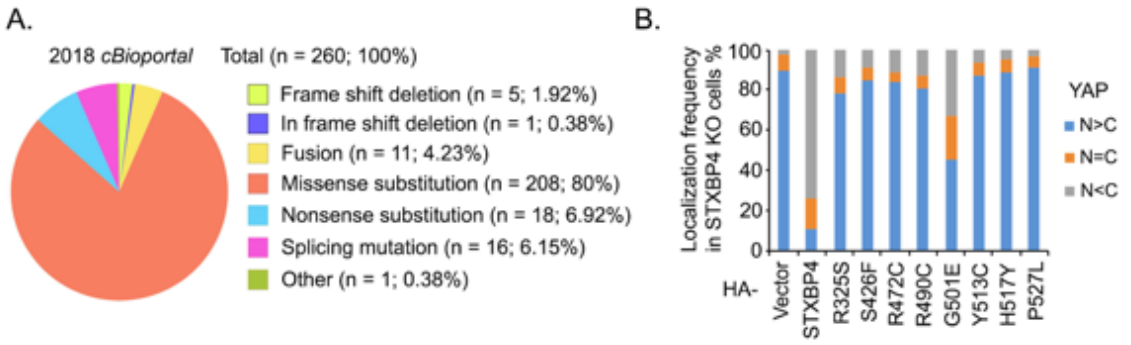


Figure C.12: **STXBP4 binds α-catenin and AMOT to regulate YAP. (This figure is related to Figure 4.4).**
(A) Summary of STXBP4 mutations in cBioportal web database (http://www.cbioportal.org).
(B) Interactions with α-catenin and the Hippo PY motif-containing proteins are both required for the STXBP4-mediated YAP suppression. The indicated STXBP4 mutants were expressed in the STXBP4 KO cells and immunofluorescent staining was performed. HA-positive cells from 30 different views (200 cells in total) were randomly selected and quantified for YAP localization.

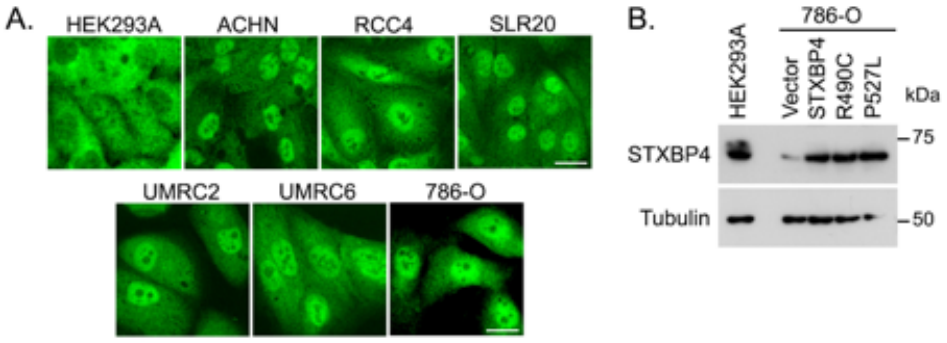


Figure C.13: **STXBP4 is a potential tumor suppressor in kidney cancer. (This figure is related to Figure 4.5).**
(A) YAP is highly enriched in ccRCC cancer cell lines. YAP cellular localization is detected by immunofluorescent staining. Scale bar, 20 μm.
(B) STXBP4 protein expression is examined in the 786-O cells overexpressing STXBP4 and its cancer-derived mutants.

Table C.1: Simulation Conditions

Structure	Number of Simulations	PDB ID	Temperature (K)	Start - End Time per sim. (Âţs)	Ions & Waters
(WT) YAP- WW1 & SMAD7	3	2LTW	300	0.4-1	1 Na+, 3147waters
(WT) STXBP4-WW & SMAD7	3	2YSG, 2LTW	300	0.6-1	3 Na+, 4222waters
(WT) APBB3-WW & SMAD7	3	2YSC,	300	0-1	1 Cl-, 3897-5433 waters
(Mutant) YAP-WW1 L173A/P174A & SMAD7	3	2LTW	300	0-1	1 Na+, 3048-4496 waters
(Mutant) YAP-WW1 G176A & SMAD7	3	2LTW	300	0-1	1 Na+, 2880-3528 waters
(Mutant) YAP-WW1 W177A & SMAD7	3	2LTW	300	0-1	1 Na+,3235-4118 waters
(Mutant) YAP-WW1 E178A & SMAD7	3	2LTW	300	0-1	3365-3546waters
(Mutant) YAP-WW1 Y188A & SMAD7	3	2LTW	300	0-1	1 Na+, 3292-4430 waters
(Mutant) YAP-WW1 F189A & SMAD7	3	2LTW	300	0-1	1 Na+, 3260-3831 waters
(Mutant) YAP-WW1 H192A & SMAD7	3	2LTW	300	0-1	1 Na+, 2916-4162 waters
(Mutant) YAP-WW1 T197A & SMAD7	3	2LTW	300	0-1	1 Na+, 2872-3139 waters
(Mutant) YAP-WW1 W199A & SMAD7	3	2LTW	300	0-1	1 Na+, 3608-3974 waters
(Mutant) YAP-WW1 P202A & SMAD7	3	2LTW	300	0-1	1 Na+, 2872-3139 waters
(WT) apo YAP-WW1	3	2LTW	300	0-1	2870 waters
(Mutant) apo YAP-WW1 L173A/P174A	3	2LTW	300	0-1	2669-2739waters
(Mutant) apo YAP-WW1 G176A	3	2LTW	300	0-1	2761-2998 waters
(Mutant) apo YAP-WW1 W177A	3	2LTW	300	0-1	2839-2988 waters
(Mutant) apo YAP-WW1 E178A	3	2LTW	300	0-1	1Cl-, 2813-3088 waters
(Mutant) apo YAP-WW1 Y188A	3	2LTW	300	0-1	2661-3097waters
(Mutant) apo YAP-WW1 F189A	3	2LTW	300	0-1	2882-2996 waters
(Mutant) apo YAP-WW1 H192A	3	2LTW	300	0-1	2834-2974 waters
(Mutant) apo YAP-WW1 T197A	3	2LTW	300	0-1	2822-3022 waters
(Mutant) apo YAP-WW1 W199A	3	2LTW	300	0-1	2788-2934 waters
(Mutant) apo YAP-WW1 P202A	3	2LTW	300	0-1	2773-3008 waters

Appendix D

Supplement: Computational Studies of Intrinsically Disordered Proteins

D.1 Cumulative Averages of Observables

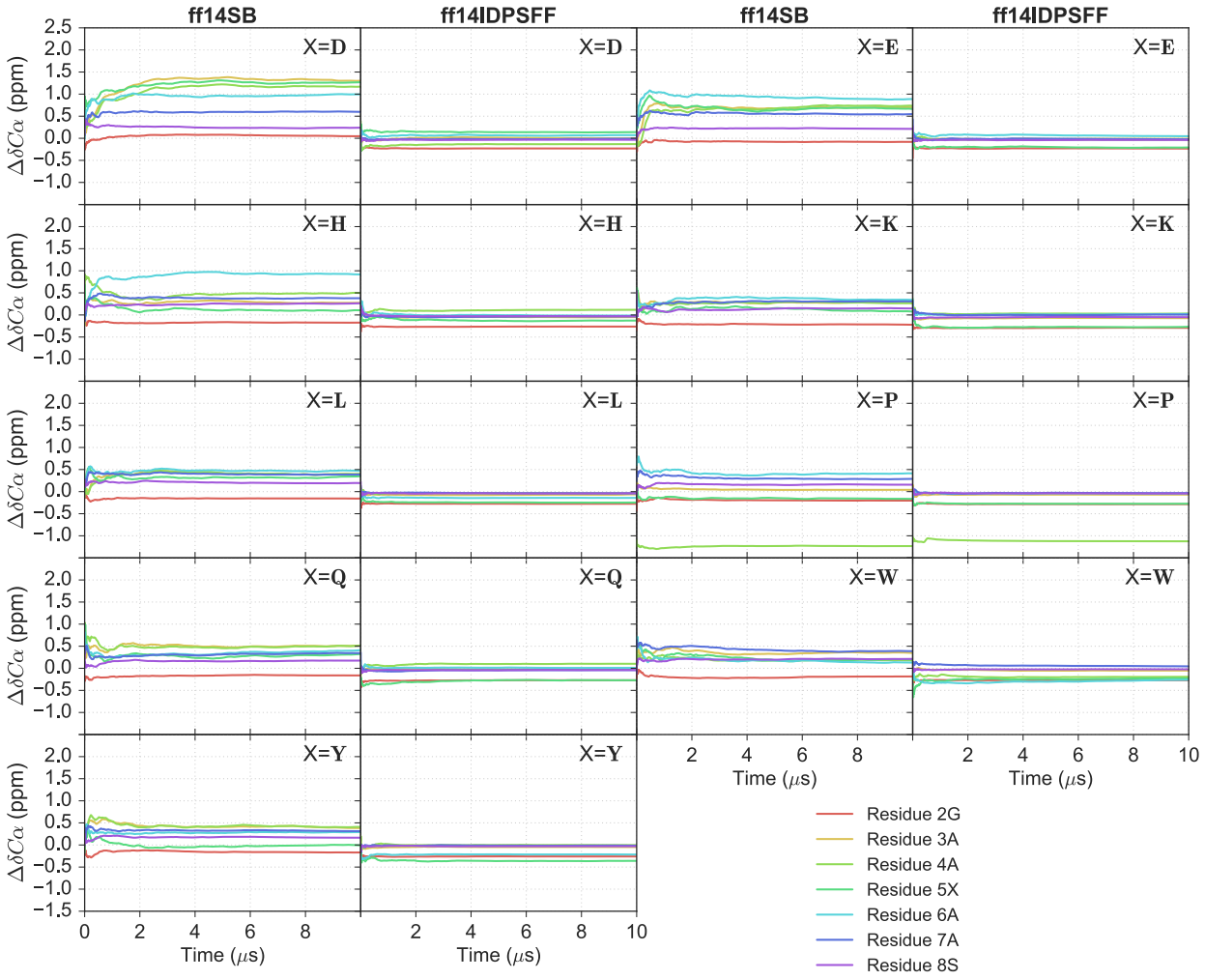


Figure D.1: The $\Delta\delta C\alpha$-derived cumulative averages per EGAAXAASS peptide and force field type were calculated and averaged between the 10 simulations. The first/third column is populated with short peptides simulated using the ff14SB and the second/fourth column is populated by the corresponding peptide simulated using the ff14IDPSFF. Each row represents an EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) peptide.

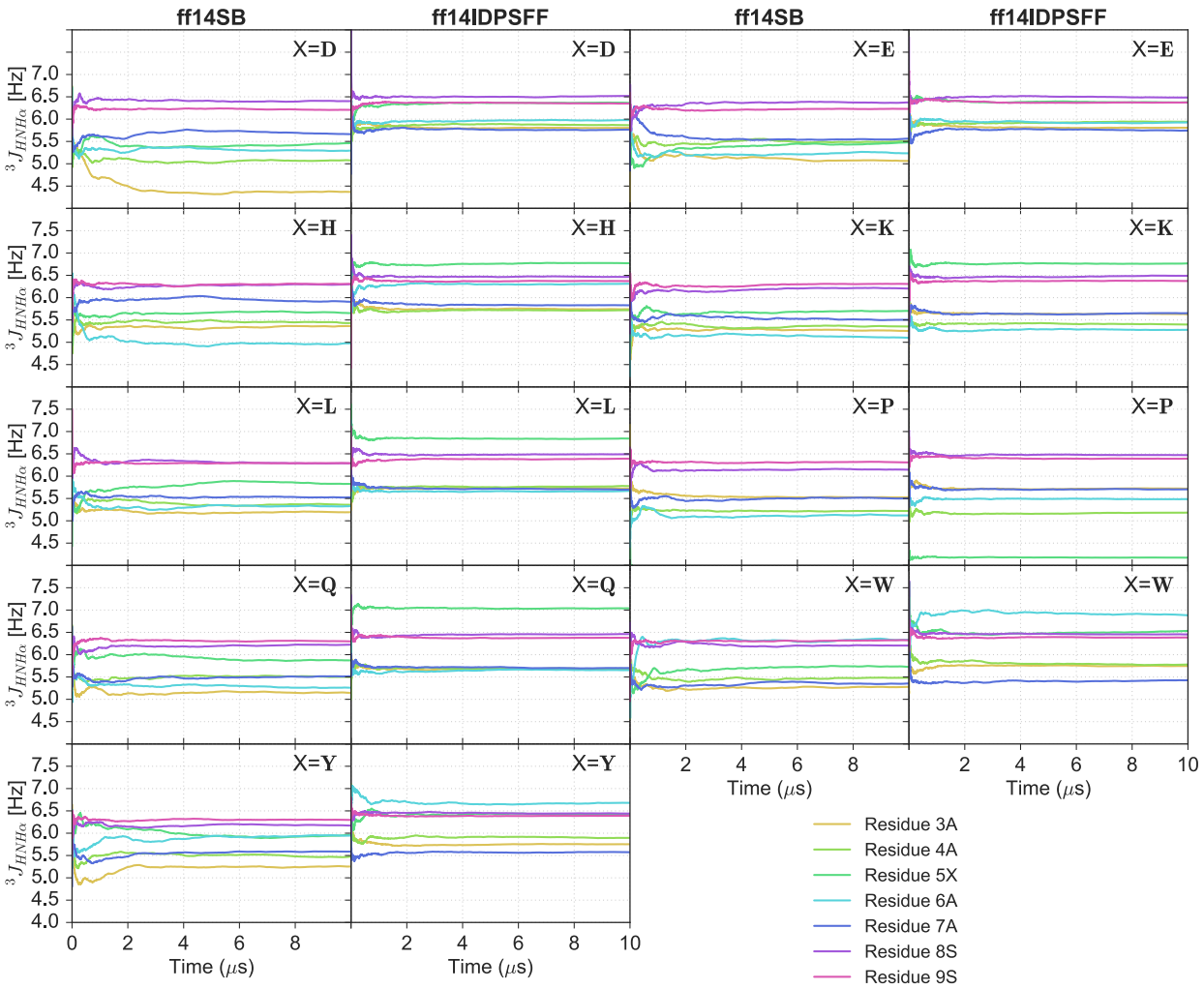


Figure D.2: The $^3J_{HNH\alpha}$-derived cumulative averages per EGAAXAASS peptide and force field type were calculated and averaged between the 10 simulations. The first/third column is populated with short peptides simulated using the ff14SB and the second/fourth column is populated by the corresponding peptide simulated using the ff14IDPSFF. Each row represents an EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) peptide.

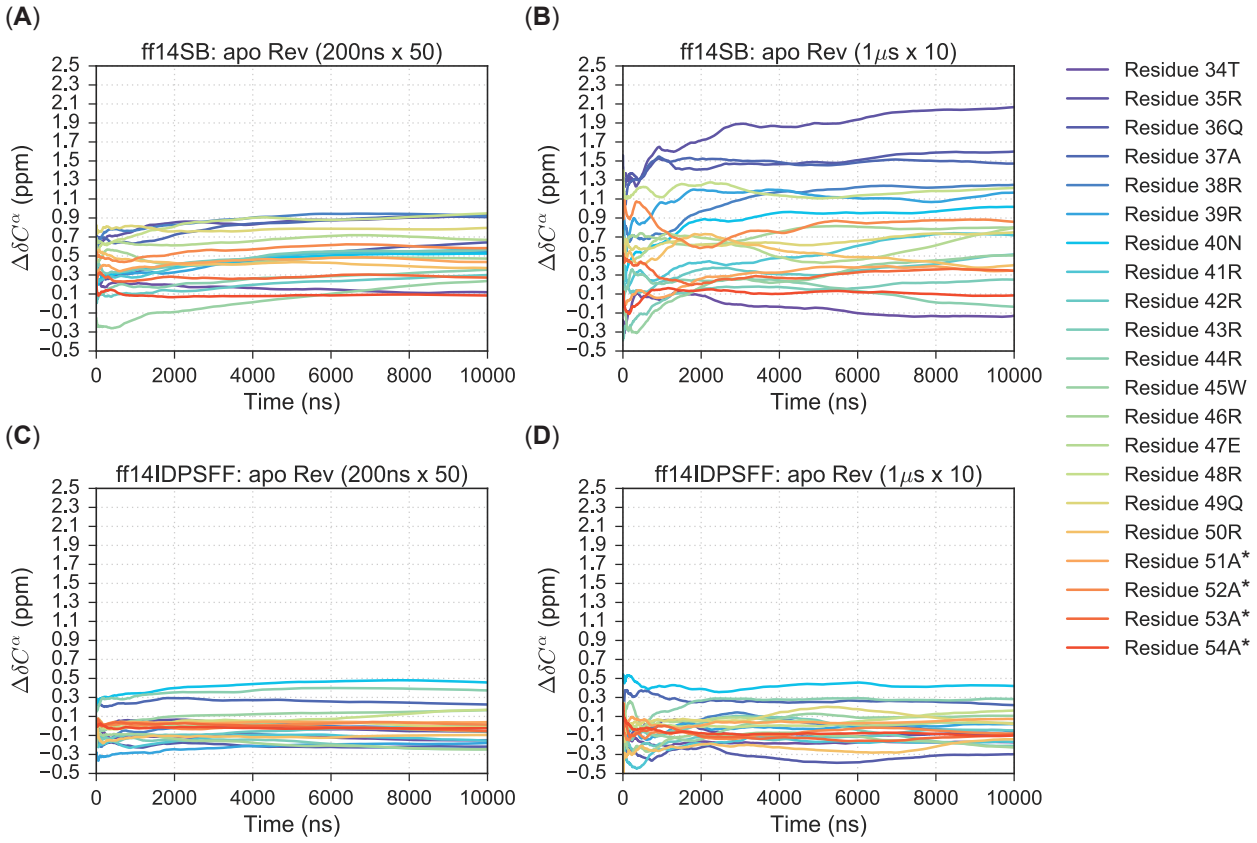


Figure D.3: The $\Delta\delta$Cα-derived cumulative averages per apo Rev peptide and force field type were calculated and averaged between the 10/50 simulations. Two simulations types were generated: fifty 200ns simulations using (A) ff14SB (B) and ff14IDPSFF, (C) and ten 1 μs simulations using ff14SB (D) and ff14IDPSFF. Residues are colored according to the legend with an asterisk (*) indicating non-native residues.

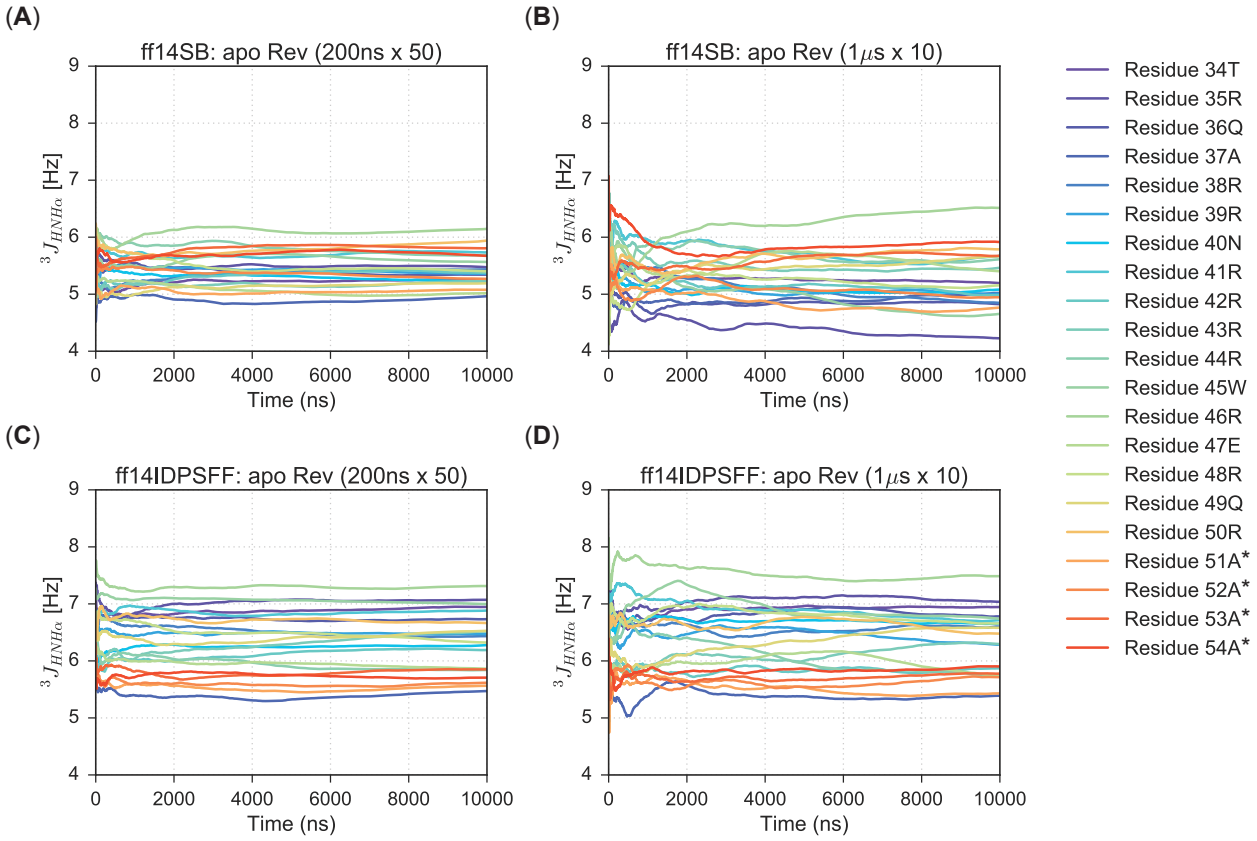


Figure D.4: The $3J_{HNH\alpha}$-derived cumulative averages per apo Rev peptide and force field type were calculated and averaged between the 10/50 simulations. Two simulations types were generated: fifty 200ns simulations using (A) ff14SB (B) and ff14IDPSFF, (C) and ten 1 μs simulations using ff14SB (D) and ff14IDPSFF. Residues are colored according to the legend with an asterisk (*) indicating non-native residues.

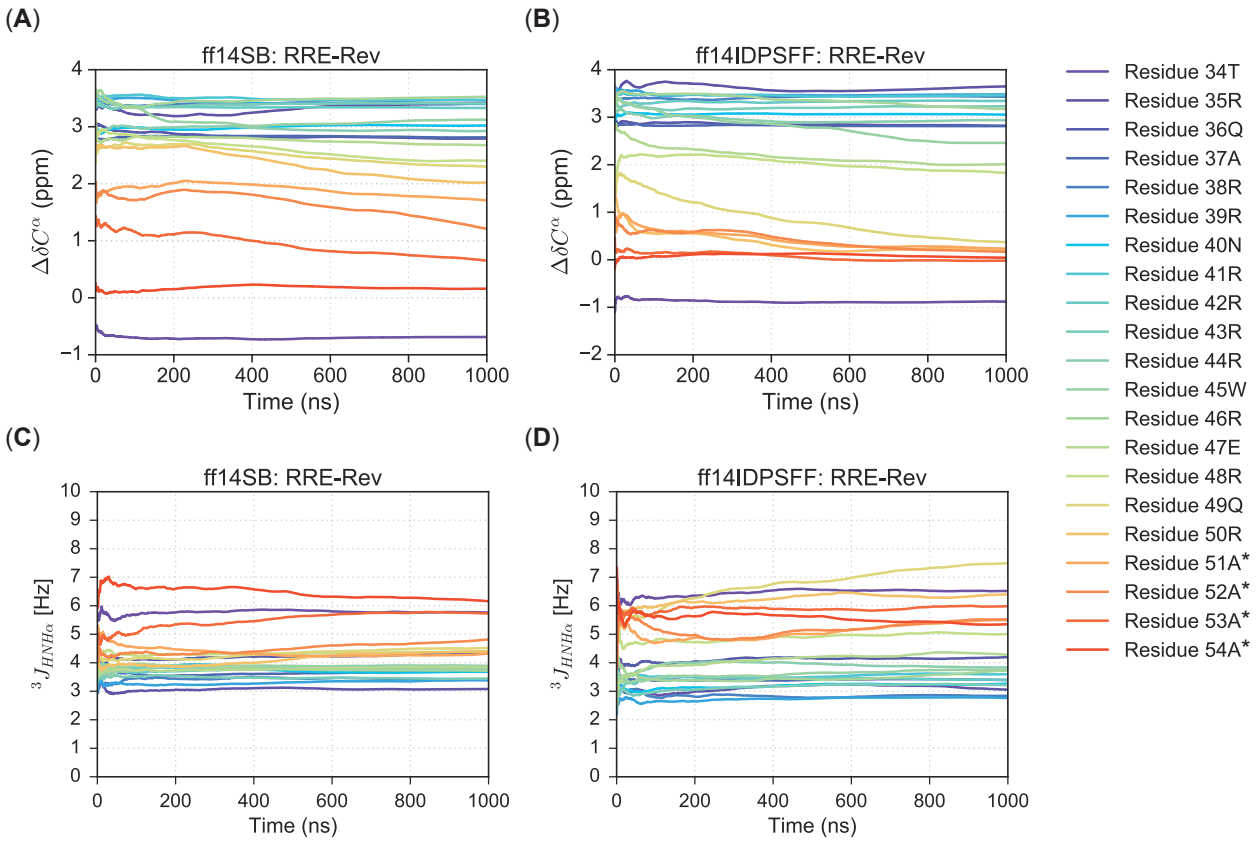


Figure D.5: The $\Delta\delta$Cα- and $3J_{HNH\alpha}$-derived cumulative averages per RRE-Rev complex and force field type were calculated and averaged between the 5 simulations. Secondary chemical shifts occupy the first row from (A) ff14SB-generated simulations and (B) ff14IDPSFF-generated simulations. $^{3}J_{HNH\alpha}$-coupling constants occupy the second row from (C) ff14SB-generated simulations and (D) ff14IDPSFF-generated simulations. Residues are colored according to the legend with an asterisk (*) indicating non-native residues.

D.2 Biphasic Exponential Fitting of $\Delta\Delta\delta C\alpha$ Datasets

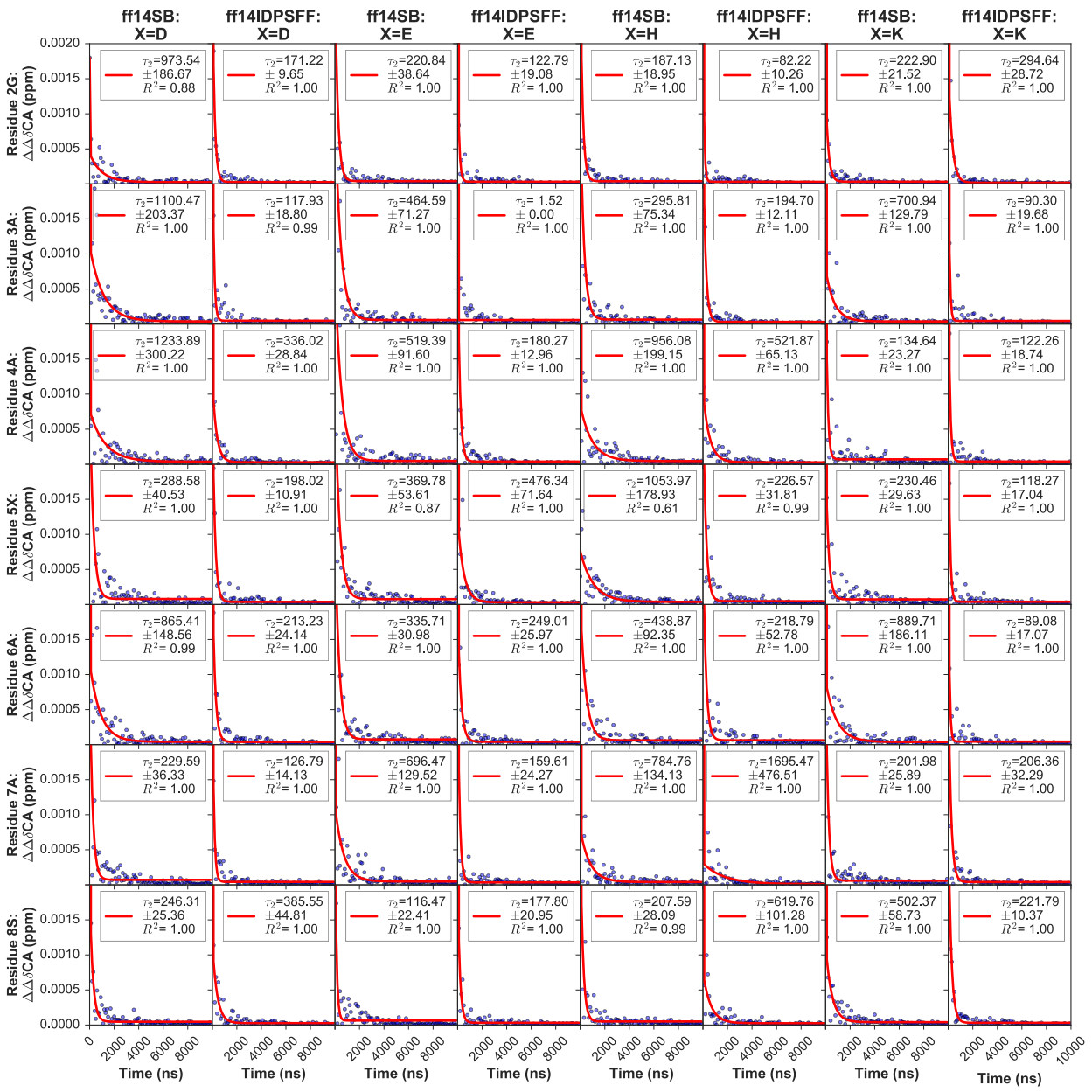


Figure D.6: Biphasic exponential fittings were generated using $\Delta\Delta\delta C\alpha$ from cumulative average data in Figure D.1 for EGAAXAASS (X= D, E, H, K) peptides and force field types. Each average cumulative $\Delta\Delta\delta C\alpha$ (blue dots) 100-ns increment was plotted per residue. Datasets were fitted to the following exponential decay function: $\Delta\Delta\delta C\alpha = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$ (red line). Each column represents a peptide and force field, and each row represents a single residue. Only residues 2G-8S are fitted.

Figure D.7: Biphasic exponential fittings were generated using $\Delta\Delta\delta C\alpha$ from cumulative average data in Figure D.1 for EGAAXAASS (X= L, P, Q, W, Y) peptides and force field types. Each average cumulative $\Delta\Delta\delta C\alpha$ (blue dots) 100-ns increment was plotted per residue. Datasets were fitted to the following exponential decay function: $\Delta\Delta\delta C\alpha = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$ (red line). Each column represents a peptide and force field, and each row represents a single residue. Only residues 2G-8S are fitted.

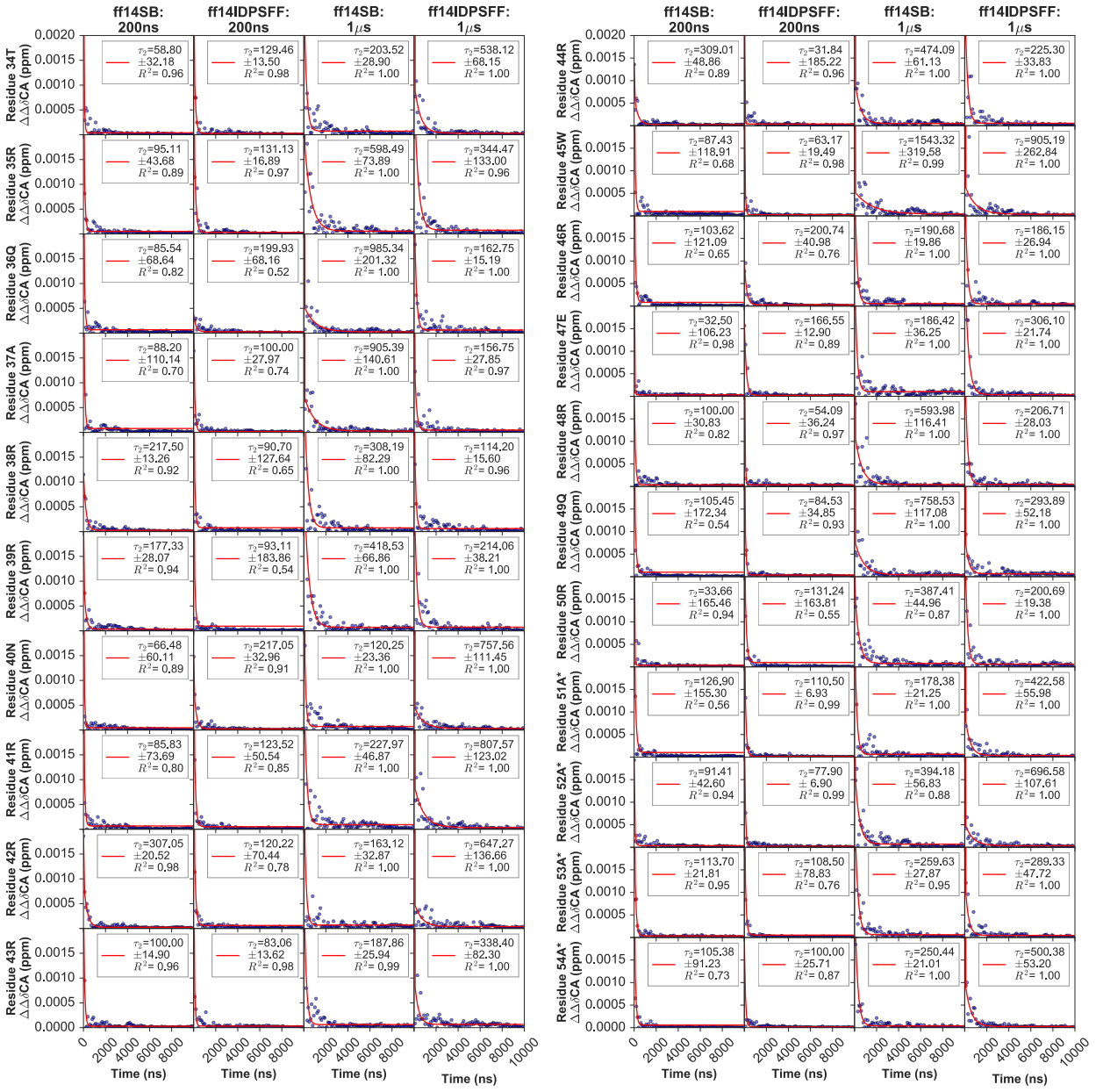


Figure D.8: To evaluate cumulative average convergence of apo Rev simulations from Figure D.3, a scatter plot of $\Delta\Delta\delta$Cα values (blue dots) and corresponding biphasic exponential fit were generated for each simulation (long, short) and force field (ff14SB, ff14IDPSFF) types. Datasets were fitted to the following exponential decay function: $\Delta\Delta\delta\text{C}\alpha = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$ (red line). The above subplot columns are titled according to simulation and force field type and rows labeled according to residue, with non-native residues marked with an asterisk (*) on the y-axis.

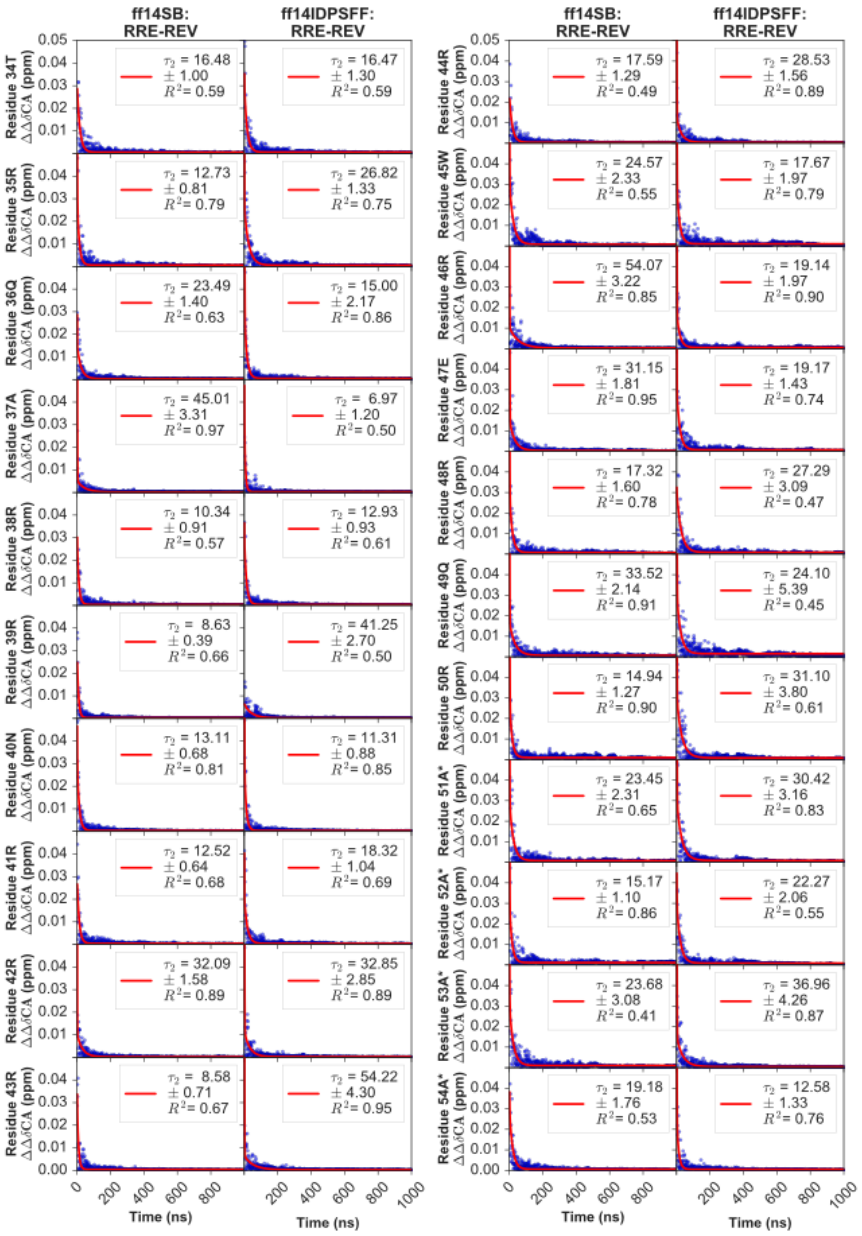


Figure D.9: Biphasic exponential fittings were generated using $\Delta\Delta\delta C\alpha$ from cumulative average data in Figure D.5 for RRE-Rev complexes and force field types. We applied the same fitting to the following exponential decay function: $\Delta\Delta\delta C\alpha = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$ (red line). Each average cumulative $\Delta\Delta\delta C\alpha$ (blue dots) 1-ns increment was plotted per residue. Each column represents a peptide and force field, each row is labeled to its corresponding residue, and non-native residues marked with an asterisk (*).

D.3 Biphasic Exponential Fitting of $\Delta^3 J_{HNH\alpha}$ Datasets

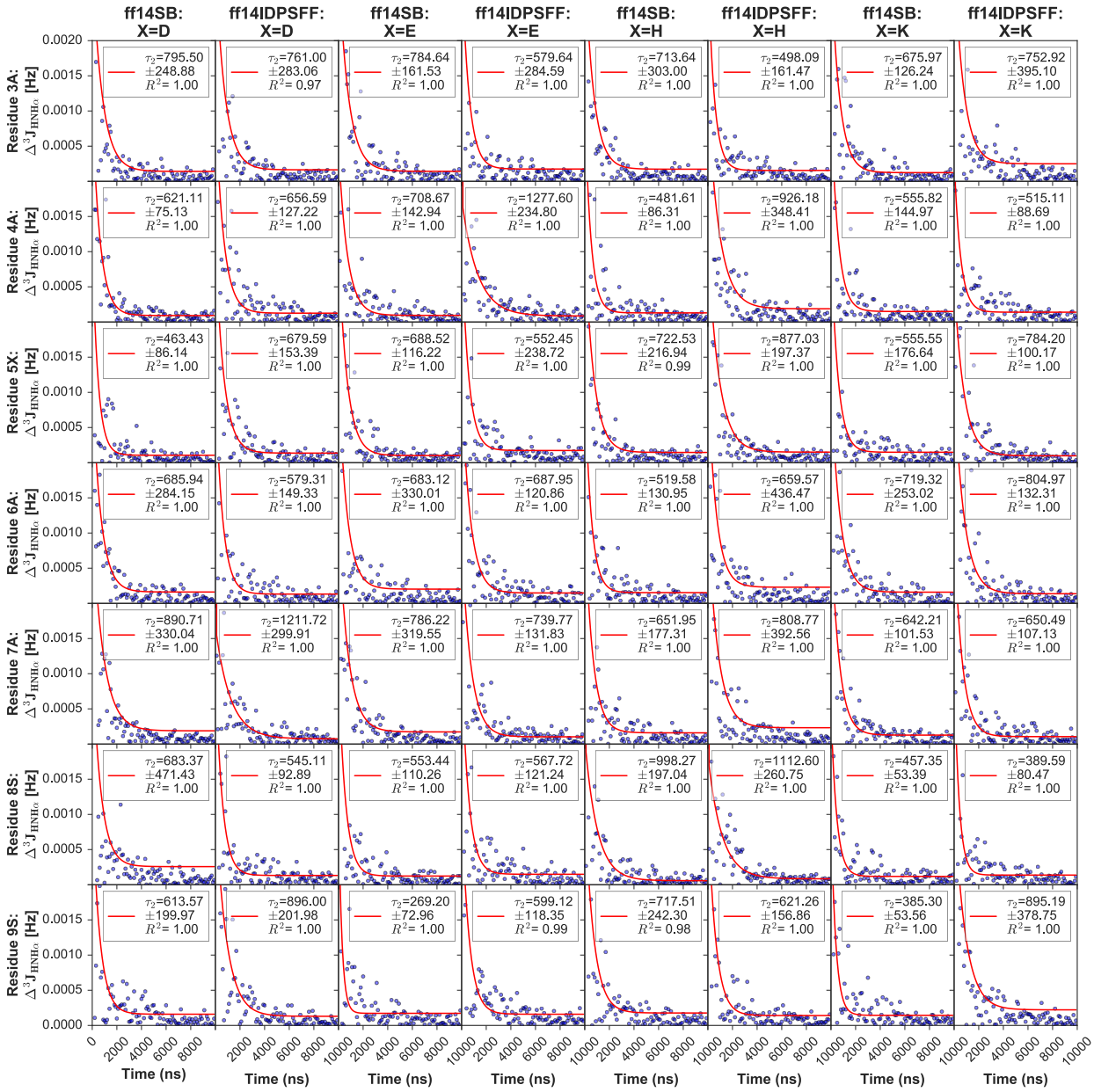


Figure D.10: Biphasic exponential fittings were generated using $\Delta^3 J_{HNH\alpha}$ from cumulative average data in Figure D.2 for EGAAXAASS (X= D, E, H, K) peptides and force field types. Each average cumulative $\Delta^3 J_{HNH\alpha}$ (blue dots) 100-ns increment was plotted per residue. Datasets were fitted to the following exponential decay function: $\Delta^3 J_{HNH\alpha} = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$ (red line). Each column represents a peptide and force field and each row represents individual residues. Only residues 3A-9S are fitted.

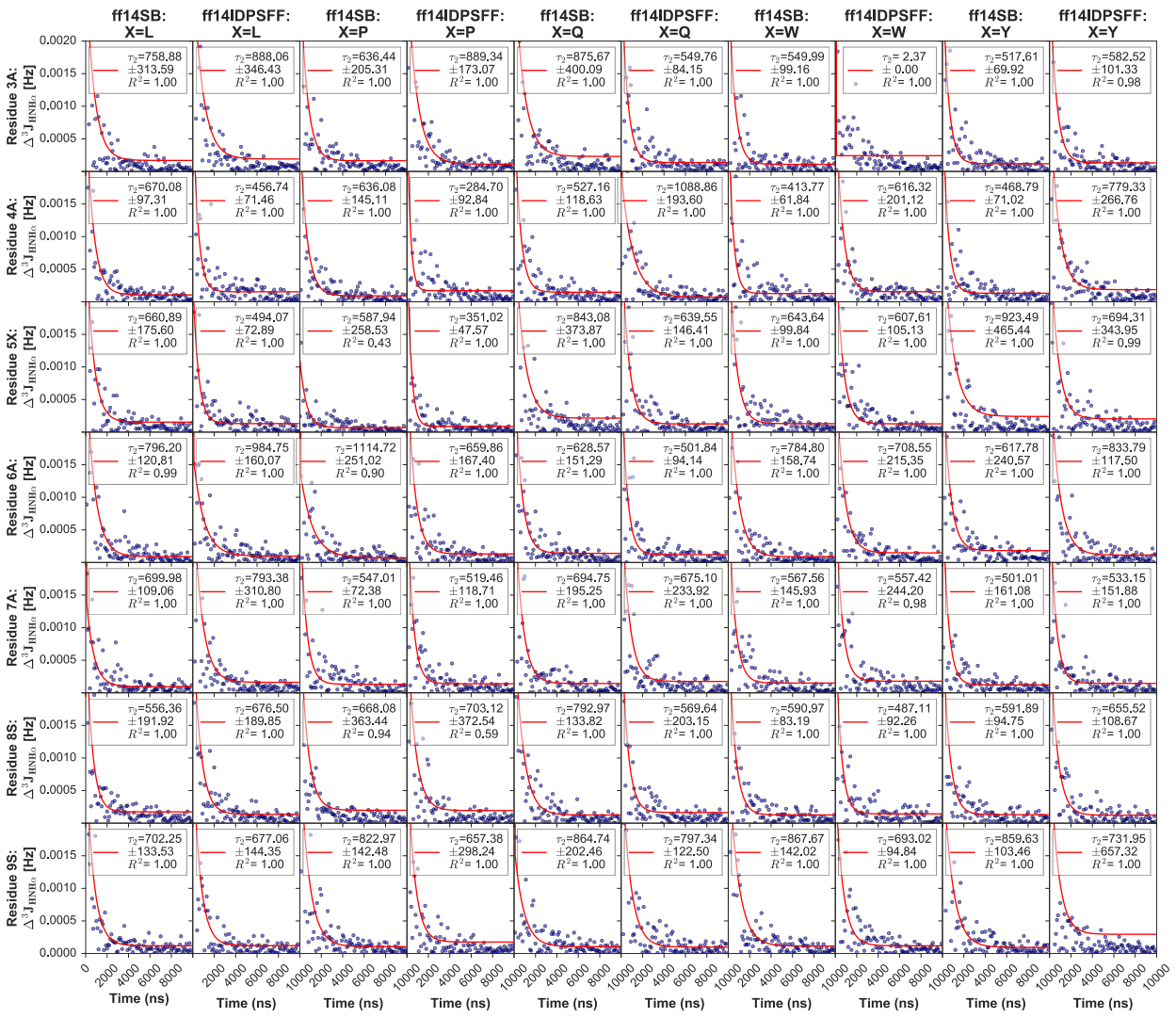


Figure D.11: Biphasic exponential fittings were generated using $\Delta^3J_{HNH\alpha}$ from cumulative average data in Figure D.2 for EGAAXAASS (X= L, P, Q, W, Y) peptides and force field types. Each average cumulative $\Delta^3J_{HNH\alpha}$ (blue dots) 100-ns increment was plotted per residue. Datasets were fitted to the following exponential decay function: $\Delta^3J_{HNH\alpha} = A_1e^{\frac{-x}{\tau_1}} + A_2e^{\frac{-x}{\tau_2}} + c$ (red line). Each column represents a peptide and force field, and each row represents individual residues. Only residues 3A-9S are fitted.

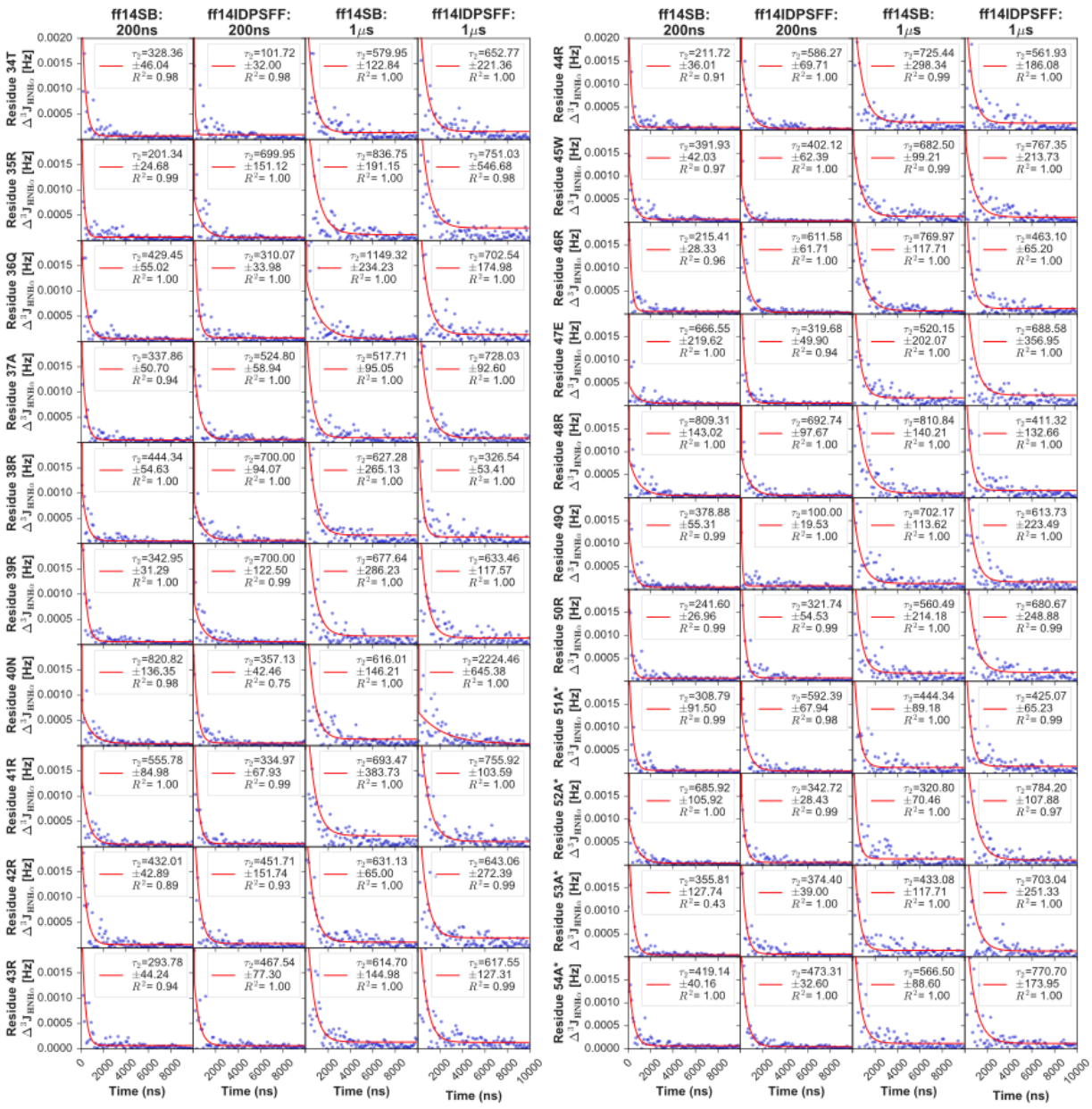


Figure D.12: To evaluate cumulative average convergence of apo Rev simulations from Figure D.4, a scatter plot of $\Delta^3 J_{HNH\alpha}$ values (blue dots) and corresponding biphasic exponential fit were generated for each simulation (long, short) and force field (ff14SB, ff14IDPSFF) types. Datasets were fitted to the following exponential decay function $\Delta^3 J_{HNH\alpha} = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$. The above subplots are titled according to simulation and force field type. Each column represents a peptide and force field, each row is labeled to its corresponding residue, and non-native residues marked with an asterisk (*).

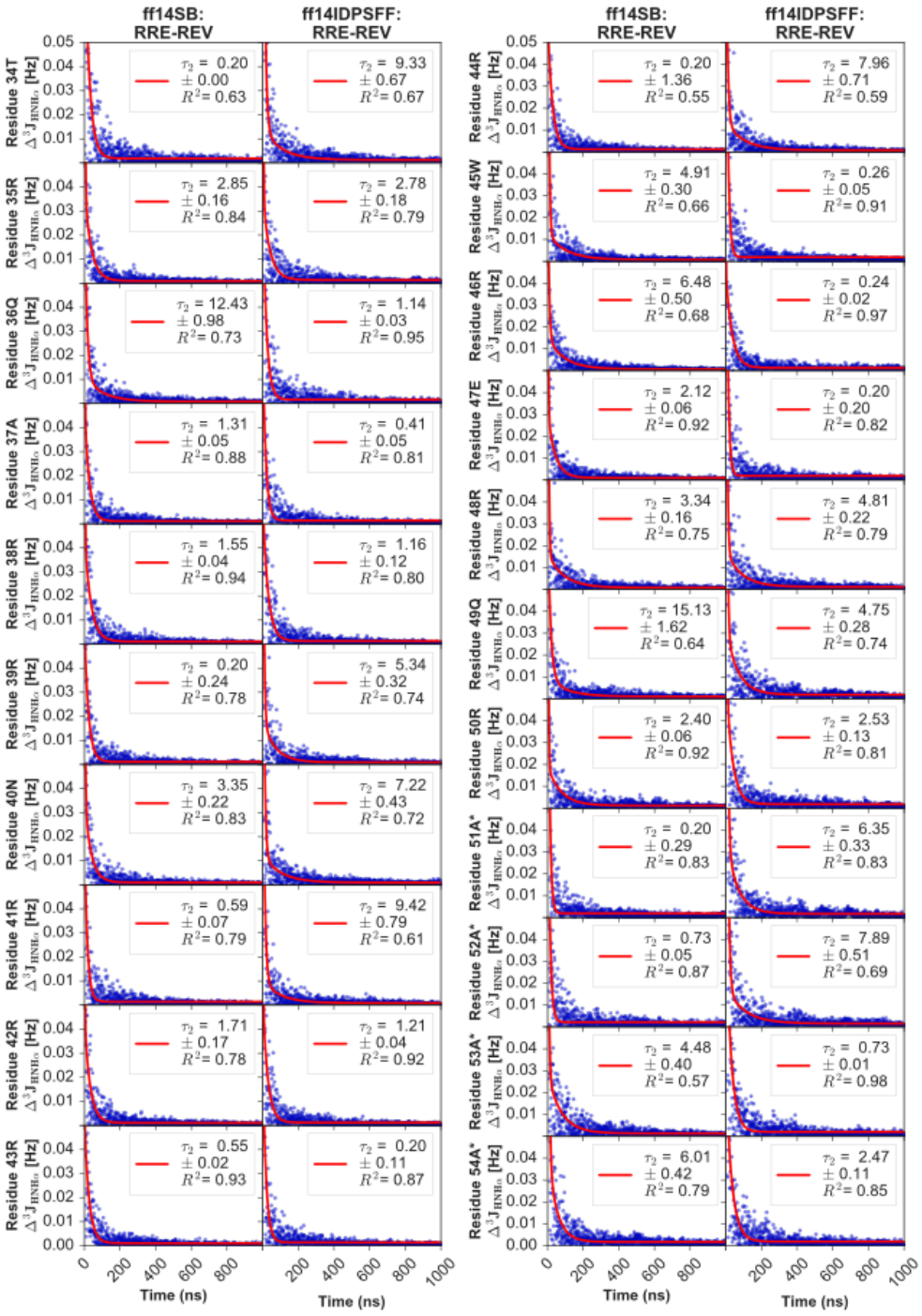


Figure D.13: Biphasic exponential fittings were generated using $\Delta^3 J_{HNH\alpha}$ from cumulative average data in Figure D.5 for RRE-Rev complexes and force field types. We applied the same fitting to the following exponential decay function: $\Delta^3 J_{HNH\alpha} = A_1 e^{\frac{-x}{\tau_1}} + A_2 e^{\frac{-x}{\tau_2}} + c$ (red line). Each average cumulative $\Delta^3 J_{HNH\alpha}$ (blue dots) 1-ns increment was plotted per residue. Each column represents a peptide and force field, each row is labeled to its corresponding residue, and non-native residues marked with an asterisk (*).

D.4 Clustering (apo Rev)

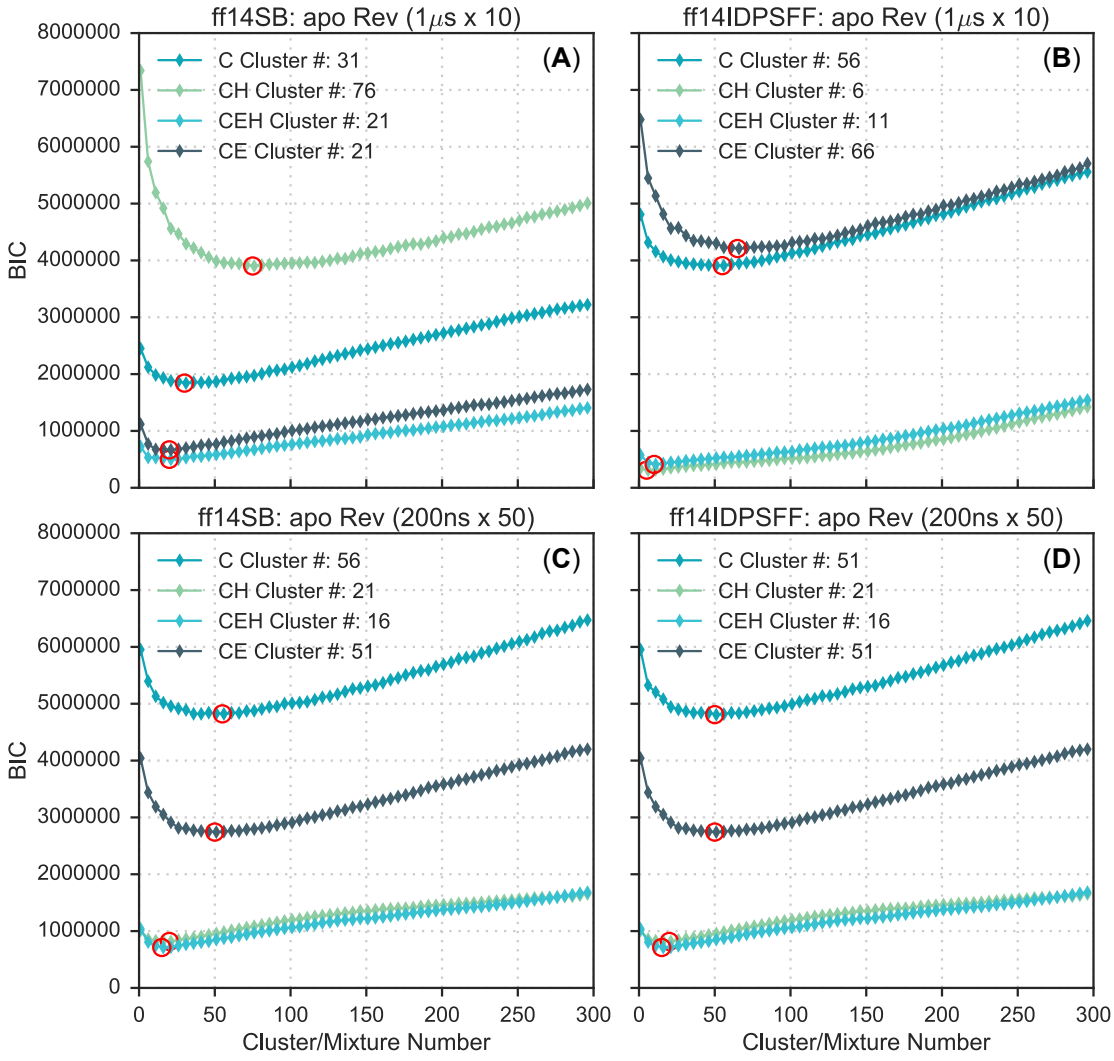


Figure D.14: Determination of appropriate cluster/mixture number using the Bayesian information criterion (BIC) for apo Rev simulations. We calculated the BIC score between 1 to 300 mixtures, and the mixture/cluster number with the lowest BIC was selected for GMM generation. Chosen cluster numbers are indicated in the legend according to secondary structure categories from DSSP pre-clustering. (A) BIC plot of ten 1μs simulations using the ff14SB force field. (B) BIC of ten 1μs simulations using the ff14IDPSFF force field. (C) BIC plot of fifty 200ns simulations using the ff14SB force field. (D) BIC plot of fifty 200ns simulations using the ff14IDPSFF force field.

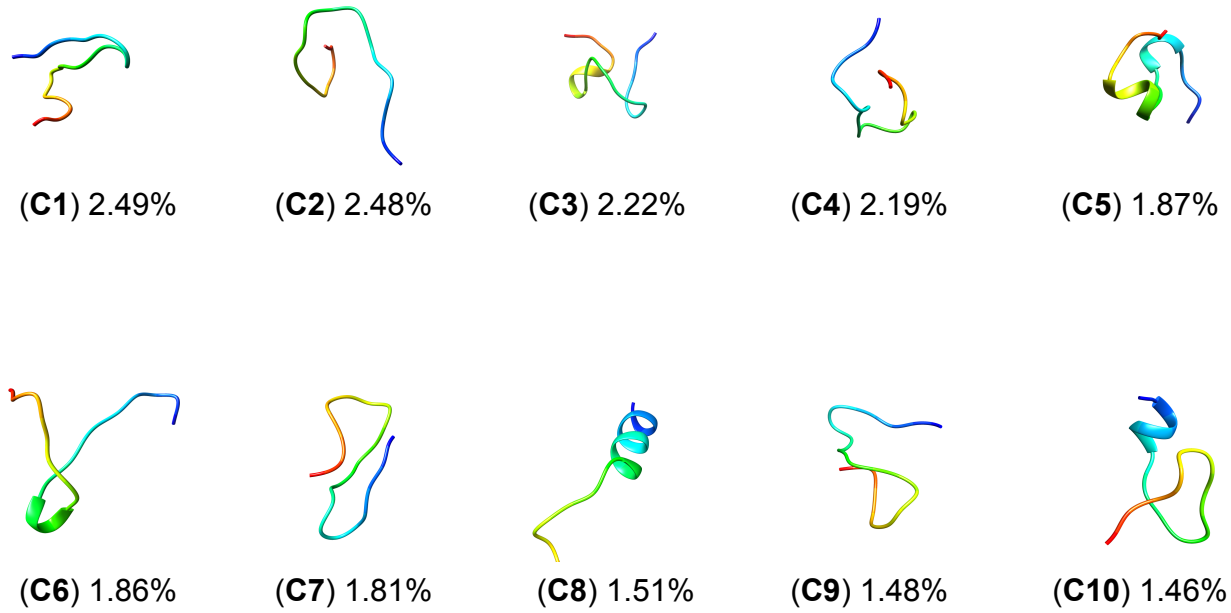


Figure D.15: Top 10 clusters of ff14SB-parameterized simulations (200ns x 50) encompass 19.36% of all frames. Clusters are labeled C1-C10 and colored according to N- to C-termini sequence (red to blue).

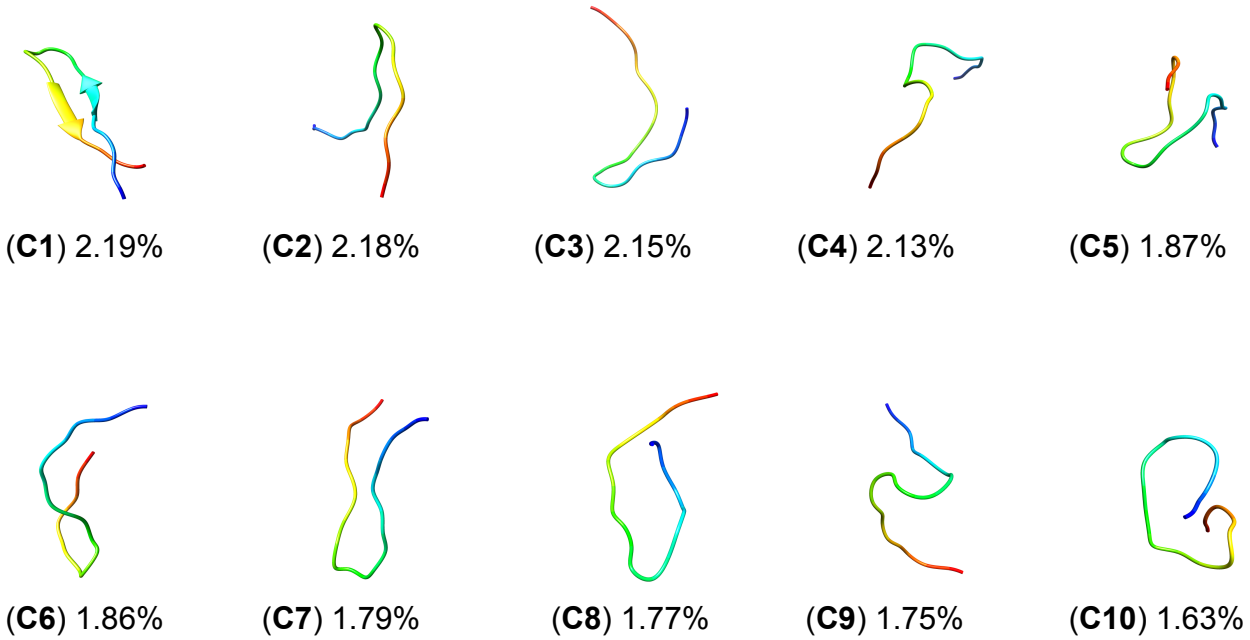


Figure D.16: Top 10 clusters of ff14IDPSFF-parameterized simulations (200ns x 50) encompass 19.32% of all frames. Clusters are labeled C1-C10 and colored according to N- to C-termini sequence (red to blue).

D.5 DSSP

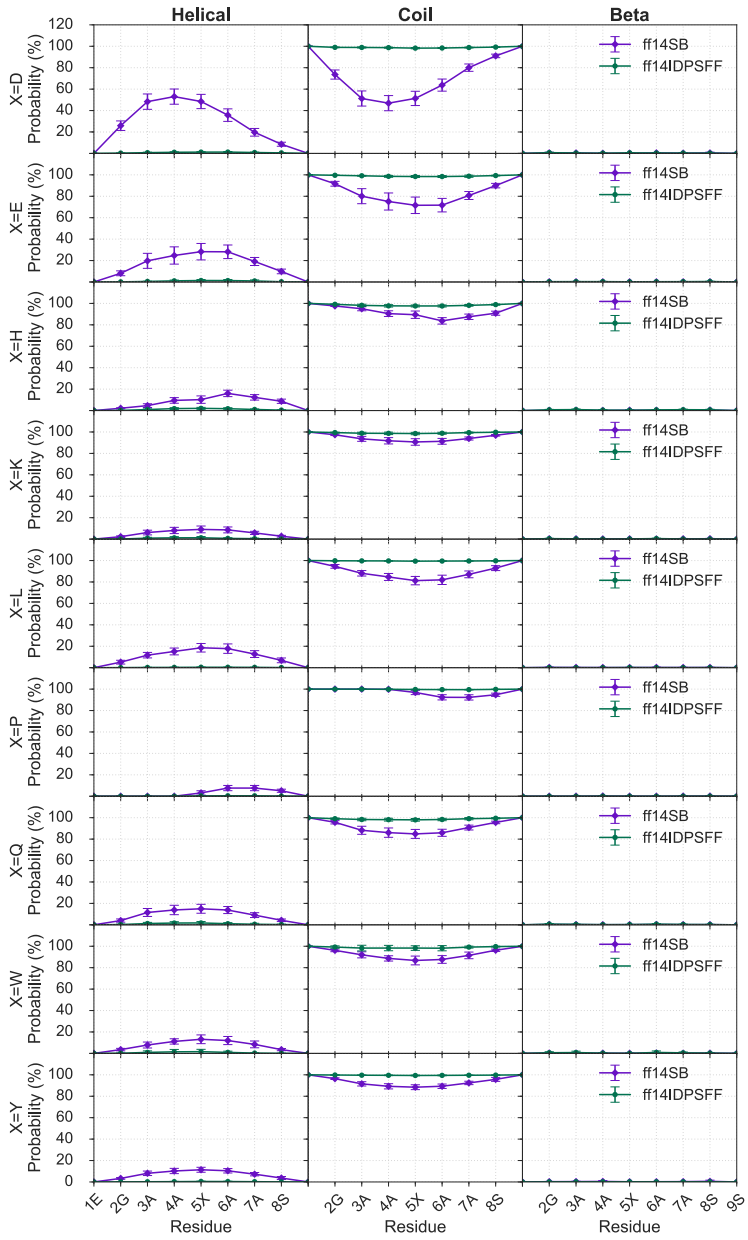


Figure D.17: The average secondary structure propensity of each disordered short peptide. Colors correspond to force fields: purple – ff14SB, green – ff14IDPSFF. All values were calculated using the DSSP1 program and MDtraj[187]. Rows indicate peptide (X = D, E, H, K, L, P, Q, W, Y) and columns indicate one of the three generalized secondary structures (helical, coiled, beta).

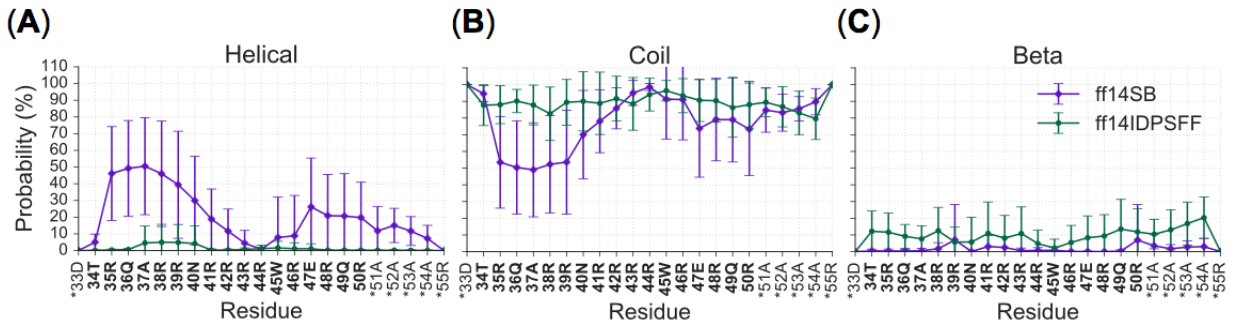


Figure D.18: The average secondary structure propensity of each apo Rev residue was quantified from long simulation (1μs x 10) datasets. Colors correspond to force fields: purple – ff14SB, green – ff14IDPSFF. All values were calculated using the DSSP[131] program and MDtraj.2 (A) The probability of a residue exhibiting helical content. (B) Probability of coil content per residue. (C) Displays the beta-sheet helical propensity per residue. Non-native residues are indicated with an asterisk (*).

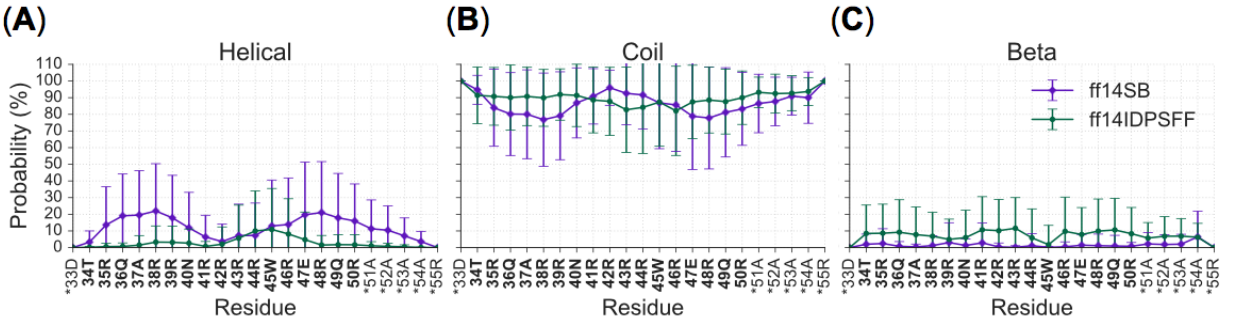


Figure D.19: The average secondary structure propensity of each apo Rev residue was quantified from short simulation (200ns x 50) datasets. Colors correspond to force fields: purple – ff14SB, green – ff14IDPSFF. All values were calculated using the DSSP[131] program and MDtraj[187]. (A) The probability of a residue exhibiting helical content. (B) Probability of coil content per residue. (C) Displays the beta-sheet helical propensity per residue. Non-native residues are indicated with an asterisk (*).

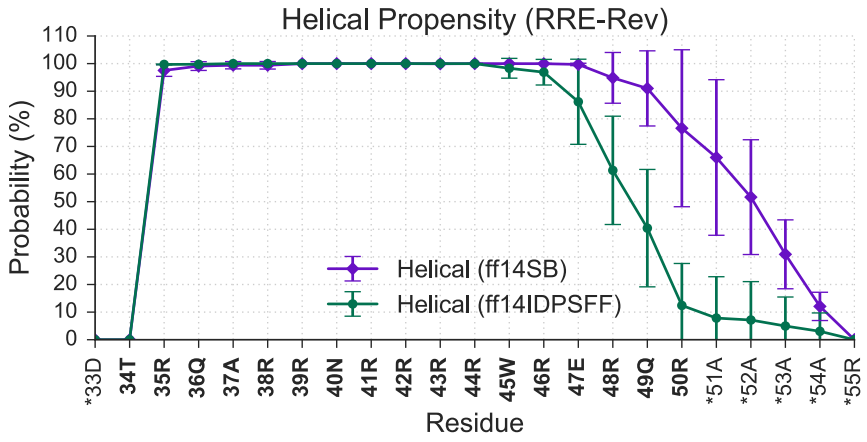


Figure D.20: Average helical propensity of Rev from bound RRE-Rev simulations using the DSSP[131] program. Colors indicate force field: purple – ff14SB, green – ff14IDPSFF.

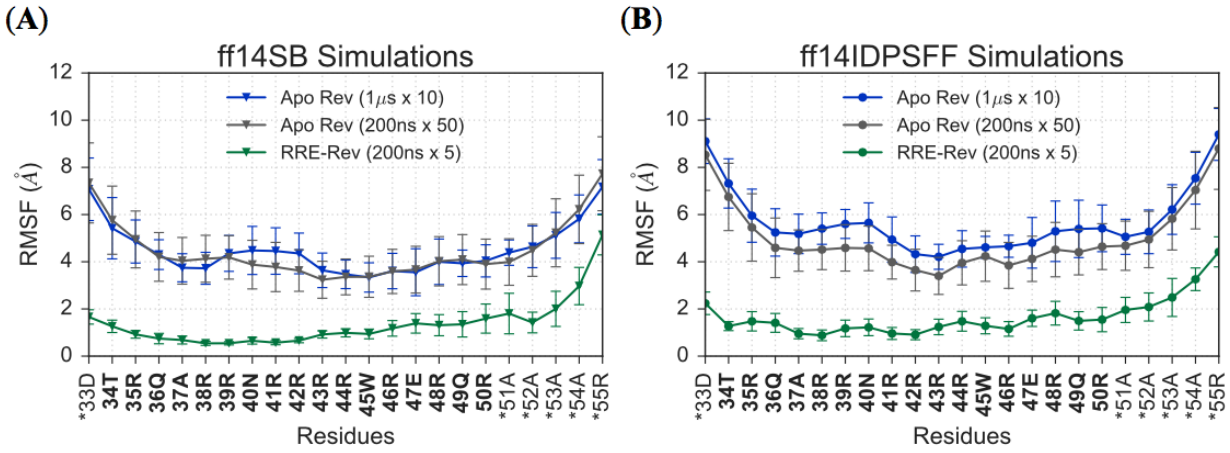


Figure D.21: RMSF analyses of backbone CïĄą atoms Rev-related simulations. (A) Average RMSF of backbone atoms in apo and bound Rev ff14SB-parameterized simulations. (B) Average RMSF of backbone atoms in apo and bound Rev ff14IDPSFF-parameterized simulations. Non-native residues contain an asterisk (*).

Appendix E

Supplement: Neural upscaling from coarse protein structure networks to atomistic structures

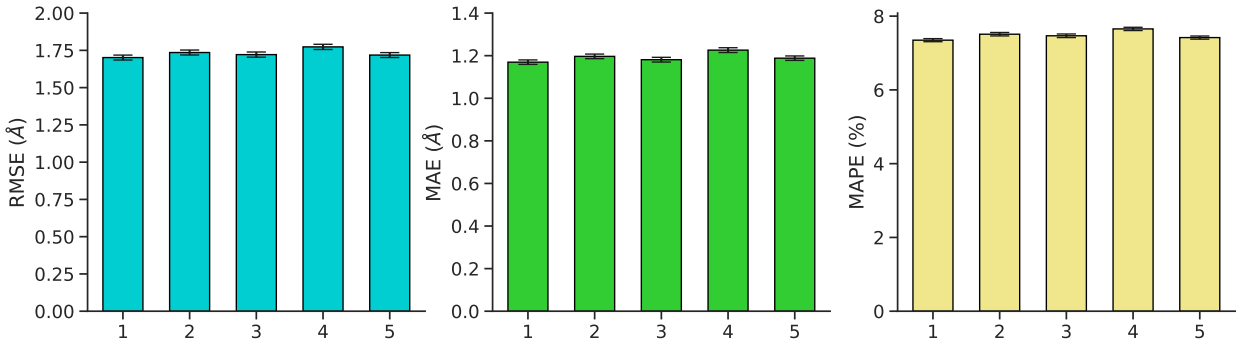


Figure E.1: K-fold cross-validation (K=5) results of the test set for each split. The average root-mean squared error/deviation (RMSD), mean absolute error (MAE), and mean absolute percentage error (MAPE) for each test fold is shown, with 95% confidence intervals represented in the error bars.

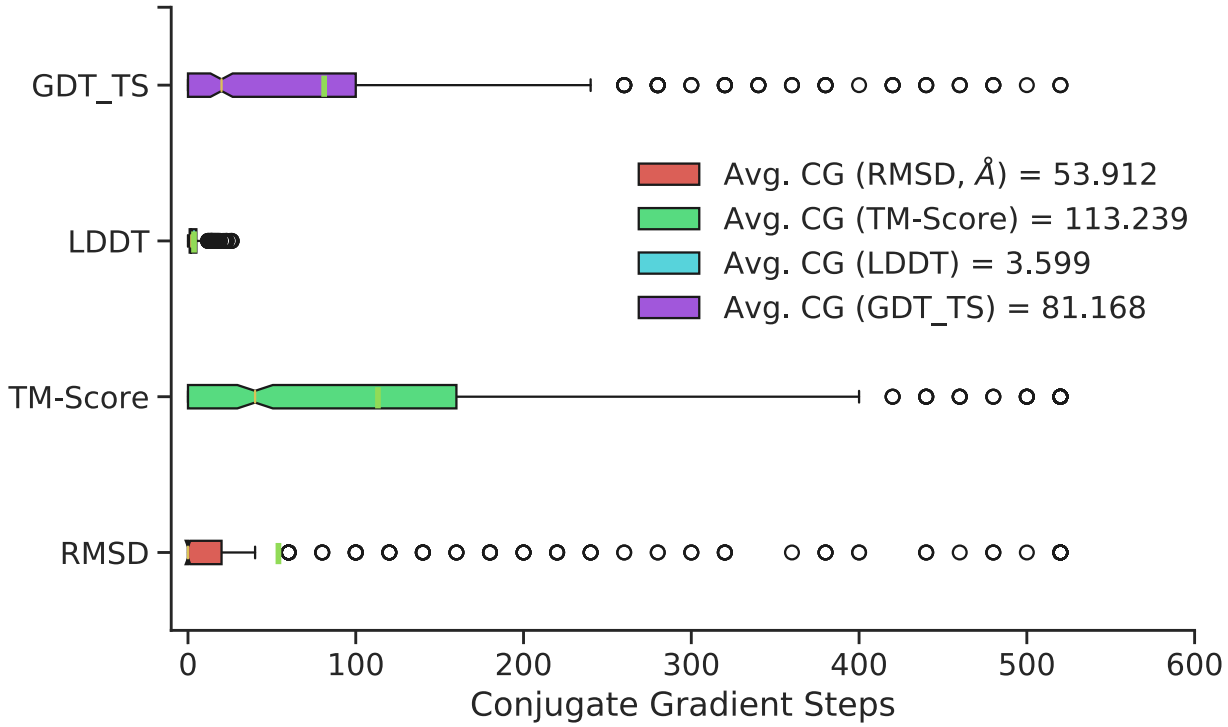


Figure E.2: Boxplot distribution plot detailing at the best conjugate gradient step each protein exhibits for each of the following metrics: RMSD, TM-Score, and LDDT. Averages for each distribution of score types are shown in the legend. Means are represented by a green line, and median represented via a notch.

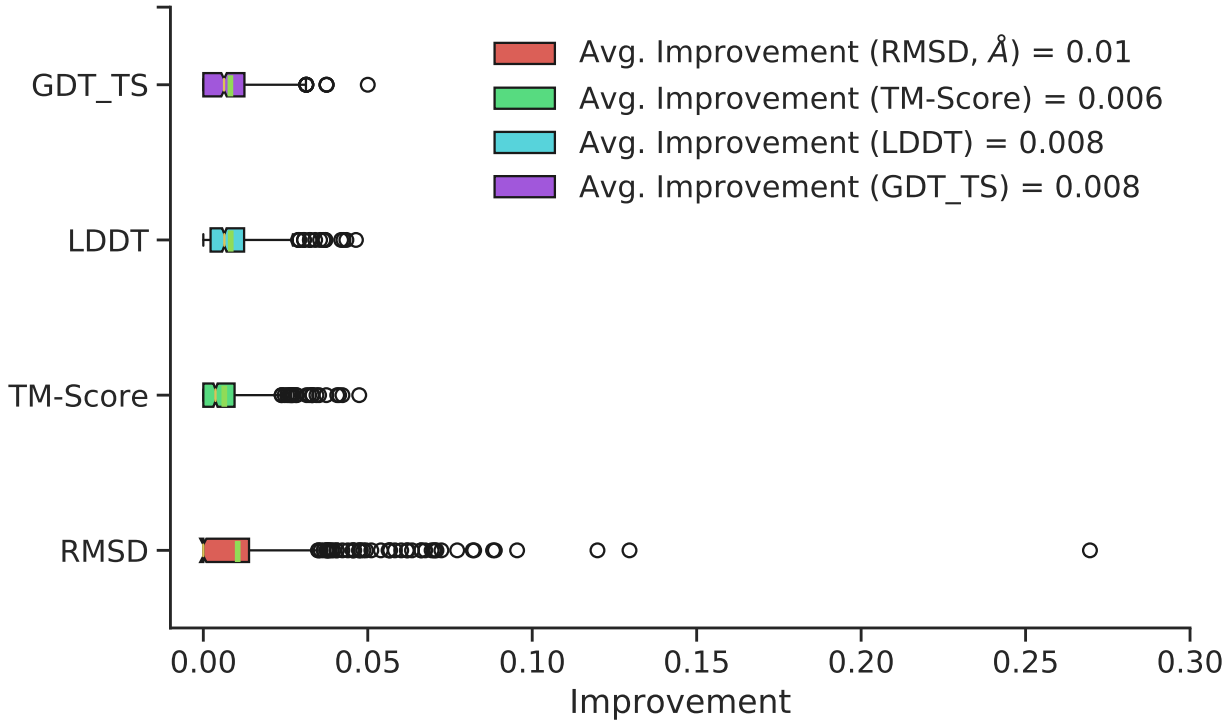


Figure E.3: Boxplot distribution plot detailing “improvement" in metric according to the best conjugate gradient step each protein exhibits, consisting of RMSD, TM-Score, and LDDT. Averages for each distribution of score types are shown in the legend. Means are represented by a green line, and median represented via a notch.