PREDICTING THE EFFECT OF GENETIC VARIANCE ON THE SEQUENCE-ENSEMBLE RELATIONSHIP OF INTRINSICALLY DISORDERED PROTEINS

BY

RUCHI LOHIA

A dissertation submitted to the

Graduate School—Camden

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational And Integrative Biology

Written under the direction of

Dr Grace Brannigan

and approved by

Dr Grace Brannigan

Dr Jinglin Fu

Dr Eric Klein

Dr Cameron Abrams

Camden, New Jersey

October2019

ABSTRACT OF THE DISSERTATION

Predicting the effect of genetic variance on the sequence-ensemble relationship of intrinsically disordered proteins

by RUCHI LOHIA

Dissertation Director:

Dr Grace Brannigan

A hierarchical sequence-based framework for analysis and conceptualization of intrinsically disordered proteins (IDPs) is presented. This framework was further used to develop a novel test for enrichment of higher-order (tertiary) structure in a disordered protein using Molecular Dynamics (MD) simulations and Monte Carlo simulations. Finally, we show that the developed framework can also serve as a useful tool in predicting the consequence of an amino acid substitution on the IDPs function using a bioinformatics approach.

In structured proteins, contacts between residues distant along the sequence are reflected in the tertiary structure, but developing a framework for describing the analogous property in IDPs has not been straightforward. The distribution of hydrophobic residues within the sequence was used to identify 4-15 residues 'blobs' representing local globular regions or linkers. We use this framework within a novel test for enrichment of higher-order (tertiary) structure in disordered proteins; the size and shape of each blob is extracted from MD simulation of the real protein (RP) and used to parameterize a self-avoiding heterogeneous polymer (SAHP). In our study on the 91-residue disordered prodomain of brain derived neurotrophic factor (BDNF), we find that the long 15 residue linker itself creates a segmentation in contact pair map for both SAHP and RP. We find that in RP

only the contact between the segmented region is enriched relative to SAHP. We further quantified the enrichment observed for several other hydrophobic substitutions within the disordered prodomain, including the disease-causing Val66Met substitution. We find that in RPs the enrichment observed in the contact between the segmented region is sensitive to amino acid substitution as well. Only the disease-associated Met66 substitution enriches these contacts significantly, due to a preferred Met-Met interaction. Furthermore, we find several properties of the blobs identified with the sequence-based framework which are enriched in disease-associated SNPs relative to non disease-associated SNPs. This allowed us to present the first systematic, bottom-up, attempt to both identify and annotate subdomains within disordered proteins that are enriched for functional effects.

Acknowledgements

I would like to thank Dr. Grace Brannigan for her advice and mentorship during my journey at Rutgers. Her thoughtful feedback, immense support, guidance, constant encouragement, and appreciation always motivated me to push my boundaries and try harder. Her kindness and patience towards my mistakes always made me humble. She has been a role model and I still have so much to learn from her. Her advice has been invaluable for my research and personal development.

I would like to thank Dr. Matt Hansen for collaborating on the Bioinformatics project (Chapter 3 in the thesis). I am thankful for his guidance and thoughtful feedback during the last few months.

I would like to thank Dr. Reza Salari, for providing his guidance during the first few years in the lab. He taught me all the basics from running my first molecular dynamics simulations to writing cleaner codes.

I would like to thank my thesis committee: Dr. Cameron Abrams, Dr. Jinglin Fu, and Dr. Eric Klein, for being a part of my committee and providing insightful comments and feedback. I would like to thank Dr. Jérôme Hénin and Dr. Tom Joseph for their insightful discussions and suggestions. I would like to thank Dr. Robert Best for his suggestions on my project during Biophysical meetings.

I would like to thank the entire team of Rutgers Discovery Informatics Institute (RDI2) and The Rutgers Office of Advanced Research Computing (OARC) for providing the computational time and support in the most critical time which made the data collection possible.

I would like to thank Dr. Sunil Shende and Daniel Russo for their suggestions

on the web tool development. I would like to thank undergraduate student Kaitlin Bassi for assisting me with the web tool development and for allowing me to be a mentor myself.

I would like to thank my fellow graduate students Sruthi, Liam, Shashank, Rulong, and Kristen for being wonderful lab-mates and for providing their constant help and feedback on my presentations. I am thankful to Sruthi and Nastassia for being incredible roommates. I am thankful to my friends Arushi, Rupali, Shourya, Varun, Mukta, Suprita, Avani, Priya, Marlene and Monica for always ensuring fun trips in the last few years which kept me sane.

Finally, I would like to thank my family for constantly encouraging me and always believing in me. I am especially thankful to my mother, sister, dad, and brother, this journey would not have been possible without their love and support.

Dedication

Dedicated to my loving mother Annu Lohia.

Table of Contents

List of Tables

List of Figures

Introduction

Proteins are biological macromolecules composed of amino acids which are essential for cell function. In 1958, Kendrew et. al (Kendrew et al. 1958) used x-ray crystallography to solve the first structure of myoglobin, leading to the structural biology era. Since then, the number of protein structures solved is growing exponentially, with more than 100,000 protein structures deposited in the Protein Data Bank (PDB). In parallel, the structure-function paradigm was developed, which states that the function of a protein is encoded in its unique 3D structure, which is determined through its sequence. For a few decades, proteins were considered to be functional only if they had a highly structured state (folded conformation) under physiological conditions.

Until the late 1990s, if a biologically active protein was found to lack structure, it was considered to be an exception and was mostly ignored. In the late 1990s, it was recognized that the structure-lacking functional proteins constituted a new class of biologically active proteins, and they were termed intrinsically disordered proteins (IDPs) (Wright & Dyson 1999). The disorder in the protein sequence could range from a few residues to its entire sequence. The disordered protein segments are referred to as intrinsically disordered regions (IDRs). IDPs/IDRs (collectively referred as IDPs hereafter) are prevalent in organisms from all kingdoms of life and more than 33% of eukaryotic proteins contain long IDRs (more than 30 residues) (Ward et al. 2004). Seventy percent of protein structures deposited in the PDB have portions of their sequence of missing electron density indicating disorder (DeForte & Uversky 2016b). At present, the disordered protein database DISPROT (Sickmeier et al. 2007) has more than 2000 IDR sequences deposited, and approximately 1150 non-redundant validated IDPs have been found (DeForte & Uversky 2016a). Many of the IDPs are involved in critical biological functions, including transcriptional regulation (Minezaki et al. 2006), cell signaling (Dunker

et al. 2005; Wright & Dyson 2015) and control pathways (Fuxreiter et al. 2008). They mediate post-translational modifications (PTMs) (Darling & Uversky 2018) and function as hubs in protein-protein interaction networks (Oldfield et al. 2008).

IDPs exist as ensembles of rapidly interconverting heterogenous conformations and mediate their function through disorder (Dyson & Wright 1998; Mukhopadhyay et al. 2007; Abeln & Frenkel 2008). High conformational flexibility of these proteins gives them several advantages, like increased interaction area per residue ('fly-casting') (Shoemaker et al. 2000), ability to bind multiple partners (Fuxreiter et al. 2008) and ability to bind with high specificity and low affinity (Iakoucheva et al. 2002; Uversky & Dunker 2010). These proteins might act as linkers for two globular domains (Akimoto et al. 2013), tails that modulate the function of the structured domain (Erler et al. 2014; Peng et al. 2012), undergo coupled folding and binding in the presence of a ligand/receptor (Wright & Dyson 2009; Staneva et al. 2012) or remain disordered even in the bound state (Fuzzy complexes) (Tompa & Fuxreiter 2008; Jin et al. 2013). In parallel to the emerging role of IDPs in biological function, studies on the formation of membrane-less organelles and their physical properties are fundamentally changing our views of cellular biology. Together, in vitro and in vivo experiments have demonstrated the important role of IDPs in the formation of these liquid-like organelles (Li et al. 2012; Holehouse & Pappu 2018).

It has been shown that 21.7 % of missense disease mutations are found in disordered regions (Vacic & Iakoucheva 2012). Due to their signaling and regulatory roles, PTMs are generally found in disordered regions and altered or disrupted PTMs often have been linked to diseases (Darling & Uversky 2018). IDPs are potential drug targets (Bhattacharya & Lin 2019).

Cellular conditions, post-translational modifications, binding events, and disease-associated SNPs can affect the relative free energies of individual conformations (Boehr et al. 2009). As a result, the populations of individual conformations

within the ensemble change under different conditions. These individual conformations are often important for function (van der Lee et al. 2014). Therefore, to understand the functional mechanisms of IDPs, it is essential to understand their sequence-ensemble relationship. This is particularly challenging due to the highly dynamic and heterogeneous conformational ensembles IDPs occupy.

NMR is the widely used experimental tool for characterizing the solution structure and dynamics of IDPs (Mittag & Forman-Kay 2007; Habchi et al. 2014). However, IDPs typically convert between conformations faster than nanoseconds-millisecond time scale of NMR experiment, leading to an averaging of the NMR observables across structural subpopulations. The uniform average hinders the structural characterization of all the conformations in its ensemble (Ball et al. 2014). Molecular Dynamics (MD) simulations therefore can play an important role in understanding their dynamics (Stanley et al. 2015; Ithuralde et al. 2016; Knott et al. 2012; Invernizzi et al. 2013; Yedvabny et al. 2015; Levine & Shea 2017). It has the potential to provide detailed structure and dynamics information at the atomic level. Extensive MD simulations, especially replica-exchange at both all-atom and coarse grained levels have been performed to probe the effects on IDPs (Patel et al. 2014a,b; Ojeda-May & Pu 2013; Kurcinski et al. 2014; Kovalskyy & Ivanov 2014). MD ensembles are validated through direct back calculation of NMR observables.

The lack of structure in IDPs is encoded in its sequence, and sequence properties such as mean hydrophobicity (H), the net charge per residue (NCPR) and the fraction of charged residues (FCR) have been useful in predicting disorder and phase behavior from sequence. Uversky et al. (Uversky et al. 2000) found that proteins with disorder are generally characterized by low H and high NCPR (Fig 1a,b). This result was explained by a simple model: repulsion between like charges favors unfolding while increased hydrophobicity favors folding. They showed that the ratio of just these two sequence parameters can segregate ordered

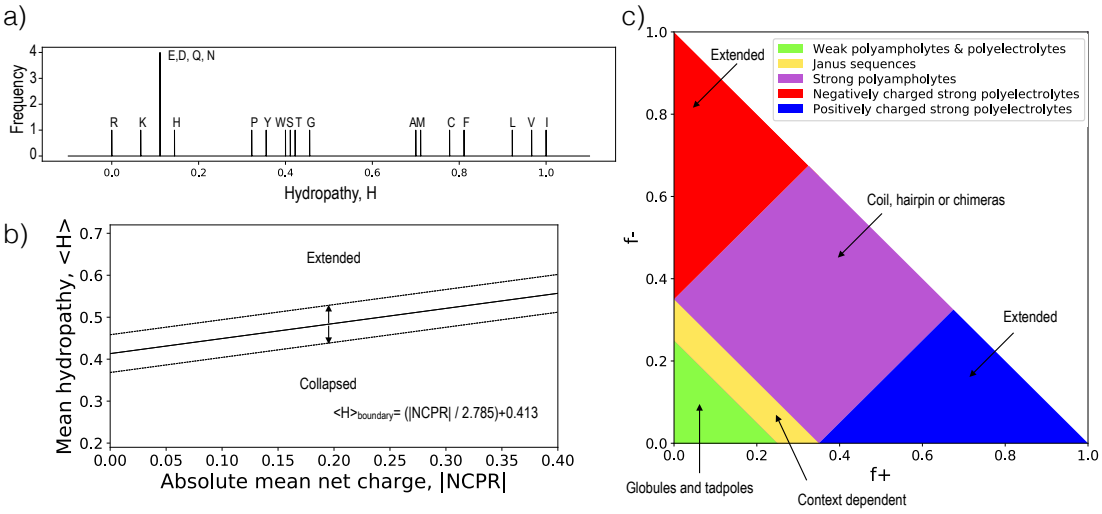

Figure 1: **Schematic of theories predicting sequence ensemble relationship of IDPs.** a) The frequency of hydrophobicity (H) values for all twenty amino acids on the Kyte-Dolittle (Kyte & Doolittle 1982) hydropathy scale. The H values were scaled to fit between 0 and 1. b) Uversky diagram (Uversky et al. 2000) of IDPs and globular proteins, as a function of absolute net charge per residue ($|NCPR|$) and $\langle H \rangle$, with the boundary line between folded and disordered proteins given by the equation in the legend. The arrows point to the lines delimiting the zone with a prediction accuracy of 95% for disordered proteins and 97% of ordered proteins, at the expense of discarding 50% of all proteins (Oldfield et al. 2005). c) The diagram of IDP states proposed by Das and Pappu (Das & Pappu 2013), based on the fraction of positive (f^+) and negative (f^-) charged residues.

and disordered proteins. The proteins on either side of the boundary will be collapsed or expanded. However, the predicted method works well if the proteins lie far away from the boundary. A large number of proteins lie within a small margin (about 50%) at the boundary of Uversky plot and therefore aren't distinguishable when projected onto the Uversky plot (Oldfield et al. 2005).

Ten years later, Das et al (Das & Pappu 2013) found that along with NCPR,

the fraction of positive charges (f+), the fraction of negative charges (f-) and segregation of opposite charges (reflected in the metric κ) can be useful in predicting the sequence-phase behavior relationship of IDPs (Fig 1c). They showed that since IDP sequences include both types of charges, and at least 75% of known IDPs are polyampholytes, they typically possess the expected behavior of flexible and charged polymers (Mao et al. 2010; Das & Pappu 2013). The FCR discriminates between weak and strong polyampholytes. Weak polyampholytes show conformational preference for compact globules. Conformational preferences for strong polyampholytes (FCR >0.35) further depend on the distribution of oppositely-charged residues within the linear sequence. For well-mixed charged distributions, electrostatic repulsions and attractions are counterbalanced, leading to conformational preferences that resemble self-avoiding random walks or generic Flory random coils. Segregation of oppositely-charged residues form hairpin-like conformations caused by long-range electrostatic attractions. However, in the Das and Pappu phase diagram (Das & Pappu 2013), 40% of IDPs lie at the boundary region, and thus the phase diagram does not give us any insights regarding their expected phase behavior (Das et al. 2015).

Understanding the effect of an amino acid substitution (disease-associated or otherwise) has been a classical way of understanding a protein's sequence function relationship. A similar approach has been used for IDPs as well. Numerous structural and simulation studies (Larini et al. 2013; Ganguly & Chen 2015; Viet et al. 2014, 2013; Truong et al. 2014; Zhan et al. 2013; Xu et al. 2013) have demonstrated clear effects of single charged-residue insertion, deletion, or substitutions on conformational ensemble and aggregation of IDPs monomers. It has been frequently observed that post-translational regulations such Ser/Thr phosphorylation can change the FCR and NCPR properties of IDPs and can thus tune the sequence-phase behavior relationships of IDPs due to their polyampholytic nature (Firman & Ghosh 2018; Das & Pappu 2013).

However, the current theories work well for predicting the sequence-phase behavior relationship of strong polyampholytes, shorter sequences, and mutations that change charge. For a strong polyampholyte, charge permutation can modulate charge segregation parameter κ, but for weak polyampholytes charge permutations do not affect the FCR, and thus no change in sequence conformation is predicted. For IDPs longer than 30 residues, a single residue insertion or deletion often does not change its predicted location in the phase diagram, because such a small fraction of the total number of residues are affected. Moreover, the Das and Pappu diagram (Das & Pappu 2013) is completely based on the charged residues within the sequence and does not predict the effect of a charge neutral mutation.

Although IDP's do not have a unique 'tertiary structure' and structure-based coupling between distant residues in IDPs is expected to be weak, conformations of IDP's are governed by several non-specific and weak long range interactions. In structured proteins, contacts between residues distant along the sequence are reflected in the tertiary structure, but developing a framework for describing the analogous property in IDPs has not been straightforward. As the length of IDPs increases, the number of possible residue contact pairs increases exponentially while the frequency of contact formation between these residue pairs decreases. A large number of weak contacts makes it statistically challenging to analyze all the possible contact pairs in IDPs.

In my dissertation, I have aimed to address the following questions in order to understand the role of subtle sequence differences in differentiating the conformational space and function of homologous IDPs: 1) How can we meaningfully detect tertiary interactions in a long disordered protein? 2) Can we predict the effect of amino-acid substitutions, including charge-neutral substitutions on the tertiary interactions in a long disordered protein? I focused my investigation of these questions on a long disordered protein (the prodomain of brain-derived neurotrophic factor, or BDNF) which contains a well-studied disease-associated

charge-neutral mutation. Although the BDNF prodomain has not been previously studied computationally, it proved to be a useful model system. From my simulations of the BDNF prodomain, I developed new hypotheses that were also supported by a general survey of disease-associated SNPs in disordered proteins.

During my Ph.D. work, we developed a novel hierarchical sequence-based framework for analysis and conceptualization, which further helped in characterizing tertiary interactions in a long disordered IDP. The current theories have identified NCPR, H, opposite charges segregation in determining the sequence-ensemble relationship of IDPs. In our current approach, we identified a new sequence property, the distribution of hydrophobic residues within the sequence, to identify 5-15 residues 'blobs' representing local globular regions or linkers. The properties of these blobs were further identified on the Uversky diagram and Das and Pappu diagram. We use this framework within a novel test for enrichment of higher-order (tertiary) structure in disordered proteins; the size and shape of each blob are extracted from MD simulation of the real protein (RP), and used to parameterize a self-avoiding heterogeneous polymer (SAHP). In our study, we find that the long linker blobs create a segmentation in the contact-pair map for both SAHP and RP. In RP the contact between the segmented region is enriched and is sensitive to amino acid substitution. Furthermore, we find several properties of the blobs which are enriched in disease-associated SNPs relative to non disease-associated SNPs, further helping us to develop a scalable relationship between the consequence of a single amino acid substitution and its location within the disordered protein.

In Chapter 1, we develop the hierarchical sequence-based framework and test for enrichment of higher-order (tertiary) structure to study the effect of hydrophobic disease-associated SNP Val66Met on a 91 residue disordered domain of brain derived neurotrophic factor (BDNF). We find that M66 form of the BDNF

prodomain increases tertiary interactions within the protein due to preferred Met-Met interactions. In Chapter 2, we further apply the tertiary test for detecting the effect of several other charge neutral mutations at residue 66 and the effect of adding charge at residue 65 within the BDNF prodomain. In Chapter 3, we explore the ability of our developed sequence-based framework to predict a scalable relationship between the consequence of a single amino acid substitution and its location within the disordered protein. Finally, realizing the significance of the developed sequence-based framework for IDPs, in general, we developed a web tool for interactive identification of sequence topology for any given protein sequence.

Chapter 1

Sequence specificity despite intrinsic disorder: how a disease-associated Val/Met polymorphism rearranges tertiary interactions in a long disordered protein

1.1 Abstract

The role of electrostatic interactions and mutations that change charge states in intrinsically disordered proteins (IDPs) is well-established, but many disease-associated mutations in IDPs are charge-neutral. The Val66Met single nucleotide polymorphism (SNP) in precursor brain-derived neurotrophic factor (BDNF) is one of the earliest SNPs to be associated with neuropsychiatric disorders, and the underlying molecular mechanism is unknown. Here we report on over 250 μs of fully-atomistic, explicit solvent, temperature replica-exchange molecular dynamics (MD) simulations of the 91 residue BDNF prodomain, for both the V66 and M66 sequence. The simulations were able to correctly reproduce the location of both local and non-local secondary structure changes due to the Val66Met mutation when compared with NMR spectroscopy. We find that the change in local structure is mediated via entropic and sequence specific effects. We developed a hierarchical sequence-based framework for analysis and conceptualization, which first identifies "blobs" of 4-15 residues representing local globular regions or linkers. We use this framework within a novel test for enrichment of higher-order (tertiary) structure in disordered proteins; the size and shape of each blob is extracted from MD simulation of the real protein (RP), and used to parameterize a self-avoiding heterogenous polymer (SAHP). The SAHP version of the

BDNF prodomain suggested a protein segmented into three regions, with a central long, highly disordered polyampholyte linker separating two globular regions. This effective segmentation was also observed in full simulations of the RP, but the Val66Met substitution significantly increased interactions across the linker, as well as the number of participating residues. The Val66Met substitution replaces β-bridging between Val66 and Val94 (on either side of the linker) with specific side-chain interactions between Met66 and Met95. The protein backbone in the vicinity of Met95 is then free to form β-bridges with residues 31-41 near the N-terminus, which condenses the protein. A significant role for Met/Met interactions is consistent with previously-observed non-local effects of the Val66Met SNP, as well as established interactions between the Met66 sequence and a Met-rich receptor that initiates neuronal growth cone retraction.

Author summary

Intrinsically disordered proteins are proteins that have no well-defined structure in at least one functional form. Mutations in one amino acid may still affect their function significantly, especially in subtle ways with cumulative adverse effects on health. Here we report on molecular dynamics simulations of a protein that is critical for neuronal health throughout adulthood (Brain-derived Neurotrophic Factor). We investigate the effects of a mutation carried by 30% of human population, which has been widely studied for its association with aging-related and stress-related disorders, reduced volume of the hippocampus, and variations in episodic memory. We identify a molecular mechanism in which the mutation may change the global conformations of the protein and its ability to bind to receptors.

1.2 Introduction

The physiological significance of intrinsically disordered proteins (IDPs), which can explore a wide range of conformational ensembles in their functional form, is now well-established (Uversky 2013; Panchenko & Babu 2015; Ward et al. 2004; Dyson & Wright 2005; Uversky 2019). More than 33% of eukaryotic proteins contain disordered regions longer than 30 residues (Ward et al. 2004), many of which are involved in critical biological functions, including transcriptional regulation (Minezaki et al. 2006) and cell signaling (Dunker et al. 2005; Wright & Dyson 2015; Vucetic et al. 2007). Long intrinsically disordered regions are particularly abundant among cancer (Iakoucheva et al. 2002) and neurodegenerative-associated proteins (Habchi et al. 2014; Buée et al. 2000).

IDP amino acid sequences tend to be low-complexity (Weathers et al. 2006; Romero et al. 2001) and include numerous charged residues, often in long repeats (Uversky 2013; Jorda et al. 2010). In contrast to ordered proteins, in which a complex sequence encodes a well-defined tertiary structure, an IDP sequence determines a heterogeneous conformational ensemble (Dyson & Wright 1998; Mukhopadhyay et al. 2007; Abeln & Frenkel 2008). More than 35% of IDPs reported in DISPROT (Sickmeier et al. 2007) are strong polyampholytes, and their ensemble properties can be predicted using statistical theories of polyampholytes from polymer physics and global properties of the sequence, including the fraction of charged residues and the separation of oppositely charged residues (Fig 1c) (Das et al. 2015; Das & Pappu 2013; Sawle & Ghosh 2015; Firman & Ghosh 2018). This role is consistent with the long-range nature of electrostatic interactions, which can affect coupling between distant residues in an otherwise disordered structure.

Although IDP sequences are low-complexity and do not encode a well-defined structure, single residue substitutions can still have functional effects that are

significant for the organism (Uversky et al. 2008). More than 25% of disease-associated missense single nucelotide polymorphisms (SNPs) are found in IDPs (Vacic et al. 2012). Although detectable, the relatively subtle functional effects of these SNPs may lead to relatively weak selection pressure, whether positive or negative, allowing the mutation to persist at high frequencies within a population. Numerous structural and simulation studies (Larini et al. 2013; Ganguly & Chen 2015; Viet et al. 2014, 2013; Truong et al. 2014; Zhan et al. 2013; Xu et al. 2013) have demonstrated clear effects of single charged-residue insertion, deletion, or substitutions on conformational ensemble and aggregation of IDPs monomers. Simple electrostatic models predict that modifications of residue charge will directly affect ensemble properties. (Das et al. 2015; Larini et al. 2013; Bah & Forman-Kay 2016; He et al. 2015). Locally, such mutations can modulate residual secondary structure preferences via forming or breaking local salt-bridges or by introducing helix breaking residues (Conicella et al. 2016; Ganguly & Chen 2015; Zhan et al. 2013).

For IDPs with a relatively low fraction of charged residues, typical of the Janus region of the state diagram proposed by Das and Pappu (Das et al. 2015; Das & Pappu 2013) (Fig 1c), more subtle differences among neutral amino acids play an increasingly important role in determining the ensemble. More than 40% of disease-associated IDP polymorphisms annotated in the human UniProtKB/Swiss-Prot database (Yip et al. 2008) are substitutions between two charge-neutral residues. The extent to which such substitutions in IDPs can affect non-local aspects of the conformational ensemble is uncertain; these substitutions directly affect short-range interactions, and structure-based coupling between distant residues in IDPs is expected to be weak. Nonetheless, correlations between secondary structure of distant residues has been frequently observed in IDPs (Ganguly & Chen 2015; Iešmantavičius et al. 2013; Feuerstein et al. 2012); for example, several cancer mutations in transactivation domain of tumor suppressor p53 can lead to

helicity changes in residues sequentially far away from the mutation sites (Ganguly & Chen 2015).

In structured proteins, contacts between residues distant along the sequence are reflected in the tertiary structure, but developing a framework for describing the analogous property in IDPs has not been straightforward. Among traditional structural biology techniques, NMR has been most useful for characterizing IDPs, but is frequently limited to residual secondary structure (Ref. (Mittag & Forman-Kay 2007; Habchi et al. 2014) and references therein). Molecular dynamics (MD) simulations have played a significant role in understanding IDP structure and dynamics (Stanley et al. 2015; Ithuralde et al. 2016; Knott et al. 2012; Invernizzi et al. 2013; Yedvabny et al. 2015; Levine & Shea 2017), but face limitations on chain length similar to those incurred in simulations of protein folding; most unbiased simulations have been performed in implicit solvent and/or involve chains too short to meaningfully sample contacts between residues far apart on the peptide chain. Studies of aggregation among multiple shorter monomeric IDPs (Levine et al. 2015; Pappu et al. 2008) have provided some of the most useful frameworks for considering tertiary contacts between residues which are distantly connected along the peptide backbone. Point mutations are also known to affect these contacts via differential salt-bridge and hydrogen-bonding formations, with mutations that change charge states affecting conformational ensemble via altered salt-bridge networks (Levine et al. 2015).

Many SNPs in IDPs are associated with neurological, aging-associated neurodegenerative, or psychiatric disorders; despite an exponential increase in the amount of available genetic data, identifying the genetic origins of such disorders has proven remarkably challenging, with few variants identified as replicable predictors of disease. One of the earliest identified variants is the Val66Met SNP (rs6265) in precursor brain-derived neurotrophic factor (BDNF), a signaling

protein that retains a critical role in neurogenesis and synaptogenesis throughout adulthood (Korte et al. 1995; Davies 2003) (Fig 1.1a). It has been implicated in maintenance of the hippocampus (Pezawas et al. 2004; Benjamin et al. 2010), orientation selectivity in the visual system (Huang et al. 1999; Liu et al. 2011; Gao et al. 2014) and the mechanism underlying action of numerous antidepressants (Autry & Monteggia 2012; Björkholm & Monteggia 2016), including rapidly acting low-dose ketamine (Autry et al. 2011). An extensive library of genome-wide association studies (GWAS) (and even earlier)have repeatedly identified the Val66Met SNP as reducing hippocampal volume and episodic memory, as well as predicting increased susceptibility to neuropsychiatric disorders including schizophrenia, bipolar, and unipolar depression, but associations have been inconsistent and population dependent (Soliman et al. 2010; Chen et al. 2008; Verhagen et al. 2010; Notaras et al. 2015; Autry et al. 2011).

Difficulties in obtaining unambiguous disease associations at the precursor BDNF Val66Met SNP using GWAS are paralleled by challenges in characterizing its effects on the properties of the BDNF prodomain using structural techniques. A crystal structure of a homologous neurotrophic factor in complex with a shared receptor revealed a well-defined volume corresponding to the prodomain, but lacked resolvable density (Feng et al. 2010). The prodomain sequence falls in the Janus sequence region in the phase diagram proposed by Das and Pappu (Das et al. 2015; Das & Pappu 2013).

It was subsequently revealed that the cleaved prodomains (91 residues) are found in monomeric states *in vivo*, and the M66 (but not V66) form binds to SorCS2 (sortilin-related VPS10p domain containing receptor 2), leading to axonal growth cone retraction (Anastasia et al. 2013) and eliminated synapses in hippocampal neurons (Giza et al. 2018). NMR measurements on the prodomain confirmed significant intrinsic disorder for both forms, with differential secondary structure preference around residue 66 (Anastasia et al. 2013). Tertiary contact

distances from NOEs were not accessible, however, an uncertainty in interpretation of the NMR signal obscured non-local effects on secondary structure. Additional NMR experiments implicated residue 66 in binding of M66 prodomain to SorCS2 (Anastasia et al. 2013).

In this work, we aimed to provide insight into the following questions: (1) What interactions drive the secondary structure change local to residue 66 observed through NMR? (2) How can we meaningfully detect tertiary interactions in a long disordered protein? (3) Do effects on tertiary interactions explain the non-local secondary structure changes previously observed through NMR? (4) How and why does the Val66Met mutation change tertiary interactions, especially as a charge-neutral mutation? To achieve these aims, we conducted unbiased fully-atomistic replica-exchange MD simulations of the 91 residue BDNF prodomain in explicit solvent, for V66 and M66 sequence.

We begin by identifying globular regions, or blobs within the protein using a sequence-based approach based on residue hydrophobicity; this is useful for both conceptualizing the long disordered protein in the absence of a well-defined topology, as well as focusing the analysis. We then compare our simulation results with previous NMR results of Anastasia et al. (Anastasia et al. 2013) and discuss the effects of the Val66Met SNP on residual secondary structure. We propose and apply an approach for decoupling short-range structural correlations from long-range structural correlations, by comparison with a simplified polymer model parameterized from the MD trajectories. We then discuss the effect of the Val66Met SNP on the network of correlated β strands between distant residues, illustrating how effects of the mutation propagate to tertiary contacts in which the mutation is not involved. Finally, we identify individual residue sidechains that drive the observed effects on this network. Our results suggest an important and previously-unconsidered role for specific Met-Met interactions in transducing the effects of the BDNF Val66Met SNP, and confirm the presence of weak but

long-range structural correlations in a disordered protein.

1.3 Results

1.3.1 Prodomain sequence decomposition

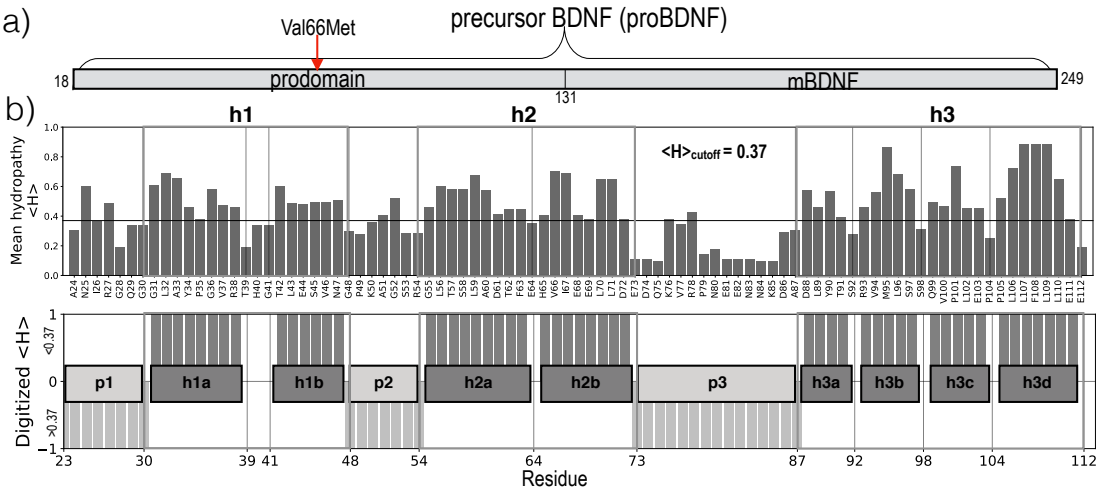

Figure 1.1: **Sequence-based decomposition of the BDNF prodomain.** a) The two functional domains of precursor BDNF: the disordered prodomain considered in this manuscript and the structured mature domain BDNF (mBDNF). b) The mean hydrophobicity $\langle H \rangle$ per residue (top), given by the Kyte-Dolittle (Kyte & Doolittle 1982) score averaged over a three residue window, and scaled to fit between 0 and 1 was digitized (bottom) according to a cutoff at $\langle H \rangle > 0.37$. Three or more contiguous residues above the cutoff were identified as forming a hydrophobic blob. Eight hydrophobic "h" blobs (darkgrey) are identified along with 3 "p" blobs of low hydrophobicity (light grey).

The region of the BDNF prodomain studied using NMR (Anastasia et al. 2013), and simulated here, is 91 residues long. Conceptualization of long structured proteins relies heavily on the consecutive secondary structure elements that form the protein's topology, allowing for a coarse cartoon-style representation.

No such approach for constructing an IDP topology has been available. Our original motivation for identifying globular segments in the sequence was to improve statistical power in analyzing contacts, but we found the resulting topological description to be broadly useful for interpretation of results. We thus present this conceptual tool upfront for clarity.

To avoid ambiguity, we restrict use of the term "domain" to refer to the two major BDNF domains (mature domain and prodomain), and instead specify three levels of hierarchy below the domain level: the prodomain contains multiple "regions", regions contain "groups", and groups contain "blobs". Blobs and groups were identified by sequence alone, as described in Methods, while regions were identified by Monte Carlo simulation of a simplified polymer representing the blobs.

The sequence-analysis approach outlined in Methods divides the sequence into alternating groups, classified as either hydrophobic (h groups) or non-hydrophobic (p groups). The prodomain is composed of six such groups, notated as p1-h1-p2-h2-p3-h3 from N-terminus to C-terminus. The h groups are further divided into blobs (Fig 1.1b), indexed with a letter. Each hydrophobic group contains two to four blobs: h1 contains h1a and h1b, h2 contains h2a and h2b, and h3 contains h3a, h3b, h3c, and h3d. We denote multiple consecutive blobs within a group by multiple letters: h3ab indicates the stretch of residues between the beginning of blob h3a and the end of blob of h3b. Each p group consists of just one blob. The results in Section "Regions of tertiary enrichment" led us to further designate Region I (containing p1 through h2), Region II (comprised of p3) and Region III (comprised of h3).

Since each blob sequence has its own properties (Table 1.1), this process also suggested a new, more tractable conceptualization of the long, disordered BDNF prodomain. Each blob can be analyzed individually according to Das and Pappu metrics (Das & Pappu 2013) (Fig 1.2a) or Uversky metrics (Uversky et al. 2000)

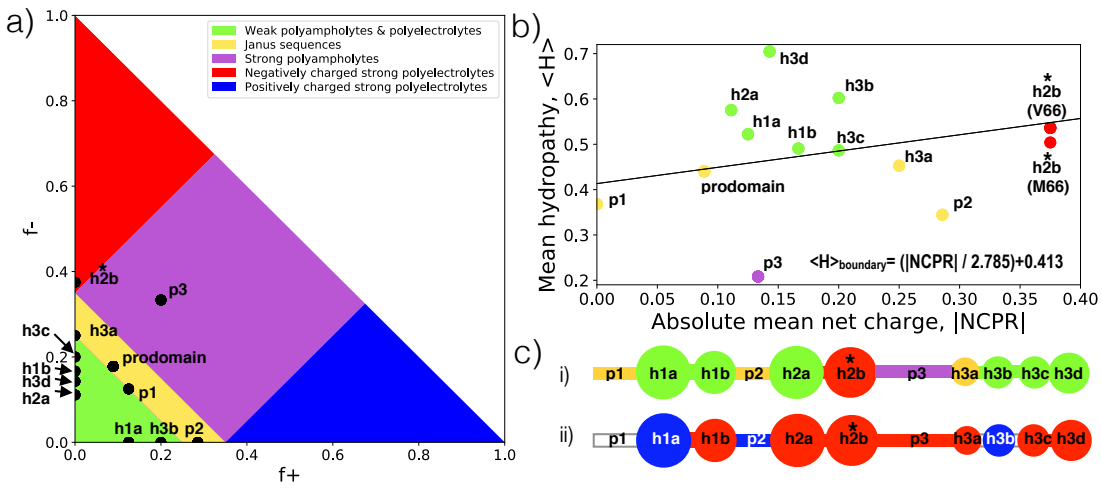

Figure 1.2: **Properties of hydrophobic and linker blobs identified in the BDNF prodomain.** a) The diagram of IDP states proposed by Das and Pappu (Das & Pappu 2013), based on fraction of positive (f^+) and negative (f^-) charged residues, and annotated by the location of the simulated BDNF prodomain and each blob identified in Fig 1.1b. b) Location of simulated BDNF prodomain and each blob on an Uversky diagram (Uversky et al. 2000) of IDPs and globular proteins, as a function of absolute net charge per residue ($|NCPR|$) and $\langle H \rangle$, with the boundary line between folded and disordered proteins given by the equation in the legend. c) Blobs identified in Fig 1.1b), colored according to the Das and Pappu (Das & Pappu 2013) diagram in a) (i) or net charge (ii), with negative charge (red), positive charge (blue) and neutral (white) or . The blob h2b contains the Val66Met mutation and is marked with star. Additional properties of the blob sequences can be found in Table 1.1.

(Fig 1.2b), while several other sequence properties of each blob are shown in Table 1.1. The Das and Pappu phase diagram (Das & Pappu 2013) predicts the compactness of IDPs based on their fraction of positively (f+) and negatively (f-) charged residues (Fig 1.2a). Hydrophobic blobs h2b and blob h3a lie in the strong polyelectrolyte and Janus sequence region respectively. All the remaining hydrophobic blobs are classified as weak polyampholytes and, as isolated peptides,

Table 1.1: Sequence based properties of hydrophobic and linker blobs identified in the BDNF prodomain, as shown in Fig 1.2.

Region	Group	Blob	N[a]	NCPR[b]	$\langle H \rangle$[c]	FCR[d]	f-[e]	f+[f]	κ[g]	Sequence	R[h]	P[i]
I	p1	p1	8	0.00	0.37	0.25	0.13	0.13	0.8	EANIRGQG	2	0.00
	h1	h1a	8	0.13	0.52	0.13	0.00	0.13	1.0	GLAYPGVR	1	0.13
		h1b	6	-0.17	0.49	0.17	0.17	0.00	0.1	TLESVN	1	0.00
	p2	p2	7	0.29	0.34	0.29	0.00	0.29	0.4	GPKAGSR	2	0.14
	h2	h2a	9	-0.11	0.58	0.11	0.11	0.00	0.7	GLTSLADTF	1	0.00
		h2b(V66)	8	-0.38	0.54	0.38	0.38	0.00	0.3	HVIEELLD	4	0.00
		h2b(M66)	8	-0.38	0.50	0.38	0.38	0.00	0.3	HMIEELLD	4	0.00
II	p3	p3	15	-0.13	0.21	0.53	0.33	0.20	0.1	EDQKVRP NEENNKDA	3	0.06
III	h3	h3a	4	-0.25	0.45	0.25	0.25	0.00	N/A	DLYT	2	0.00
		h3b	5	0.20	0.60	0.20	0.00	0.20	N/A	RVMLS	1	0.00
		h3c	5	-0.20	0.49	0.20	0.20	0.00	N/A	QVPLE	1	0.20
		h3d	7	-0.14	0.70	0.14	0.14	0.00	1.0	PLLFLLE	1	0.14
V66 Seq			91	-0.09	0.44	0.26	0.18	0.09	0.2		2	0.07
M66 Seq			91	-0.09	0.44	0.26	0.18	0.09	0.2		2	0.07

[a] Number of residues in the blob

[b] Net charge per residue

[c] Mean hydrophobicity, average of Kyte-Dolittle(Kyte & Doolittle 1982) scores for each residue in the blob scaled to fit between 0 and 1

[d] Fraction of charged residues

[e] Fraction of negatively charged residues

[f] Fraction of positively charged residues

[g] Charge distribution parameter κ as defined by Das and Pappu (Das & Pappu 2013), calculated using CIDER (Holehouse et al. 2017)

[h] Region in phase diagram proposed by Das and Pappu (Das & Pappu 2013), (Fig 1.2a)

[i] Fraction of Proline residues

would be predicted to have compact globule conformations to shield hydrophobic residues (Das & Pappu 2013). Linker blobs p1 and p2 also lie in the Janus sequence regions, while blob p3 lies in the strong polyampholyte region with low value of the charge distribution parameter κ (Das & Pappu 2013), and is predicted to have random coil conformations if present as an isolated peptide.

The Uversky diagram (Uversky et al. 2000) characterizes proteins as globular or intrinsically disordered based on their normalized mean hydrophobicity and absolute net charge per residue ($|NCPR|$) (Fig 1.2b). The proteins falling above the boundary line are classified as globular proteins, while the ones below that line are generally classified as IDPs. With the exception of hydrophobic blobs h2b and h3a, all hydrophobic blobs identified here fall in the globular protein regions. Blobs h2b, h3a and p1 fall on the disordered side of the boundary, while p2 and p3 are both deep in the disordered region.

The blob h2b containing Val66Met has several unique properties among the identified blobs: 1) it is located at the sequence midpoint 2) it is the only strong polyelectrolyte blob 3) it has the highest NCPR (-0.38) among all the blobs 4) its sequence is composed almost entirely of two competing residue types, yielding the uncommon mix of a highly-charged, hydrophobic blob. Considering mean hydrophobicity alone, Uversky et al. (Uversky et al. 2000) found $\langle H \rangle \sim 0.48 \pm 0.03$ for a set of 275 folded proteins and $\langle H \rangle \sim 0.39 \pm 0.05$ for a set of 91 unfolded proteins. By this criteria, we would expect the h2b sequence to be folded: for V66-h2b, $\langle H \rangle \sim 0.54$, while for M66-h2b, $\langle H \rangle \sim 0.50$. The full Uversky diagram also considers NCPR, and the high NCPR pushes h2b into the IDP region of the Uversky diagram (Uversky et al. 2000).

More specifically, this blob sequence (HV/MIEELLD) has hydrophobic residues at i, i+3, i+4 separated by acidic residues at i+1, i+2. Helix formation would thus segregate hydrophobic residues from acidic residues but would also increase

the density of like-charge residues. Similar sequences are observed in the activation domains of transcription factors: a motif of alternating hydrophobic and acidic residues folds into an amphipathic helix upon binding, and the interactions between the amphipathic helix and the binding partner are mediated by hydrophobic residues, not charged residues (Brzovic et al. 2011; Uesugi et al. 1997; Radhakrishnan et al. 1997; Canales et al. 2017; Staller et al. 2018). Staller et al. (Staller et al. 2018) have earlier reported that in the disordered acidic activation domain of Gcn4, the acidic residues keep key hydrophobic residues exposed to solvent and binding partners.

The blob h3a is a unique hydrophobic Janus blob with high NCPR. Janus sequences have intermediate compositional biases and their conformations are context dependent (Das & Pappu 2013; Das et al. 2015). The SNP blob h2b and the Janus blob h3a are separated by the long (15 residue) strong polyampholyte linker p3, which has well mixed charge ($\kappa = 0.1$). The blobs h1a and h3b are positively charged and all the remaining hydrophobic blobs are negatively charged (Fig 1.2c).

1.3.2 Comparison of experimental observables and their computational analogues

NMR spectroscopy (Anastasia et al. 2013) has previously confirmed the intrinsic disorder of the prodomain. Many of the common force-field and water model combinations used for MD simulations are optimized for folded proteins, and are not recommended for IDPs (Mercadante et al. 2015; Piana et al. 2015). Piana et al. (Piana et al. 2015) showed that several such force-field and water model combinations produced substantially more compact disordered states when compared with experiments. In order to predict accurate ensembles of the prodomain, we

tested several force-field and water model combinations, optimized for IDPs, including a03sbws (Best & Hummer 2009; Best et al. 2014) with Tip4p/2005 (Abascal & Vega 2005), a99sbws (Lindorff-Larsen et al. 2010; Best et al. 2014) with Tip4p/2005 (Abascal & Vega 2005), a99sb*-ildn-q (Lindorff-Larsen et al. 2010; Hornak et al. 2006) with Tip4p-D (Piana et al. 2015) and c36m (Huang et al. 2017) with Tip3p (Jorgensen 1981) on 30 residue fragments of the V66 prodomain using temperature replica-exchange molecular dynamics (T-REMD), further described in SI. To minimize the effects of loss of long-range contacts in the 30 residue fragment, only $\Delta\delta C_\alpha$ were compared; $\Delta\delta C_\beta$ is more dependent on β-pairing within the sequence. Among all the force-fields tested only a03sbws with Tip4p/2005 and a99sb-ildn with Tip3p gave significant deviations with NMR. The three remaining force-fields gave good comparison ($\Delta\delta C_\alpha$ RMSD <0.5 ppm) (Fig 1.11, Table 1.2). This is also consistent with the force-field comparison study by Robustelli et al. (Robustelli et al. 2018), which observed that for IDPs with little or no secondary structure, both c36m and a99sb*-ildn-q with Tip4p-D yielded the best agreement with experimental NMR measurements.

The a99sb*-ildn-q/Tip4p-D forcefield was used for the full prodomain MD simulations further described in Methods. Fig 1.3 shows the C_α and C_β secondary chemical shifts calculated from the full-length simulations using SPARTA+ (Shen & Bax 2010) (further described in Methods) and compares them with the NMR secondary chemical shifts obtained from Anastasia et al. (Anastasia et al. 2013) for V66 sequence and M66 sequence. We obtain good agreement with NMR secondary chemical shifts: the discrepancy at each residue is <0.7 ppm, which is less than the individual SPARTA+ prediction uncertainties of $\sim$ 1 ppm (Shen & Bax 2010).

Comparison of the simulated hydrodynamic radius (R_h) generated from MD and NMR/SAXS is an additional useful validation measure. (Mercadante et al. 2015; Rauscher et al. 2015; Meng et al. 2018) R_h was calculated from the trajectory

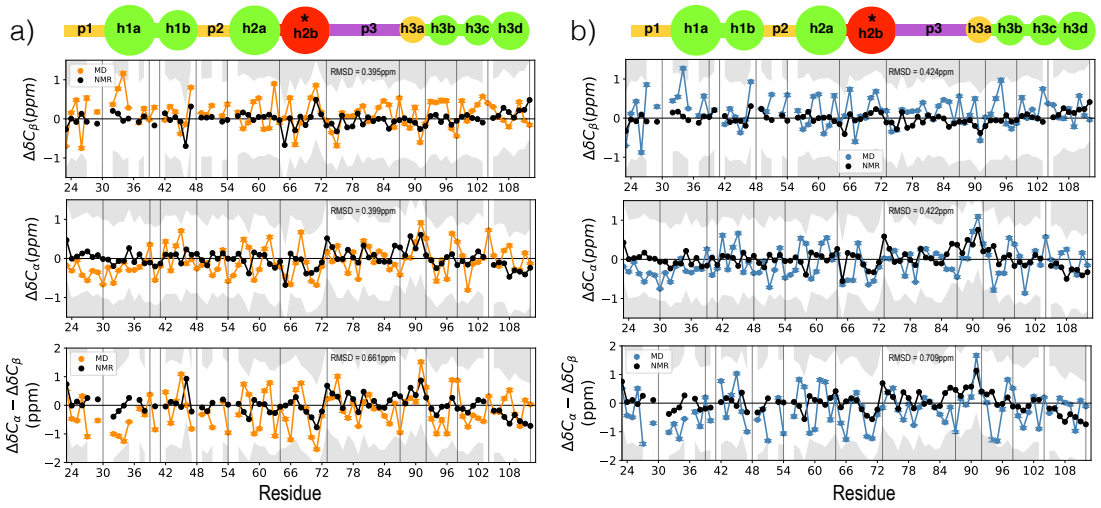

Figure 1.3: **Comparison of MD and NMR observables.** a) $\Delta\delta C_\beta$ (top), $\Delta\delta C_\alpha$ (middle), $\Delta\delta C_\alpha$-$\Delta\delta C_\beta$ (bottom) values from NMR at 280K (black lines) (Anastasia et al. 2013) and MD at 300K for V66 (a) and M66 (b). The gray region represents a discrepancy of more than 1 ppm from NMR secondary chemical shifts. Root-mean-squared deviation (RMSD) represents the deviation between the NMR and MD values. Error at each residue is calculated as the standard error in the mean, where $n = 1088$ is the product of total number of replicas simulated and average number of roundtrips per replica.

using Hydropro (Ortega et al. 2011) (further described in Methods). Hydrodynamic radii of both the V66 ($R_h = 2.202 \pm .006$ nm) and M66 ($R_h = 2.187 \pm .005$ nm) sequences are in excellent agreement with the experimental values from NMR diffusion measurements (Anastasia et al. 2013) ($R_h = 2.24 \pm 0.1$ nm and $R_h = 2.20 \pm 0.1$ nm for the V66 and M66 sequence respectively) (Fig 1.9a). (Error bars for simulation results represent statistical uncertainty and do not contain the additional systematic uncertainty of about 5% or 0.1 nm associated with use of Hydropro (Ortega et al. 2011).) Our results support the use of Tip4p-D with a99sb*-ildn-q for simulations of disordered proteins (Piana et al. 2015; Robustelli et al. 2018); use of Tip3P resulted in more compact ensembles. Although the M66 sequence is slightly more compact, the distributions of both R_h and the

simulated radius of gyration (R_g) demonstrate that the V66 and M66 sequence populate closely overlapping ensembles (Fig 1.9a).

1.3.3 Effects of Val66Met on local and non-local secondary structure

Anastasia et al. (Anastasia et al. 2013) reported an increase in helical tendency for the M66 sequence within blob h2 and h3ab and an increase in β tendency within blob h3b in the V66 sequence (Fig 1.4a). Consistent with these NMR experiments (Anastasia et al. 2013), the M66 sequence demonstrates an increased tendency of forming helices within blob h2 and h3a relative to the same blobs in the V66 sequence (Fig 1.4b). Comparing the length of secondary structure formed at each residue (Fig 1.4c) reveals an even stronger effect of the mutation that would not have been detectable via NMR: Val66Met consistently increases the frequency of long helices formed within group h2.

In general, C^{β}-branched amino acids, such as valine, have more restricted side-chain rotamers in helical conformation when compared with non-C^{β}-branched amino acids. Creamer et. al. (Creamer & Rose 1992) ranked the entropic cost of helix formation for apolar side chains using simulations of an $(Ala)_8$ sequence with the guest amino acid at the center, and reported a higher entropic cost of helix formation for valine when compared with methionine. In our simulations, the likelihood that Val66 will be in a short helix decreases with temperature, while the opposite effect is observed for Met66 (Fig 1.12). These trends are consistent with an increased entropic cost for helix formation at Val66 relative to Met66.

The helical structure within group h2 in M66 is also stabilized by local sequence, including the favorable interaction between Met66(i) and Phe63(i-3). MD simulations have previously shown the stability of a sulfur-aromatic contact in a model helix (Viguera & Serrano 1995). Fig 1.4d shows the residue level contact

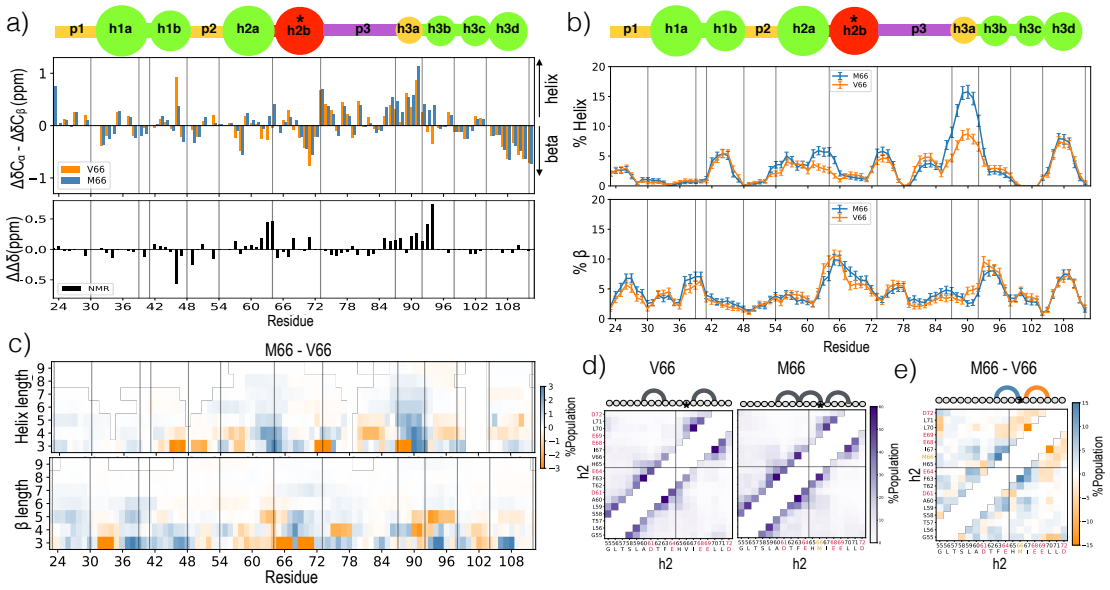

Figure 1.4: **Effects of Val66Met on secondary structure.** a) $\Delta\delta C_\alpha$ - $\Delta\delta C_\beta$ values for the V66 sequence and M66 sequence from NMR (Anastasia et al. 2013). Values on top are equivalent to the two NMR curves shown in Fig 1.3c, while the difference between the two curves is shown at the bottom. b) Helix (top) or β (bottom) propensity for each simulated residue of the 300K replica, defined as the probability of a given residue being part of a sequence of four or more consecutive residues whose dihedral angles place them in the helical (left) region or β (right) region of the Ramachandran map (further described in Methods). Errors represent standard error of a Bernoulli trial with n samples, where $n = 1088$ is the product of the total number replicas (64) and average number of roundtrips per replica (17). c) Difference (M66-V66) between probabilities of secondary structure formation of a given length, for helix (top) and β (bottom). d) Contact probability for each residue pair within the h2 group for each sequence. Each residue in group h2 is annotated with a circle representation and the contacts with at least 50% population are represented with an edge. e) Difference between the contact probabilities shown in d)

map within group h2. For the M66 sequence, Met66 (i) more frequently contacts Phe63 (i-3) than any other residue within the blob: Met66-Phe63 is formed 60% more often than Met66-Glu69 (Fig 1.4d). We find that the largest change in intrablob contacts from V66 to M66 is the gain of contact at Met66-Phe63 (40% stronger in M66 when compared with V66) followed by loss of contact at Ile67 (i+1)-Leu70 (i+4) (30% weaker in M66 when compared with V66) (Fig 1.4e). This is also consistent with previously identified Met-Phe interactions (Viguera & Serrano 1995; Faure et al. 2008; Valley et al. 2012; Gómez-Tamayo et al. 2016)

While the effects of the Val66Met mutation on secondary structure in the blob which contains residue 66 (h2b) are not unexpected, we also observed an effect on secondary structure in group h1 and blobs h3a and h3b within group h3. As shown in Fig 1.4c, the increased frequency of long helices for blob h3a in the Met66 sequence is comparable to the increase in blob h2b. We consider the possible tertiary origins of the non-local effects on secondary structure in Section "Effects of Val66Met on the β-pairing network".

1.3.4 Regions of tertiary enrichment

The potential number of residue-residue contacts in the prodomain is $91 \times 90/2 \sim 4000$, and each contact is formed infrequently. Detecting significant differences for numerous weak signals is statistically prohibitive, even given the long simulations presented here. Dividing the sequence into blobs based on sequence hydrophobicity, as described in Methods, helps address this analysis challenge (Fig 1.1b). Such coarse-graining reduces the number of potential contacts to $11 \times 10/2 = 55$, while increasing the likelihood that any given contact will be formed.

We expect that even for a freely-jointed, self-avoiding heteropolymer (SAHP), contact probability between monomers would depend on monomer shape and separation, although a SAHP does not have tertiary structure. Inspired by the Kuhn treatment of real polymers,(Rubinstein & Colby 2003) we propose that the

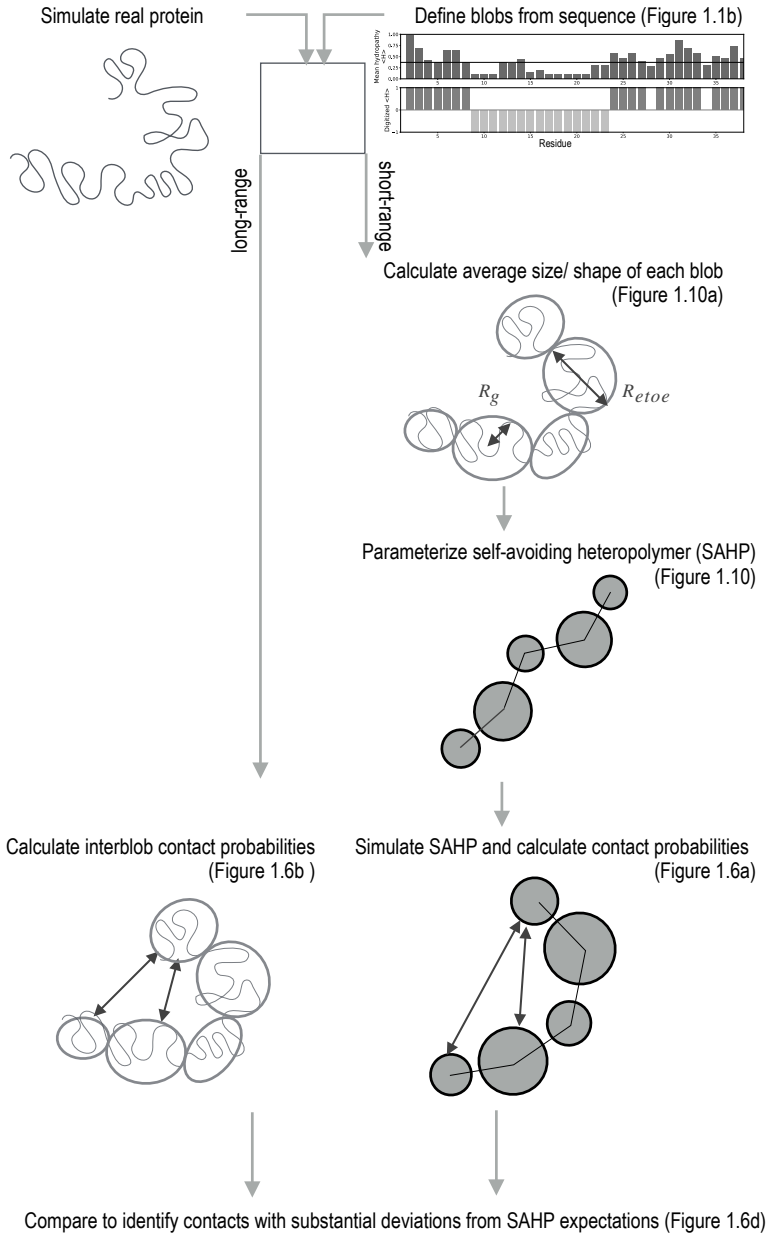

Figure 1.5: **Detection of Tertiary Enrichment** To decouple short-range from long-range structural correlations, this work grouped segments of the protein into blobs using sequence, and then compared contacts between the blobs to those expected for an analogous self-avoiding heteropolymer (SAHP). The SAHP was parameterized by extracting local properties (size and shape) of blobs from the real protein (RP) trajectory.

expected intermonomer contact frequency in a SAHP can be a useful reference for detecting specific tertiary interactions (Fig 1.5), as long as the monomers mimic the blobs of the real protein (RP). In support of this approach, we find that within a given blob, the protein obeys Flory polymer scaling laws (Section "Heterogeneous behavior of individual blobs" in SI). The exponent varies across blobs (Fig 1.13), capturing the intrinsic heterogeneity of the long polymer.

The predicted contact probabilities for this freely-jointed, self-avoiding heteropolymer are shown in Fig 1.6a. In the SAHP version of the prodomain, the chain is visibly segmented at p3 boundaries. As shown in Fig 1.14, shifting the p3 blob within the SAHP chain shifts the visible segmentation boundary, confirming that the p3 blob defines the segmentation. Based on this expectation, we define three regions: the pre-p3 blobs are "Region I", p3 is "Region II", and the post-p3 blobs are "Region III". SAHP blobs within Region I are in contact for 61% of the frames, while blobs within Region III are in contact in 81% of the frames. In comparison, the average contact probability between Regions I and III is only 10% (Fig 1.6c).

Fig 1.6b shows the probability of blob-blob contacts for both the V66 and M66 sequences of the RP, calculated analogously to those in the SAHP. The frequencies of contacts within Region I and within Region III were quantitatively consistent with the SAHP predictions. The total number of blob-blob contacts within Region I was enriched by 1.2 times the expected value for the SAHP. Within Region III, the total number was depleted by 0.9 times the expected value (Fig 1.6c). In contrast, contacts between blobs on either side of the long p3 linker are more common in the RP than in the SAHP, and are also affected by the substitution at residue 66 (Fig 1.6c,d,e). Contacts between pre-linker Region I and post-linker Region III are about three times as common in the RP as in the SAHP, indicating specific tertiary interactions beyond those expected for a polymer undergoing a random-walk. Quantitatively, enrichment in the V66 sequence is 3.0±0.1 while

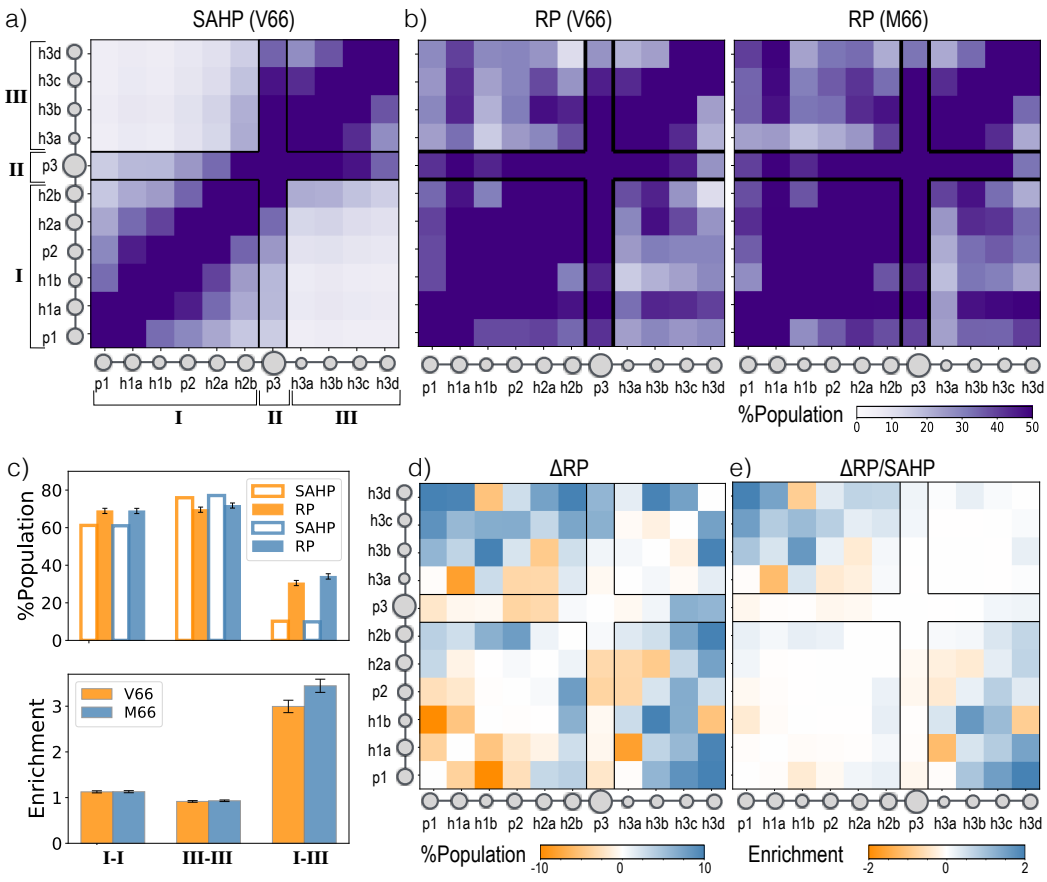

Figure 1.6: **Effect of Val66Met on contacts between blobs.** a) Blob-blob contact probability for the V66 self-avoiding heteropolymer (SAHP) from Monte Carlo simulations (further described in Methods). The black boxes mark the regions identified. b) Blob-blob contact probability shown in a) for the real protein (RP); V66 (left) and M66 (right) sequences. The x and y axes are annotated with cartoon representations of the prodomain; circles are drawn to the scale of each blob's size. c) Population of contacts (top) and enrichment in RP contacts with respect to SAHP contacts (bottom) for each region pair. The errors represent standard errors (n =1088 as described in Methods). d) Difference between the contact probabilities shown in b). e) Differences shown in d) with respect to SAHP; interactions more frequently found in M66 or V66 are in blue and orange respectively.

enrichment in the M66 sequence is 3.4±0.1. The increased number of cross linker contacts are also consistent with the lower mean R_h (Fig 1.9a) and R_g (Fig 1.9a) for the M66 sequence.

1.3.5 Effects of Val66Met on the β-pairing network

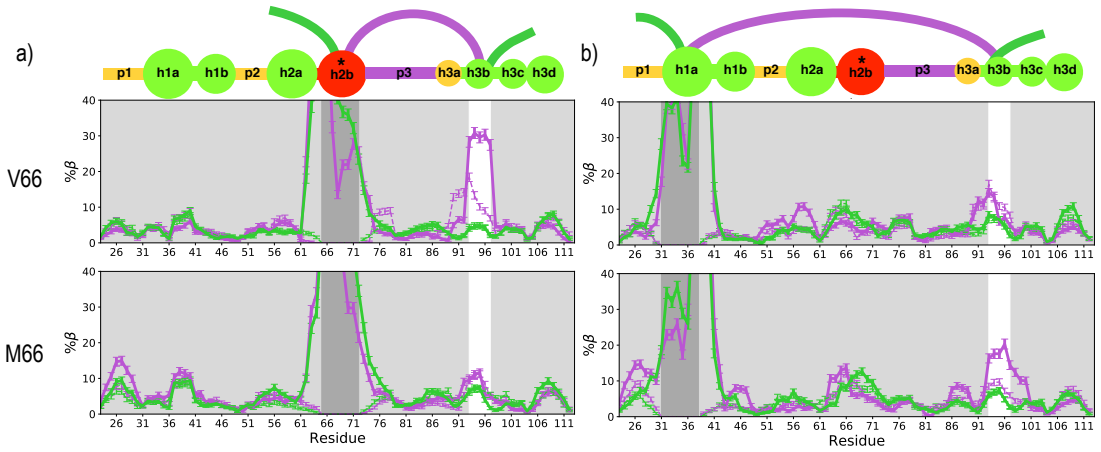

Figure 1.7: **Secondary structure coupling between blobs on each side of the p3 linker.** β propensities at each residue in V66 sequence (top) and M66 sequence (bottom) for four clusters. Frames were first clustered by whether the h3b-h2b (a) or h3b-h1a (b) contact was formed (purple) or broken (green), and then by whether β structure was present in h2b (solid) (a) or h1a (b) or absent (dashed). The dark-gray window indicates the contacting blob that is constrained to have high or vanishing values by construction of the cluster, while the white window indicates the contacting blob without constrained secondary structure. If the contact is coupled to simultaneous β-strand formation, the peak within the white window for the solid purple curve should be significantly higher than other curves. Errors represent standard error of a Bernoulli trial with n number of samples, where n is the product of total number of unique replicas in a given cluster and average number of roundtrips per replica (17).

To test whether the changes we observed in tertiary contacts at the blob,

group, or region level could be due to a change in partnering β-strands, we applied a clustering approach. All frames were divided into 4 clusters, representing two independent collective variables with two possible values each: either a certain contact between blobs X and Y is formed or broken, and any residue in blob X is found within a stretch of 4 sequential residues in β conformation. The four clusters are thus represented as (contacting,absent), (contacting,present), (distant,absent), and (distant,present).

For each cluster, we calculated β propensity across all residues (Fig 1.7). If the X-Y contact reflects correlated β-strands, we expect a peak at residues in blob Y in the (contacting,present) cluster that is significantly higher than the signal for all other clusters. If the secondary structure in Y is used for clustering instead, the reciprocal peak (at blob X) should be reproduced. Furthermore, unless there are higher-order correlations between multiple sets of β-strands, β propensity should not depend on cluster for all residues *not* in blob X or Y.

This clustering process on all frames was carried out for all possible X and Y blobs pairs (Appendix A). For most pairs, there was no correlating peak in β structure. For some pairs, a peak was present in one direction but the reciprocal peak was not present in the opposite direction. This result reflected longer β-strands that extended to a neighboring blob, which had the true peak. One symmetrically significant peak (indicating correlated β structure) involving the h3b was observed in each sequence. The partner blob shifted from h2b in the V66 sequence to h1a in the M66 sequence (Fig 1.7). A second correlated pair involving the blob p1 was also observed in each sequence. The partner blob for this pair shifted from h3d in the V66 sequence to h2b in the M66 sequence (Appendix A).

Despite loss of correlated β-pairing, the contact between h2b and h3b is actually more probable in the M66 sequence than in the V66 sequence (Fig 1.6d). As discussed in Section "Noteworthy residue-residue interactions stabilizing tertiary contacts", this result reflects a significant change at the residue level. In the M66

sequence, specific interactions between Met66 and side-chains of residues within h3b form the contact, rather than backbone-backbone interactions. As the h3b side-chains stabilize the contact with h2b, the backbone of h3b is then free to pair with h1a, increasing the number of favorable long-range contacts and condensing the M66 sequence overall.

1.3.6 Noteworthy residue-residue interactions stabilizing tertiary contacts

As shown in the previous section, the Val66Met substitution causes loss of correlated β-strands between blobs h2b and h3b, while introducing correlated β-strands between blobs h3b and h1a. We consider here the effects of the substitution on these contacts at residue level. As shown in the absolute residue-residue contact probability maps (Fig 1.8a), both sequences frequently form contacts between hydrophobic residues in blobs h2b and h3b. The residue pairs most frequently forming the contact shift from Val66-Val94 in the V66 sequence to Met66-Met95 in the M66 sequence (Fig 1.8b). The residue-level contact maps also show a high probability of contacts between Asp72 and Thr91 in the V66 but not M66 sequence. As illustrated in Fig 1.8c, these contacts (between α carbons) are stabilized by salt-bridges between Arg93 and Asp74, in a conformation that is incompatible with a side-chain contact between Val/Met66 and Met95.

Met95 is the only other methionine in the simulated sequence. The role of specific Met-Met interactions due to polarizable sulfur atoms is often underappreciated, but are common in structures of folded proteins (Faure et al. 2008). Using ab initio calculations, Gómez-Tamayo et al. (Gómez-Tamayo et al. 2016).

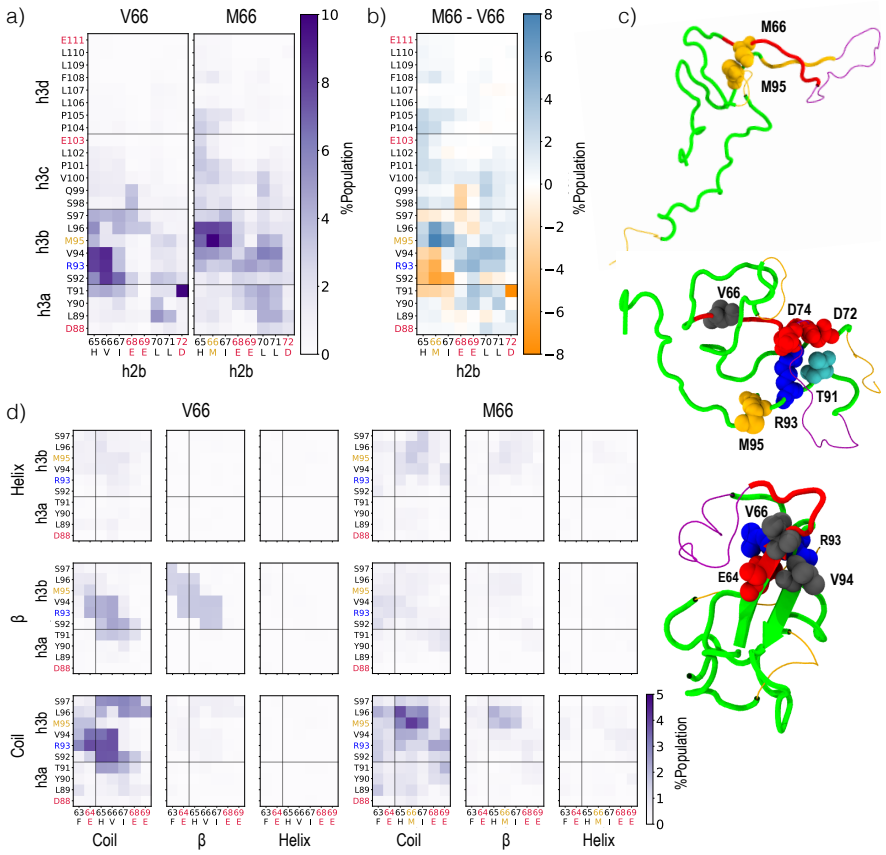

Figure 1.8: **Effect of secondary structure in group h2 on which residues form the cross-boundary h2-h3 contact.** a) Contact probability at each residue in h2b with each residue in h3 for V66 (left) and M66 (right). b) Difference between the contact probabilities shown in a). c) Representative conformation of M66 sequence and V66 sequence showing preferred residue-level contacts in VDW representations, with residues colored by residue type: blue:basic, red:acidic, cyan:polar, grey:hydrophobic, Met: yellow and chain colored as in Fig 1. Tubes represent hydrophobic "h" blobs whereas lines represent non-hydrophobic linker "p" blobs. d) Contact probability between residues 63-69 and each residue in h3ab, when respective secondary structure is formed at each residues, for both the V66 and M66 sequences. Residue labels are colored according to residue type: blue:basic, red:acidic, grey:hydrophobic/polar and Met: yellow.

predicted that Met-Met interactions are stronger than Met-aromatic or aromatic-aromatic interactions, due to the polarizability of sulfur. Although the fixed-charge forcefield we are using (a99sb*-ildn-q) cannot explicitly capture polarizability, Gómez-Tamayo et al. demonstrate that this forcefield preserves rankings of strong side-chain interactions involving methionine. In these simulations, the Met66-Met95 contact was about five times as common (10% of frames) as the analogous Val66-Met95 contact (2% of frames) (Fig 1.8a and b). Methionine-aromatic interactions also contribute to the increased number of Region I-III contacts: M66, but not V66, forms a frequent contact with F108 in blob h3d, which is also consistent with the favorable interactions between Met-Phe residues (Viguera & Serrano 1995; Faure et al. 2008; Valley et al. 2012) (Fig 1.8a and b).

To determine which residue contacts between h2b and h3ab couple the secondary structure within the two blobs, we decomposed the residue-level contact maps into nine clusters. Each cluster was specified by two collective variables with three possible values each: secondary structure (helix, β, or coil) around residue 66 and secondary structure (helix, β, or coil) in h3ab (Fig 1.8d). The β-pairing at h2b-h3ab is stabilized via a combination of backbone hydrogen bonds between Val66 and Ser92, salt-bridge between Glu64 and Arg93, and hydrophobic interactions between Val66 and Val94. The Val66-Met95 contact was only formed frequently within the (h2b - coil, h3ab - helix) cluster, and since this cluster was a very small part of the overall population, the contact overall was rare as well (Fig 1.8d). This cluster was more common in the M66 sequence, and contributes to the non-local increase in helicity around residue 95 (Fig 1.4b).

1.4 Discussion

We have carried out over 250 μs of fully-atomistic explicit solvent MD simulation of the 91 residue BDNF prodomain, with and without the disease-associated

Val66Met mutation. These long simulations successfully reproduced the experimentally observed secondary chemical shifts and R_h. The simulations also correctly reproduced the location of both local and non-local secondary changes due to the Val66Met mutation in the prodomain sequence.

We find that the highly disordered 91 residue prodomain, which as a whole falls in the Janus sequence region of the Das and Pappu phase diagram (Das & Pappu 2013), can be meaningfully divided into 11 blobs based on sequence hydrophobicity alone. Among 8 hydrophobic blobs, we identified 2 blobs in the disordered region: the strong polyelectrolyte blob h2b (which contains Val66Met), and the Janus blob h3a. These are connected via the highly disordered long linker p3. The groups containing these unique blobs have biological significance as well: The sequence h2-p3-h3 is essential for intracellular trafficking of precursor BDNF (Chen et al. 2005).

The sequence decomposition framework suggested a tractable approach for coarse-graining analysis, by reducing the initial number of potential contacts from over 4000 (Fig 1.15,Fig 1.15) to 55 (Fig 1.6), while increasing the number of observations for each contact. Furthermore, it allowed us to isolate the most sensitive regions of the protein for examination at the residue level. This method, simply based on sequence hydrophobicity, may be a generally useful strategy to identify functionally significant regions in proteomics investigations of long disordered proteins. Our conclusions further suggest an important role for disorder heterogeneity within disordered proteins.

We were able to identify mechanisms through which a charge-neutral mutation can affect a disordered protein's residual secondary structure and tertiary contacts, as well as how these effects can be propagated to non-local residual secondary structure. Within its local blob h2b, the Val66Met mutation affects local contact preference due to local sequence effects (preferred Met-Phe contacts) and the reduced entropic cost of helix formation for the methionine sidechain.

The long, disordered, exposed region II linker segregates the blob-level contact probability map: blobs within Region I or Region III have a high probability of contact, while Region I-III contacts are far less probable. We consistently observed this segregation in both simple self-avoiding hetropolymer simulations with beads mimicking identified blobs, and actual prodomain simulations. Val66Met increases the frequency of Region I-III contacts. We find here that the dominant mechanism involves replacing β-strand coupling between group h2 of Region I and group h3 of Region III with favorable Met/Met side-chain interactions between the same groups. The group h3 backbone is then exposed for interactions with the backbone of group h1, also of Region I. The non-local increase in helicity in group h3 may reflect stabilization of non-β structure by the Met-Met interactions.

Met/Met interactions have been shown to stabilize tertiary contacts in folded proteins and membrane proteins, but their role has not been investigated in disordered proteins. In general, our study supports previous observations (Gómez-Tamayo et al. 2016; Lim et al. 2019) that methionine plays a distinct role from true aliphatic residues in determining protein structure, and highlights the importance of mimicking its unique properties within fixed-charge forcefields.

Anastasia et al. (Anastasia et al. 2013) observed differential kinetics for interactions between the BDNF prodomain and SorCS2, and also observed that the SNP-containing blob h2b (H65 to L71) only interacts with SorCS2 in the M66 sequence. The increased interactions between M66 and SorCS2 could be attributed to increased helical propensity at that residue and/or specific Met-Met contacts. In the first mechanism, helix formation in the SNP blob segregates acidic and hydrophobic residues on opposite sides of the helix. It is possible that this preformed structure will stabilize binding. The second mechanism is suggested by the specific Met-Met interactions we observed in the isolated prodomain, as well as the high number of exposed methionines on the SorCS2 surface. It is also possible both mechanisms could contribute to stabilizing the complex, although

this would require a more specific protein-protein interface.

1.5 Methods

System setup

To account for differences in starting coil conformation, we included six unique structures to represent residues 23-113 of BDNF prodomain. The six structures were built using I-Tasser (Yang et al. 2014; Roy et al. 2010; Zhang 2008), Rosetta (Kim et al. 2004) or Modeller (Šali & Blundell 1993), and were simulated in a water box at 600K for 50 ns at a constant volume. From the six resulting trajectories, 64 structures with correct proline isomers were selected (based on at least 2ps time interval); in total, our study included 64 unique prodomain structures. All structures were cooled to 300K for 1ns, while prolines were restrained in trans-conformation. Each V66 replica was placed in a dodecahedron water box with approximately 30,500 Tip4p-D (Piana et al. 2015) water molecules and a 0.15M salt concentration (NaCl) for a total system size of approximately 124,000 atoms. The same volume for each replica was ensured by fixing the simulation box of each replica to the average box size (11 nm).

MD Simulations

For the simulations we use the a99sb*-ildn-q force-field(Lindorff-Larsen et al. 2010; Hornak et al. 2006) and the GROMACS 5.1.2 simulation package,(Berendsen et al. 1995; Abraham et al. 2015) with a time step of 2 fs. Long-range electrostatics are calculated using the particle mesh Ewald (PME) method (Essmann et al. 1995), with a 1 nm cutoff and a 0.12 nm grid spacing. Periodic boundary conditions are also used to reduce system size effects. System was simulated using T-REMD(Sugita & Okamoto 1999) with an exchange frequency of 1ps for 2 μs,

giving a total simulation time of 128 μs with NVT ensemble for each system. 64 replicas are used with temperatures ranging from 300-385K, with exponential spacing. A different random seed was used for the Langevin dynamics of each replica. The average exchange acceptance probability ranged between 0.19-0.23.

The minimum separation between the molecule and its image was less than 2 nm only <1% of the total simulation time for both sequences and these frames were discarded from all the analysis. Time-series of the relative measurements were generated every 100 ps. For both V66 and M66 sequences, initial 51.2 μs (800 ns $\times$ 64) trajectories were discarded for equilibration purposes, determined by plateauing of R_g (Fig 1.9a). Simulation convergence was monitored using several metrics (Fig 1.9). Over the course of remaining 76.8 μs (1.2 μs $\times$ 64) simulations, each replica completes a minimum of 5 roundtrips and an average of 17 roundtrips for each sequence (Fig 1.9e).

Time-series of the radius of gyration R_g and end-to-end distance R_{etoe} were calculated using respectively the g_gyrate and g_polystat utilities of Gromacs. We took R_{etoe} as the distance between N-termini and C-termini N and O atoms respectively. Statistical uncertainties are provided for R_g, R_h as the standard error in the mean, where $n = 1088$ is the product of total number of replicas simulated (64) and average number of roundtrips per replica (17).

Blob identification

Mean hydrophobicity ($\langle H \rangle$) at each residue is defined as the average Kyte-Dolittle(Kyte & Doolittle 1982) score with a window size of 3 residues, scaled to fit between 0 and 1. Any stretch of four or more residues with $\langle H \rangle > 0.37$ is classified as a hydrophobic or h blob and any stretch of four or more residues with $\langle H \rangle \leq 0.37$ is classified as a non-hydrophobic linker or "p" blob. Multiple consecutive hydrophobic blobs without a "p" blob separating them are classified as a single group.

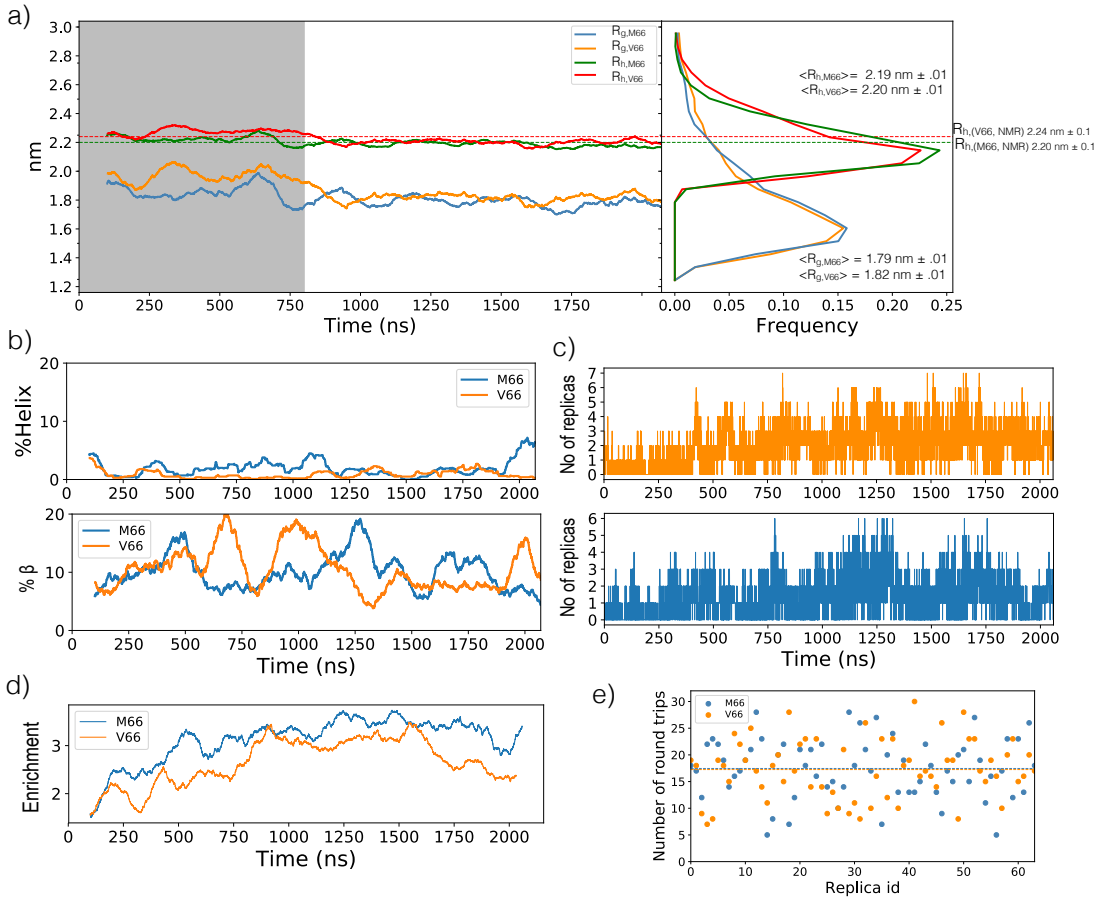

Figure 1.9: **Simulation convergence.** a) R_h, R_g vs the simulation time, using a 100 ns moving window on left and R_h,R_g distributions on the right at 300K. The R_h measured from NMR diffusion (Anastasia et al. 2013) is represented with dashed lines and the shaded region represents the discarded equilibration period. The R_h, R_g distribution and its mean does not include the simulation equilibration time. b) helix (top) and β (bottom) propensity at residue 66 vs the simulation time for both the sequence at 300K. c) Number of replicas forming Val66-Val94 contact (top) and Met66-Met95 contact (bottom) vs the simulation time. d) Enrichment of total Region I-Region III contacts relative to SAHP in the 300K replica vs simulation time. e) The number of round trips completed by each replica over the course of 76.8 μs simulation time after discarding initial 800ns for equilibration for both V66 and M66 sequence.

Secondary Chemical Shifts

Prior to the present study, Anastasia et al.(Anastasia et al. 2013) measured chemical shifts for the BDNF prodomain (residues 21-113) using NMR, and then used backbone NMR secondary chemical shifts to predict secondary structure via TALOS+ (Shen et al. 2009) and SSP (Marsh et al. 2006). For comparison with simulation data, we reinterpreted the chemical shifts directly from(Anastasia et al. 2013), deposited at Biological Magnetic Resonance Bank (BMRB). Secondary chemical shifts are calculated as follows: $\Delta\delta\mathrm{C}_{\alpha,\,MD} = (\delta\mathrm{C}_{\alpha,\mathrm{MD}}$ - $\delta\mathrm{C}_{\alpha,\mathrm{RC(300K)}})$ for MD and $\Delta\delta\mathrm{C}_{\alpha,\,\mathrm{NMR}} = (\delta\mathrm{C}_{\alpha,\,\mathrm{NMR}}$ - $\delta\mathrm{C}_{\alpha,\mathrm{RC(280K)}})$, where $\delta\mathrm{C}_{\alpha,\mathrm{MD}}$ are predicted C_α chemical shifts from MD simulation ensembles using SPARTA+ (Shen & Bax 2010) and $\delta\mathrm{C}_{\alpha,\mathrm{NMR}}$ were obtained from BMRB (Anastasia et al. 2013). Random coil chemical shifts ($\delta\mathrm{C}_{\alpha,\mathrm{RC}}$) for the 91 residue BDNF prodomain were obtained using POTENCI (Nielsen & Mulder 2018) at pH 7, 0.15 M ion concentration and 280K and 300K for NMR and MD respectively. Error at each residue are calculated as the standard error in the mean, where $n = 1088$ is the product of total number of replicas simulated (64) and average number of roundtrips per replica (17).

Hydrodynamic radius calculation

The values for the Hydropro (Ortega et al. 2011) parameters were: atomic level model with shell-method calculation, a = 0.29 nm, 6 minibead iterations, and σ = 0.1 to 0.2 nm. The temperature was taken to be 300 K, the solvent viscosity 0.01 Poise, the solvent density was 1.0 gcm^{-3}, the partial specific volume of the peptide 0.7313 cm^3g^{-1} (V66 sequence) or 0.7304 cm^3g^{-1} (M66 sequence), and molecular weight of the peptide was equal to 10044 Da (V66 sequence) or 10076 Da (M66 sequence). The resultant translational diffusion constants were then used for calculating R_h using the Stokes-Einstein equation. Error is calculated as

the standard error in the mean, where $n = 1088$ is the product of total number of replicas simulated (64) and average number of roundtrips per replica (17).

Secondary structure calculation

Helix propensity or β propensity is expressed as the probability of a given residue being part of a sequence of four consecutive residues whose dihedral angles place them in the helical region or β region of the Ramachandran space. The helical region is defined as -100° $<\phi <$-30° and -120° $\leq \psi \leq$ 50°(Nodet et al. 2009; García & Sanbonmatsu 2002; Knott et al. 2012). The β region is defined as ϕ <-80° and 50° $<\psi <$-120°. The error bars are calculated with standard error of a Bernoulli trial with n number of samples, where n is the product of total number of unique replicas in a cluster and average number of roundtrips per replica (17). The length of secondary structure (SS-map) (Iglesias et al. 2013) were calculated with the above defined helical and β region.

Blob-level contact maps

The excess distance between any two blobs i and j is defined as (Fig 1.10c)

$$d_{e,ij} = |\vec{r}_i - \vec{r}_j| - (R_{g,i} + R_{g,j}) \tag{1.1}$$

where $\vec{r}_i$ is the position vector of a blob i defined as the mean of its N-termini N atom and C-termini O atom coordinates, calculated using g_traj utility of Gromacs. Two blobs i and j are in contact if the excess distance ($d_{\mathrm{e,ij}}$) between the two is less than 0.55 nm. At residue level, two residues are in contact if the distance between C_α atoms of the two residues is 0.8 nm or less. Presented statistical uncertainties are the standard error in the mean, with n is the product of total number of replicas forming the given contact and the average number of roundtrips per replica, 17.

Self-avoiding heteropolymer simulation

The BDNF prodomain was approximated as a freely-jointed self-excluding heteropolymer with 11 monomers, each mimicking one of the blobs identified in Fig 1.1. The separation between monomers i and $i+1$ (analogous to the Kuhn length for a homopolymer (Rubinstein & Colby 2003)) was constrained to be half the end to end distance for each of the analogous blobs (Fig 1.10d):

$$|\vec{r}_{i-1} - \vec{r}_i| = \frac{\langle R_{etoe,i-1}\rangle + \langle R_{etoe,i}\rangle}{2} \tag{1.2}$$

where $\langle R_{etoe,i}\rangle$ was determined from the coordinates of blob i residues in the MD simulations, shown in Fig 1.10a.

Two monomers i and j are considered to be overlapping if

$$\frac{|\vec{r}_i - \vec{r}_j|}{\langle R_{g,i}\rangle + \langle R_{g,j}\rangle} = \frac{d_{e,ij}}{\langle R_{g,i}\rangle + \langle R_{g,j}\rangle} + 1 < a \tag{1.3}$$

where $\langle R_{g,i}\rangle$ was determined from the coordinates of residues in blob i in the MD simulations (Fig 1.10a), and a is a constant. In the MD simulations of the real protein, we observed that $\frac{d_{e,ij}}{\langle R_{g,i}\rangle + \langle R_{g,j}\rangle} \geq -0.7$ for almost all frames (Fig 1.10b), and thus we set $a = 0.3$.

The random walk was carried out using a simple Metropolis Monte Carlo, with the following move set: 1) a random bead $i > 0$ was selected, 2) a random displacement vector $\vec{\delta r}$ was generated in three cartesian dimensions, 3) $\vec{\delta r}$ was scaled so that $|\vec{r}_{i-1} - (\vec{r}_i + \vec{\delta r})| = (\langle R_{etoe,i-1}\rangle + \langle R_{etoe,i}\rangle)/2$, satisfying Eq 1.2, 4) the translation $\vec{r}_j \rightarrow \vec{r}_j + \delta_r$ was applied for all $j \geq i$.

Any trial move that caused an overlap according to Eq. 1.3 was rejected, while all others were accepted. The MC simulation was run for 500,000 steps (50,000 steps per moveable bead); additional steps did not change the outcome in Fig 1.6a.

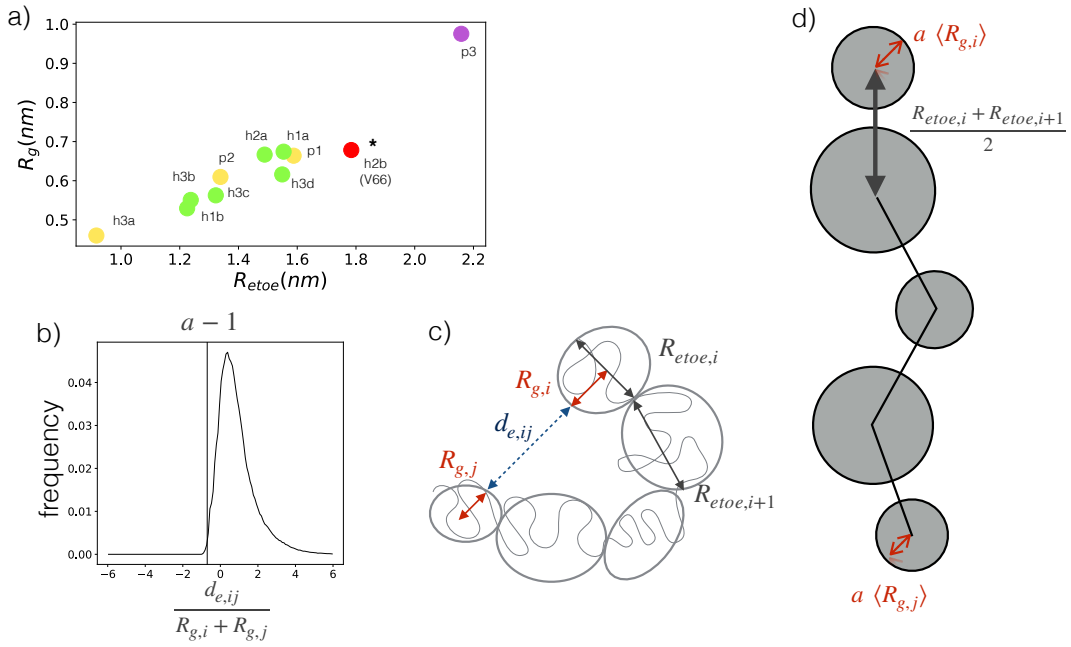

Figure 1.10: **Parameterization of self-avoiding heteropolymer.** a) Average R_{etoe} vs average R_g for each blob of V66 sequence. Error at each residue is calculated as the standard error in the mean, where $n = 1088$ is the product of total number of replicas simulated and average number of roundtrips per replica. The errors were smaller than the circles used for the representation of each blob. b) The distribution of normalized excess distances across all blob-pairs in the V66 RP, where $|i-j| > 1$. c) Relationship between the radius of gyration $R_{g,i}$, end to end distance $R_{etoe,j}$, and excess distance d_{ij}, calculated for each blob or blob pair using a RP trajectory. d) The SAHP is a chain with each monomer representing a blob of the real protein and modeled as a hard sphere. Each monomer i has radius $aR_{g,i}$ and is separated from monomer $i+1$ by bond length $(R_{etoe,i} + R_{etoe,i+1})/2$. Bond lengths are constrained and bond angles can rotate freely.

Acknowledgments

The authors are grateful to Dr. Clay Bracken and Dr. Barbara Hempstead of Weill Cornell Medical Center for helpful discussions. Computational time was provided through XSEDE resources from the National Science Foundation as well as the Rutgers Discovery Informatics Institute, which is supported by Rutgers and the State of New Jersey (Parashar et al. 2018).

1.6 Supporting Information

Table 1.2: **Summary of force-field comparison simulations.**

Force-field	$\Delta\delta C_\alpha$	$\langle R_g \rangle$	equilibration time	no of replica
aff03sbws (Tip4p/2005)	0.855	1.347 ±0.007	400 ns	36
a99sb*-ildn-q (Tip4p-D)	0.355	1.270 ±0.007	200 ns	36
a99sbws (Tip4p/2005)	0.425	1.277 ±0.007	200 ns	36
c36m (Tip3p)	0.350	1.306 ±0.007	200 ns	30
a99sb-ildn (Tip3p)	0.617	0.922 ±0.003	200 ns	32

Heterogeneous behavior of individual blobs

We also calculated the polymer properties of each blob. Disordered proteins can be well-described by Flory scaling theory $\langle R_{|i-j|} \rangle = A|i-j|^{\nu}$, where $\langle R_{|i-j|} \rangle$ is the ensemble-averaged internal distance, |i-j| is residue separation along the chain, and ν is the Flory scaling coefficient (Flory 1949). Larger values of ν correspond to swollen coils, while smaller values correspond to compact globules (Das & Pappu 2013). In particular, when ν=0.6 ("good solvent") the protein maximizes its interaction with solvent, and for ν=0.33 ("poor solvent"), the protein maximizes self-interactions. The special intermediate case of ν=0.5 is called a "theta solvent" (Flory 1949). Most IDPs that obey this scaling behavior have

ν>0.5 (Hofmann et al. 2012; Das & Pappu 2013; Zerze et al. 2015; Meng et al. 2018).

As shown in Fig 1.13 the prodomain as a whole is not well fit by a single power law: for separations of 15 or fewer residues the prodomain falls in the "theta solvent" regime, while for separations of 20 or more residues it falls in the "poor solvent" regime. Each identified individual blob does obey a power law, and we calculated A and ν for each blob as if it was isolated from rest of the protein (Fig 1.13). The highest observed value of ν was in blob h2b and h3c. This is in agreement with strong polyelectrolyte nature of h2b and high content of Proline residue (20%) in h3c.

Method: We calculated the average distance between the first atom (N) and last atom (O) for all residue pairs of a given sequence as a function of sequence separation $|i-j|$ using g_traj. Errors before fitting were calculated as the standard error in the mean, where $n = 1088$ is the product of total number of replicas simulated (64) and average number of roundtrips per replica (17). ν was calculated by linear fit of $\ln(\langle R_{|i-j|}\rangle)$ vs $\ln(|i-j|)$ weighted by each point's pre fit error with fixed A of 0.59 nm. To exclude the short-range backbone rigidity, distances with $|i-j| < 3$ were not fit.

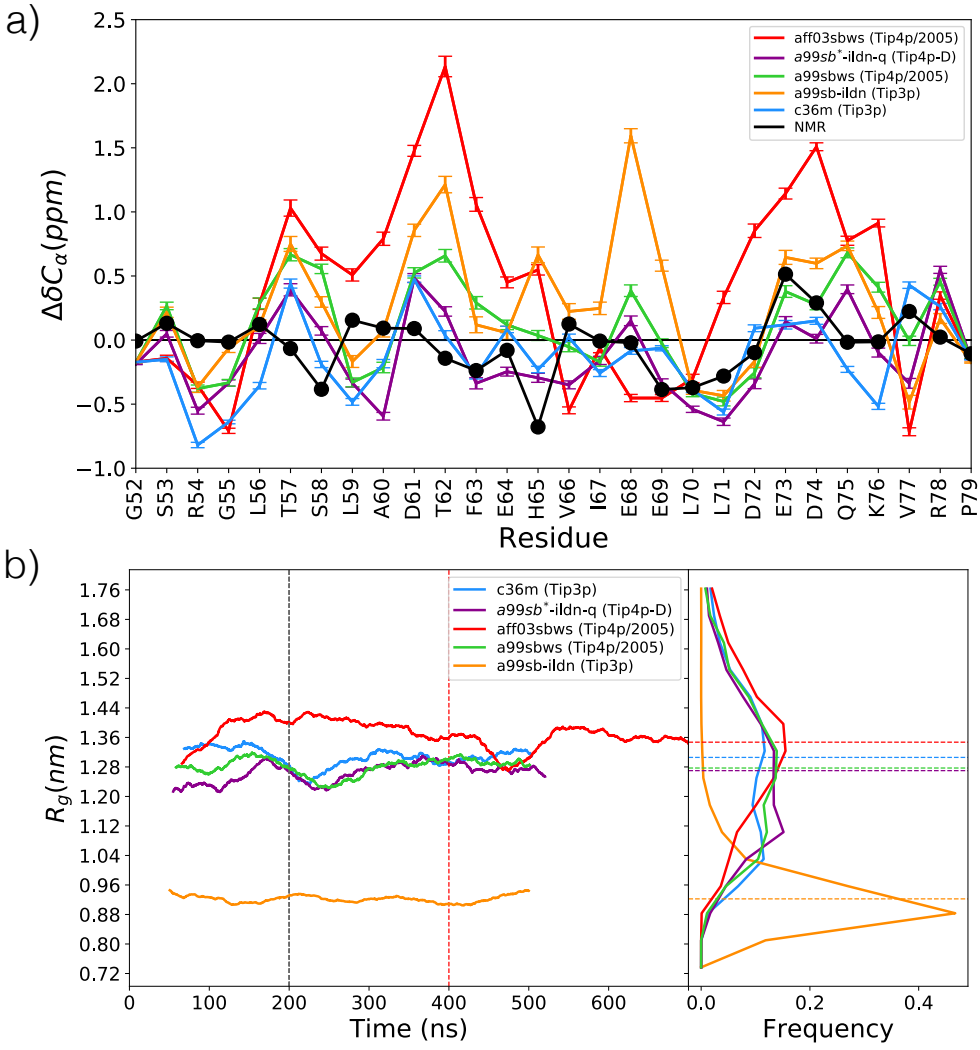

Figure 1.11: **Force-field comparison.** We ran T-REMD simulations of a 30 residue fragment of the V66 prodomain with several commonly used force-field and water model combinations. (a) Comparison of $\Delta\delta C_\alpha$ secondary chemical shifts at 280K from MD ensembles for a99sb*-ildn-q (Lindorff-Larsen et al. 2010; Hornak et al. 2006) with Tip4p-D (Piana et al. 2015), c36m (Huang et al. 2017), a99sbws (Lindorff-Larsen et al. 2010; Best et al. 2014), a03sbws (Best & Hummer 2009; Best et al. 2014), a99sb-ildn with Tip3p (Jorgensen 1981), calculated using SPARTA+ (Shen & Bax 2010) and NMR from Ref. (Anastasia et al. 2013). (b) R_g vs the simulation time, using a 100 ns moving window on left and R_g distribution for each force-field on right. Tip3p and a03sbws generates most collapsed and expanded R_g distribution respectively. The equilibration time and $\langle R_g \rangle$ is shown with vertical and horizontal dashed lines for each force-field. The R_g distribution and its mean does not include the simulation equilibration time.

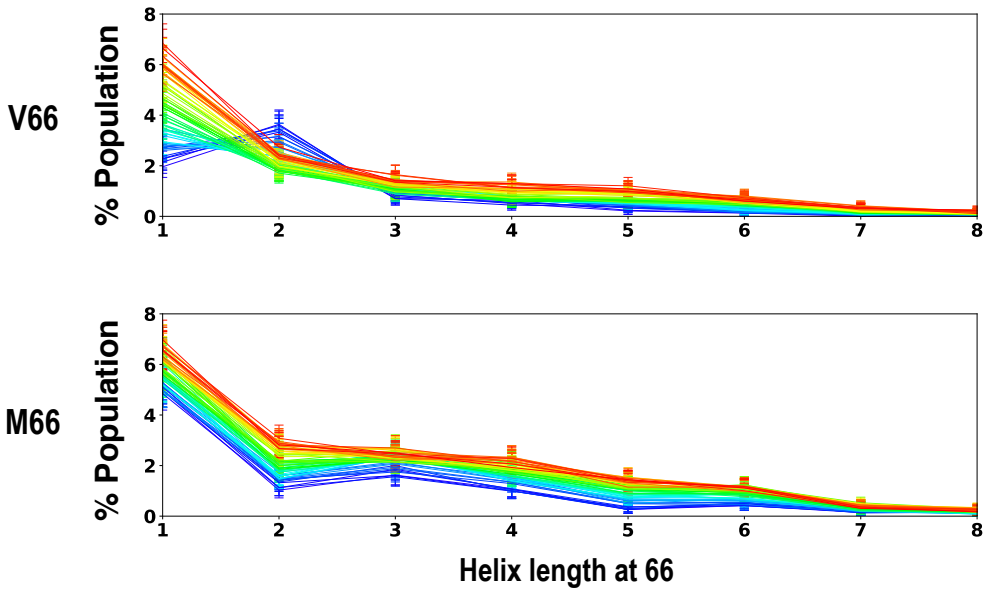

Figure 1.12: **Effects of temperature and Val66Met mutation on helix propensity around residue 66.** Frequency of helix of a given length at residue 66 in V66 (top) and M66 (bottom) in the temperature range of 300K to 385 K. With the increase in temperature the color transitions from cooler (blue) to hotter (red). It is entropically unfavorable for V66 and its neighboring residue to be simultaneously in the helical region of the Ramachandran map, as indicated by the decreasing helical propensity with increasing temperature. For longer helices, the trend will depend more on the additional side-chains in the helix, and the trend with temperature is reversed, but it remains weaker than the analogous trend for the M66 sequence. Errors represent standard error of a Bernoulli trial with n number of samples, where n is the product of total number unique replicas forming the helix of given length at residue 66 at a given temperature and average number of roundtrips per replica, 17.

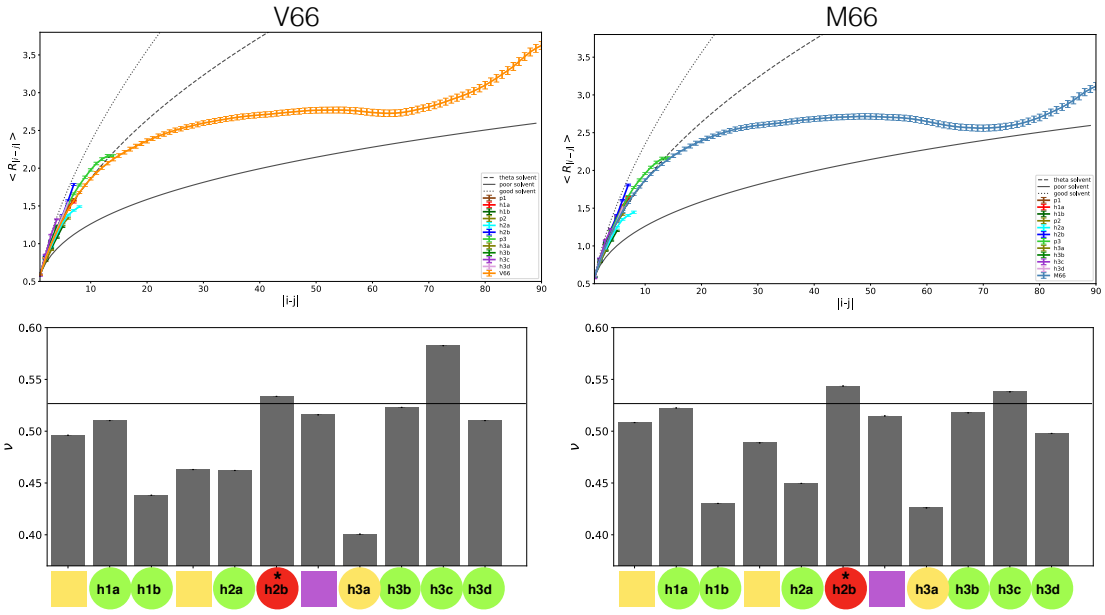

Figure 1.13: **Scaling behavior of each identified blob.** Ensemble averaged interchain distance profiles for the entire V66 and M66 prodomain and each blob in the sequence. Theoretical polymer scaling limits are shown with grey lines (prefactor A = 0.59 nm) (top). Flory exponents for each blob (bottom).

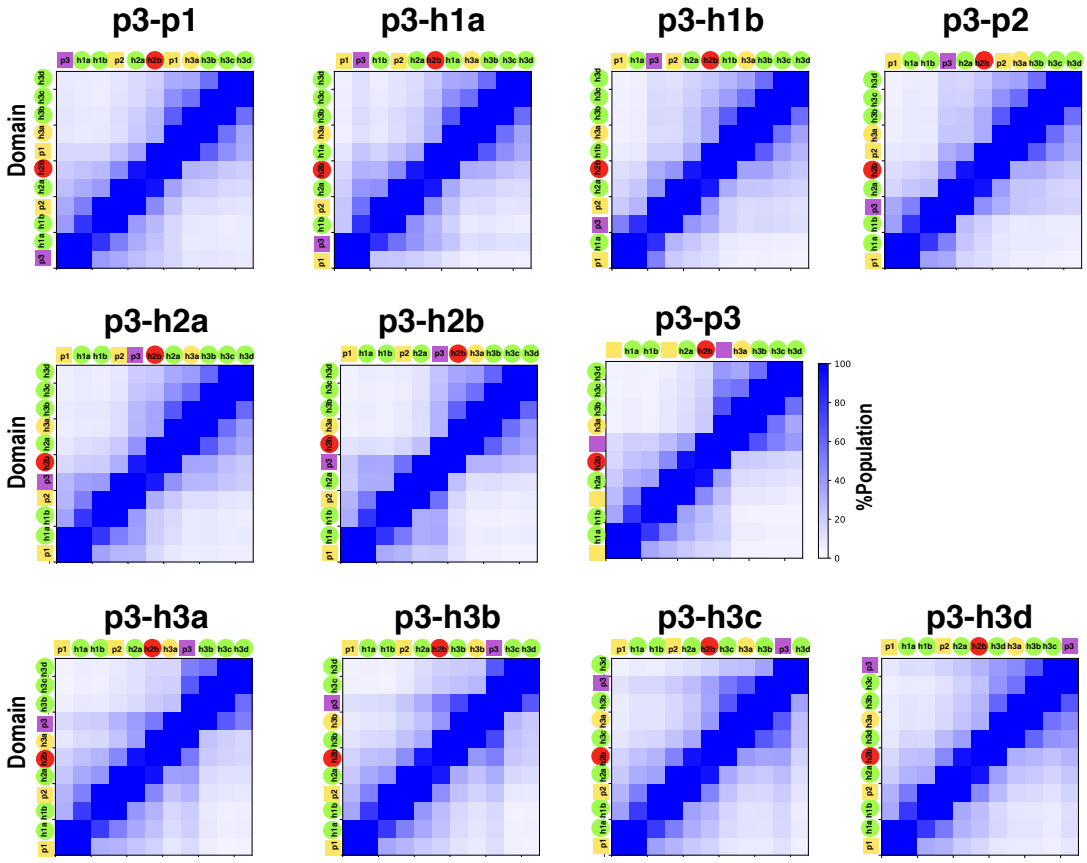

Figure 1.14: **Effect of perturbing monomer properties on freely-jointed, self-avoiding heteropolymer** Contact probability maps from MC simulations, analogous to those in Fig 5a of the main text, in which the blob p3 is swapped with every other blob in the chain, with the new location represented by the purple square in the graph annotation. As the p3 blob is shifted along the chain, p3 and p1 consistently bound a white "forbidden" region that has little interaction with the rest of the protein.

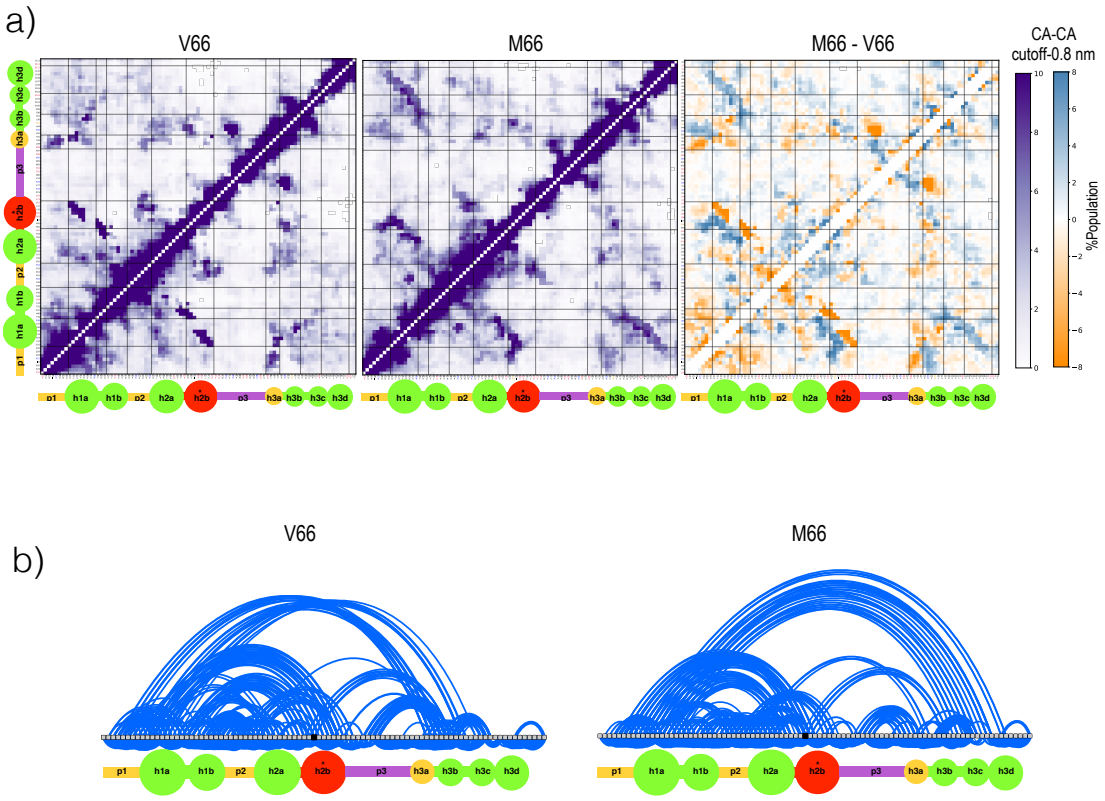

Figure 1.15: **Residue level contacts for the entire prodomain.** Contact probability between every residue pair for V66 (left) and M66 (middle) and M66-V66 (right). Two residue pairs are in contact if the distance between C_α-C_α atoms between the two residues are 0.8nm or less. b) A linear network of transient tertiary contacts shown in a). The contact networks were build using Cytoscape (Ahlstrom et al. 2013) with a linear representation of residues. Each protein residue comprises a node in the network, with interactions between residues represented as edges. The strength of individual interactions can be interpreted by the thickness of the edge line on the network diagram. If the separation between residues forming the contact is more than 3, its edge is drawn above the node; otherwise, the edge is drawn at the bottom of the node. To focus on significant interactions, interactions showing more than 4% persistence were considered in network visualization.

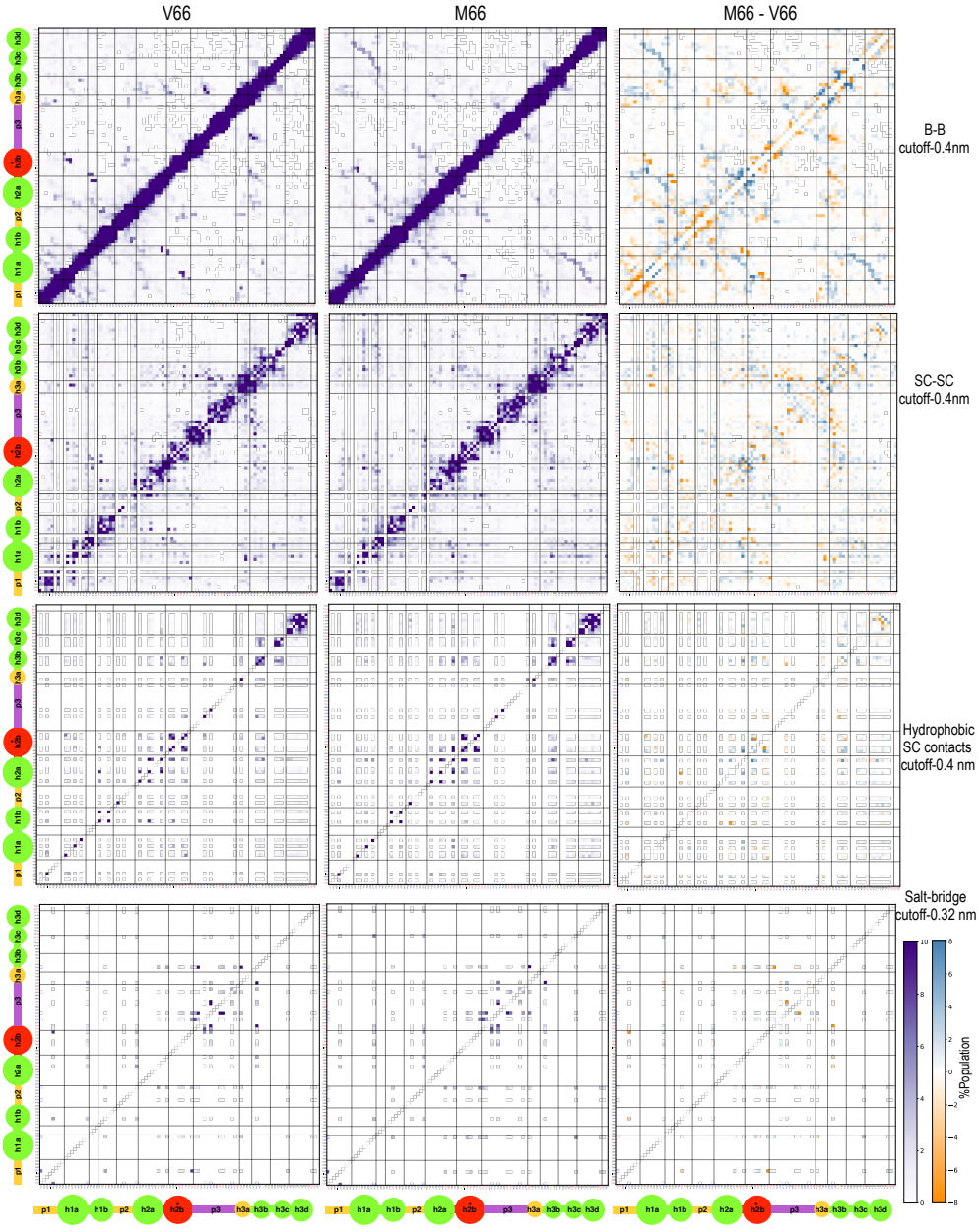

Figure 1.16: **Residue level contacts for the entire prodomain.** Contact probability between every residue pair for V66 (left) and M66 (middle) and M66-V66 (right). Two residue pairs are in contact if the distance between backbone-backbone atoms between the two residues are 0.4nm or less (1st row), if the distance between non hydrogen sidechain-siechain atoms between the two residues are 0.4nm or less (2nd row), if the distance between non hydrogen sidechain-siechain atoms between the two hydrophobic residues are 0.4nm or less (3rd row), if the two residue pairs are forming a salt-bridge with the distance between the donor and acceptor atoms < 0.32nm (4th row).

Chapter 2

Application of hierarchical analysis to other mutations.

2.1 Introduction

Although the role of electrostatic interactions and mutations that change charge states in intrinsically disordered proteins (IDPs) is well-established, many disease-associated mutations in IDPs are charge-neutral. In Chapter 1, we studied the effects of the disease-associated Val66Met substitution at the midpoint of the prodomain of precursor brain-derived neurotrophic factor (proBDNF) using fully atomistic molecular dynamics simulations. The Val66Met substitution is found in more than 30% of the human population (McGregor & English 2019), and has been widely studied for its association with aging-related and stress-related disorders, reduced volume of the hippocampus, and variations in episodic memory (McGregor & English 2019; Pezawas et al. 2004; Benjamin et al. 2010; Autry & Monteggia 2012; Björkholm & Monteggia 2016). We found that the Val66Met substitution changes the local and non local secondary structure and transient tertiary contacts in the BDNF prodomain. We developed a novel hierarchical sequence-based framework for analysis and conceptualization for measuring tertiary enrichment in disordered proteins.

To gain further insight into the generalizability of the identified mechanism and to further establish the relevance of the tertiary enrichment comparison, we performed two more sets of studies: 1) We tested the effect of 5 more hydrophobic substitutions (A66, I66, L66, F66, Y66) at residue 66 in BDNF prodomain and

2) we tested the effect of changing charge by protonating histidine at residue 65 (H^{65+}) in V66 and M66 sequence. In this chapter, we report on fully- atomistic temperature replica exchange molecular dynamics simulations of the 91 residue BDNF prodomain for these 7 substitutions; A66, I66, L66, F66, Y66, $V66^{65+}$ and $M66^{65+}$sequence.

First we compare the secondary structure and tertiary enrichment in all 7 hydrophobic substitutions at residue 66 on BDNF prodomain ensemble. Then we compare the effect of protonating histidine at residue 65 in V66 and M66 sequence on BDNF prodomain ensemble. Finally we compare the chain properties from across all the 9 sequences simulated.

2.2 Results

2.2.1 Comparing the effect of 7 hydrophobic mutations at residue 66 on BDNF prodomain ensemble.

Effects of other hydrophobic mutations on local and non-local secondary structure

In Chapter 1, we found local and non-local secondary structure changes due to the Val66Met substitution at residue 66. V66 has the presence of increased β in the h3ab blob while M66 had increased helicity within blob h2ab and h3a.

Fig 2.1 compares secondary structure at every residue for each sequence simulated. We find that M66 still has the highest helical propensity at residues around the SNP when compared with every other sequence, followed by L66. Earlier we found that the M66 increased helicity at residue 66 was both due to lower entropic cost of helix formation and stabilized Met66 (i)-Phe63 (i-3) contact.

To further understand the origin of helix stabilization, we looked at residue level intrablob contacts within h2 and compared it with M66 sequence. Fig 2.2a

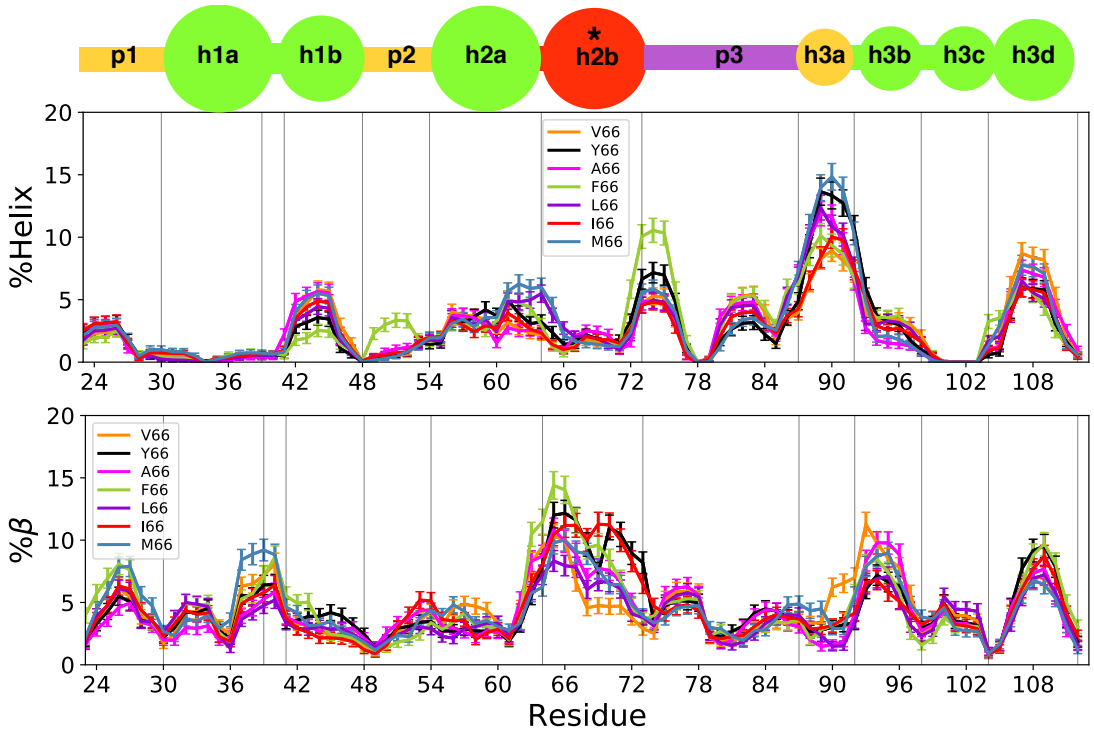

Figure 2.1: **Effects of hydrophobic substitution at residue 66 on secondary structure.** Helix (top) or β (bottom) propensity for each simulated residue of the 300K replica, defined as the probability of a given residue being part of a sequence of four or more consecutive residues whose dihedral angles place them in the helical (left) region or β (right) region of the Ramachandran map (further described in methods). Errors represent standard error of a Bernoulli trial with n samples, where n is the product of the total number replicas and average number of roundtrips per replica.

shows the residue level contact map within group h2. We find that the largest intra-blob contact difference is again found at residue 66-F63 in A66, I66, and F66 sequences. The remaining sequences did not show any large ($> 15\%$) intra-blob contact differences (Fig 2.2b). To closely look at the type of contact formed between 66-Phe63, we zoomed in the contact maps for backbone contact and sidechain contacts between these residues (Fig 2.2b). We find that M66, Y66, L66, and F66 have a higher probability of forming sidechain contacts with Phe63.

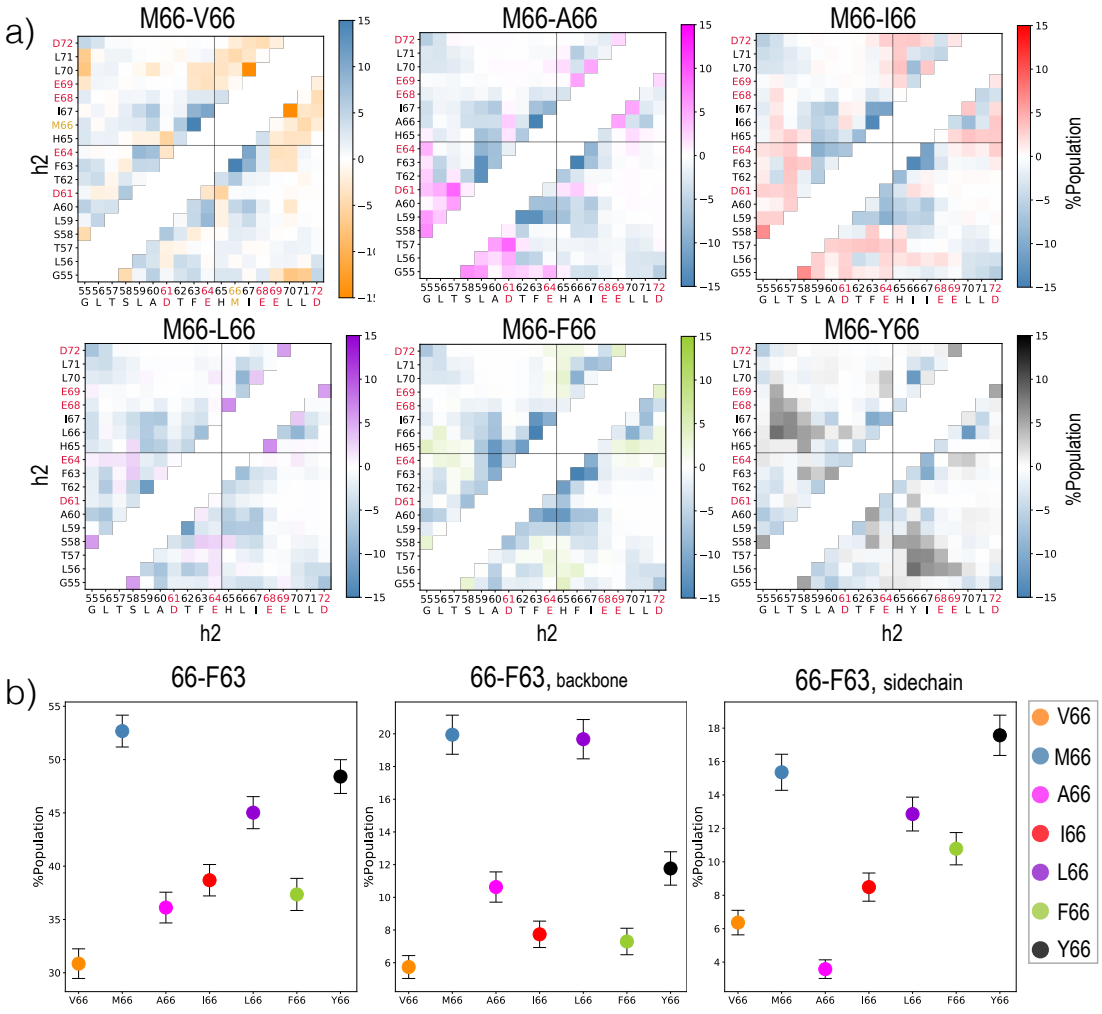

Figure 2.2: **Helix stabilization at residue 66.** a) Contact probability for each residue pair within the h2 group for each sequence relative to M66 sequence (across panels, blue indicates a greater contact likelihood for the M66 sequence than the other tested sequence). b) Contact probability only for 66-F63 pairs.

This is consistent with the preferred amino acid interaction pairs. Aromatic side chains of Tyr66 and Phe66 form preferred sidechain contacts with aromatic residue F63 due to preferred π-π interaction (Burley & Petsko 1985).

Creamer et. al. (Creamer & Rose 1992) ranked the entropic cost of helix formation for apolar side chains using simulations of an (Ala)8 sequence with the guest amino acid at the center, and reported the following order for entropic cost

of helix formation Val>Tyr>Ile>Phe>>Met>>>Leu>Ala.

Even though Y66 forms frequent Tyr66-Phe63 contacts, it does not show high helicity within h2, probably due to higher entropic cost of helix formation (Creamer & Rose 1992). Ala66 has the lowest entropic cost of helix formation but does not form any preferred contact with the local sequence. Analogous to helix stabilization of M66 sequence due to frequent Met66 (i)-Phe66 (i-3) contacts, we find that L66 helix is stabilized due to frequent Leu66 (i)-Phe63 (i-3) (Faure et al. 2008) side chain interaction and lower entropic cost of helix formation.

We find that local sequence and entropic cost of helix formation together define helix propensity. Apart from group h2, we find an increase in helicity within group p3 and p2 in F66 sequence. In Chapter 1 we found that Val66 has an increased β tendency within blob h3ab when compared with M66. We find this observation holds when compared with other mutations at residue 66 as well.

Regions of tertiary enrichment

In Chapter 1 we proposed the tertiary enrichment test to measure changes in tertiary contacts in IDPs. We find that intermonomer contact frequency in a SAHP was a useful reference for detecting specific tertiary interactions, as long as the monomers mimic the blobs of the real protein (RP).

Fig 2.3 shows the probability of blob-blob contacts for V66, M66, A66, I66, L66, F66 and Y66 sequences of the RP, calculated analogously to those in the SAHP.

We find consistent segmentation of contact maps into regions at p3 boundary for all the 7 sequences simulated. The frequencies of contacts within Region I and within Region III were quantitatively consistent with the SAHP predictions. We consistently find that the total number of blob-blob contacts within Region I was enriched by 1.2 times that expected for the SAHP. Within Region III, the total number was depleted by 0.9 times the expected value.

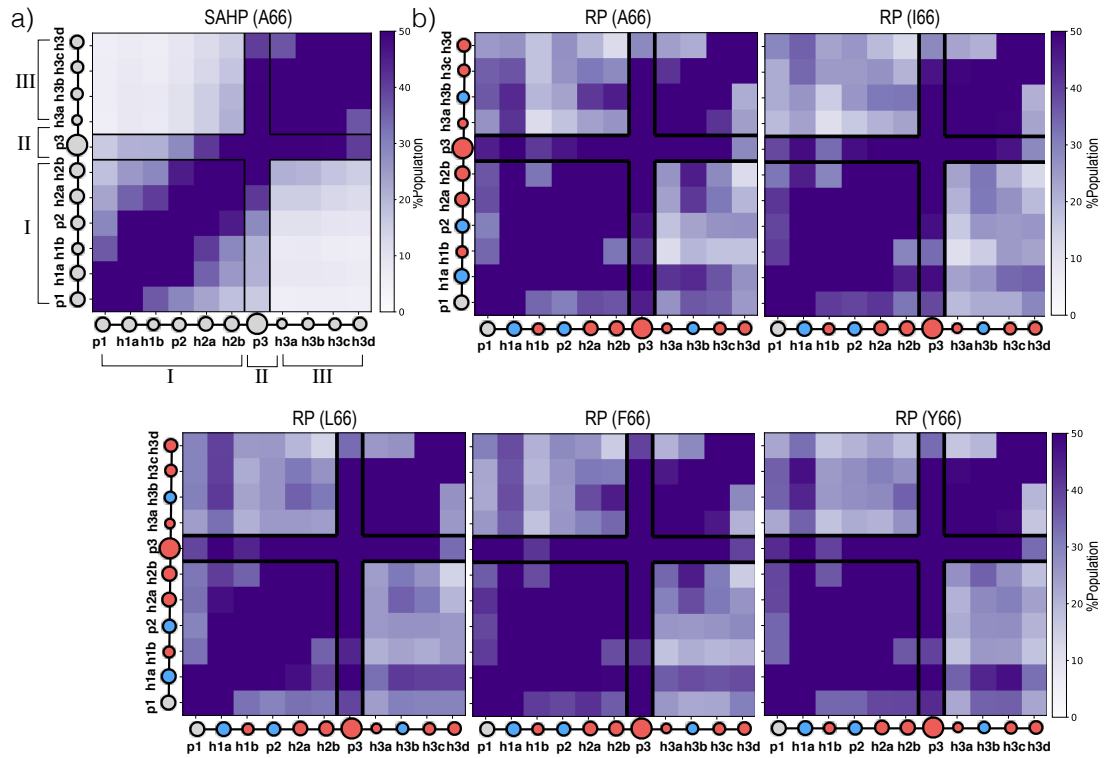

Figure 2.3: **Comparing inter-blob conatcts.** a) Blob-blob contact probability for the A66 self-avoiding heteropolymer (SAHP). The black boxes mark the regions identified. b) Blob-blob contact probability shown in a) for the real protein (RP) for every sequence simulated. The x and y axes are annotated with cartoon representation of the prodomain; circles are drawn to the scale of each blob's size and colored according to their NCPR.

Contacts between pre-linker Region I and post-linker Region III are about three times as common in the RP as in the SAHP, indicating specific tertiary interactions beyond those expected for a polymer undergoing a random-walk. We find consistently that in every other sequence of RP simulated the interactions between region I-III are consistently enriched relative to SAHP. These interactions are also sensitive to mutations as observed earlier. Interestingly we find that the enrichment in M66 RP is significantly higher than every other mutation. The enrichment in all other mutations are within the error and the order is the

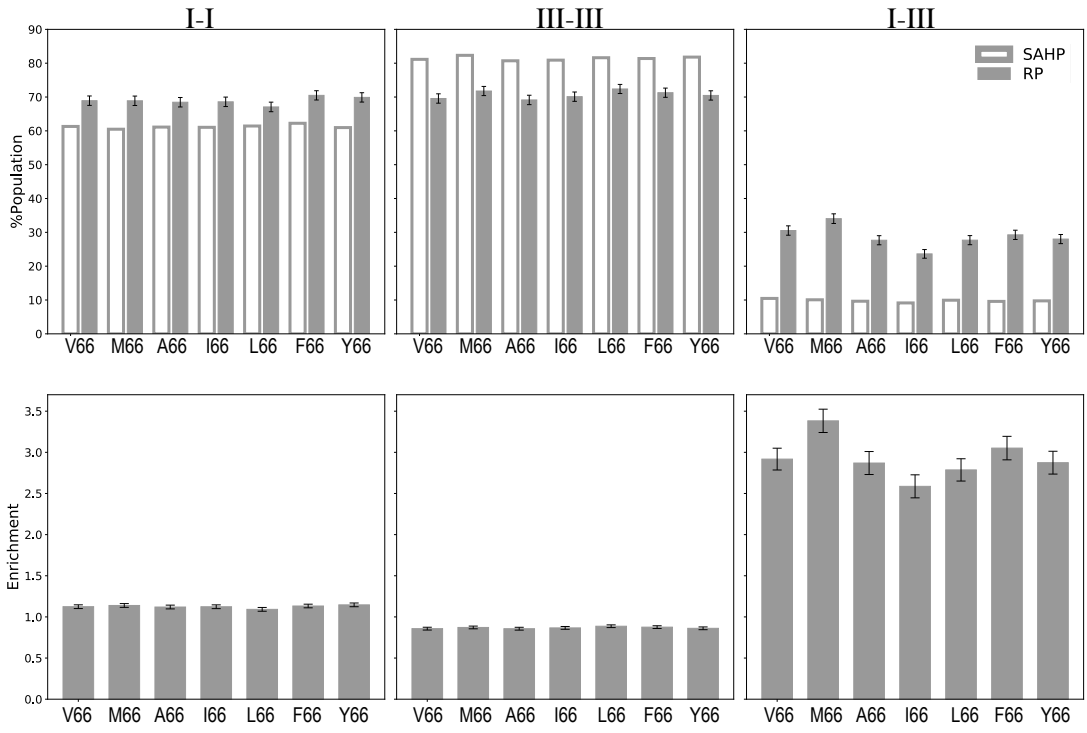

Figure 2.4: **Comparing tertiary enrichment in all 7 hydrophobic mutation sequence simulated.** %Population of contacts (top) and enrichment in RP contacts with respect to SAHP (bottom) for each region. The errors represent standard errors as described in Methods).

following: Met66> F66 > V66 > A66 = Y66 > L66 >I66 (Fig 2.4).

We find that for the interactions between region I and region III, the SNP containing blob h2b in the region I always forms the most frequent contact with the only positively charged blob h3b in region III (Fig 2.3).

Since the M66 sequence has the highest enrichment we compared the blob-blob contacts in each sequence relative to the M66 sequence (Fig 2.5). We find that the blob h2b in the M66 sequence always forms more frequent contacts with group h3 relative to every other sequence simulated. To further understand the origin of frequent contact between blob h2b and group h3 we compared the residue level contacts between h2b and group h3.

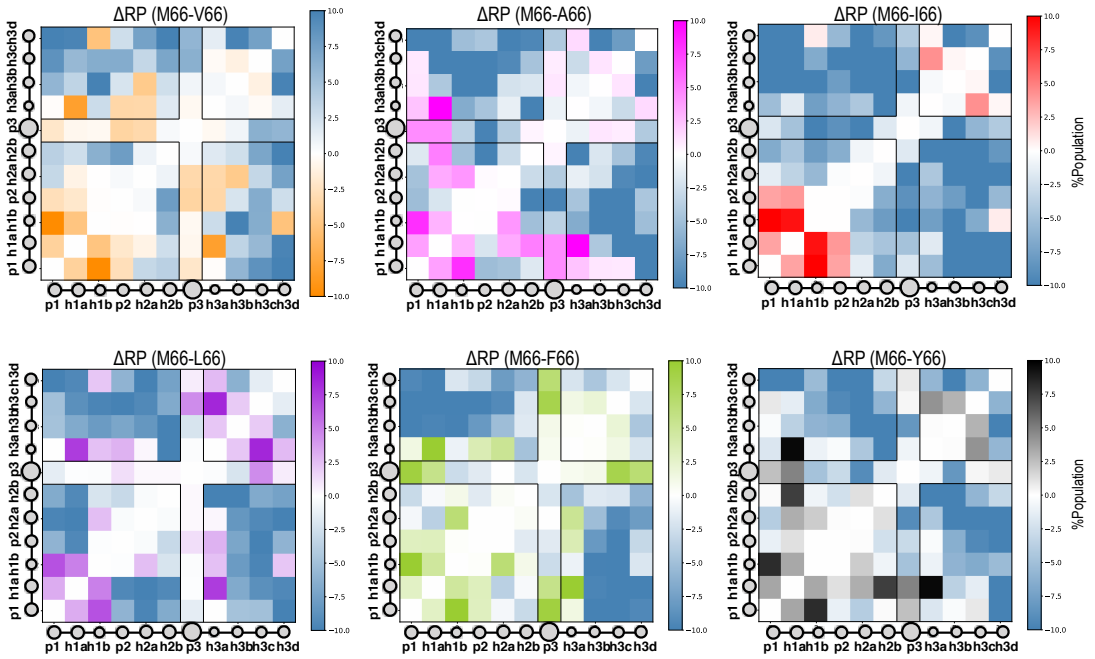

Figure 2.5: **Comparing inter-blob contact maps relative to M66 sequence.** Difference between the contact probabilities shown in Fig 2.3 relative to M66 sequence.

Residue-specific interactions at residue 66 and group h3

To further understand the origin of the significant tertiary enrichment observed in region I-III only in M66 sequence, we further zoomed into residue level contacts between blob h2b and group h3 in every sequence simulated (Fig 2.6). We find that residue 66 itself forms frequent interactions with group h3 in M66 sequence followed by F66 and V66 (Fig 2.6b). In the remaining sequences, residue 66 doesn't form any preferred contact. As found in Chapter 1, the origin of frequent M66 h3 contact is the preferred Met-Met interactions between Met66 and Met95. Following M66, the L66 sequence forms the next largest fraction of preferred contact with M95 (Fig 2.6c). This is consistent with the ab initio calculations by Gómez-Tamayo et al (Gómez-Tamayo et al. 2016), which showed that Met-Met interactions are stronger than Met-Leu or Met-Phe interactions.

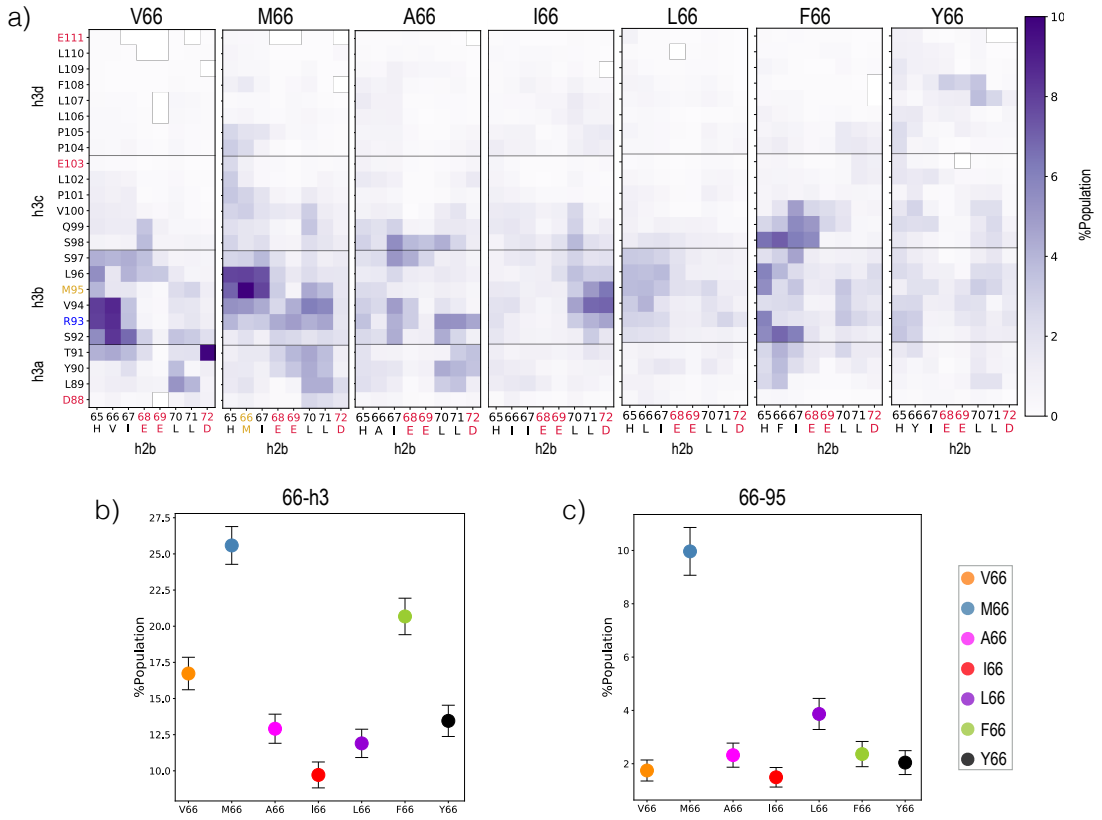

Figure 2.6: **Residues forming the cross-boundary h2b-h3 contact.** a) Contact probability at each residue in h2b with each residue in h3 for each sequence. b) and c) are same as a) but are only for residue 66-h3 contact (b) and residue 66 - Met95 contact (c).

To further understand the contribution of each amino acid at residue 66 and the high enrichment observed in RP relative to SAHP, we looked at sidechain-sidechain and backbone-backbone contacts between blob h2b and group h3. We find that following Met66, the two aromatic residues Phe66 and Tyr66 form frequent side-chain interactions with h3 (Fig 2.7). The role of aromatic-aromatic, cation-π and amino-π interactions in structural stabilization of proteins is well studied (Sundaralingam et al. 1985). We find frequent contacts between Tyr66-Phe108 and Phe66-Tyr90.

For backbone contacts we find the V66 sequence forms frequent backbone

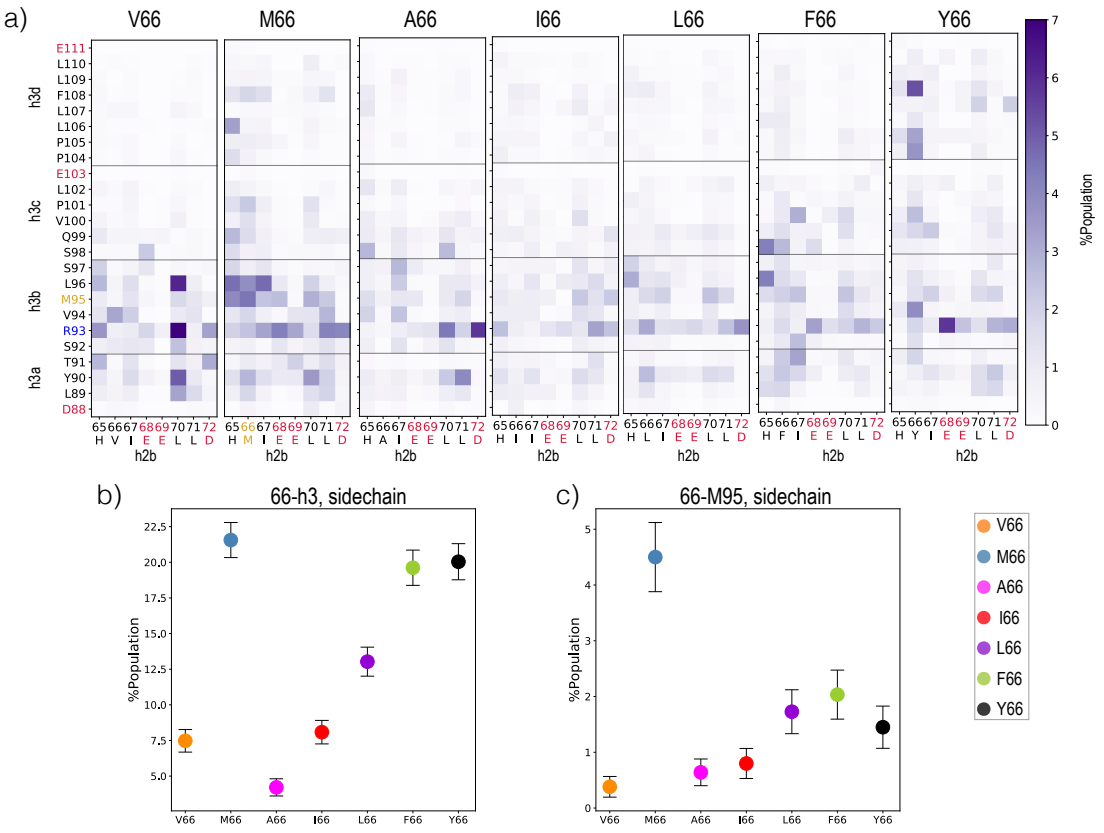

Figure 2.7: **Residues forming the cross-boundary h2b-h3 sidechain contact.** Same as Fig 2.6 but only for sidechain contacts.

contacts with group h3, followed by the F66 sequence (Fig 2.8). We earlier found that the frequent backbone contact at Val66-Ser92 is also stabilized by β - β pairing between h2b-h3b in the V66 sequence.

We further tested β - β pairing between 63-69-h3ab blobs in all the sequences. Apart from the V66 sequence no other form significant β - β pair (Fig 2.9). This is also consistent with the observation that only the V66 sequence forms significantly higher β at h3ab (Fig 2.1) and no significant differences are observed in the β propensity at h3ab blob for any other sequence simulated.

To summarize, we find the two sequences found in humans (V66 and M66) form the most frequent residue specific interactions between residue 66 and h3 group. Only V66 frequently forms β-pairs between group h3 and residue 66,

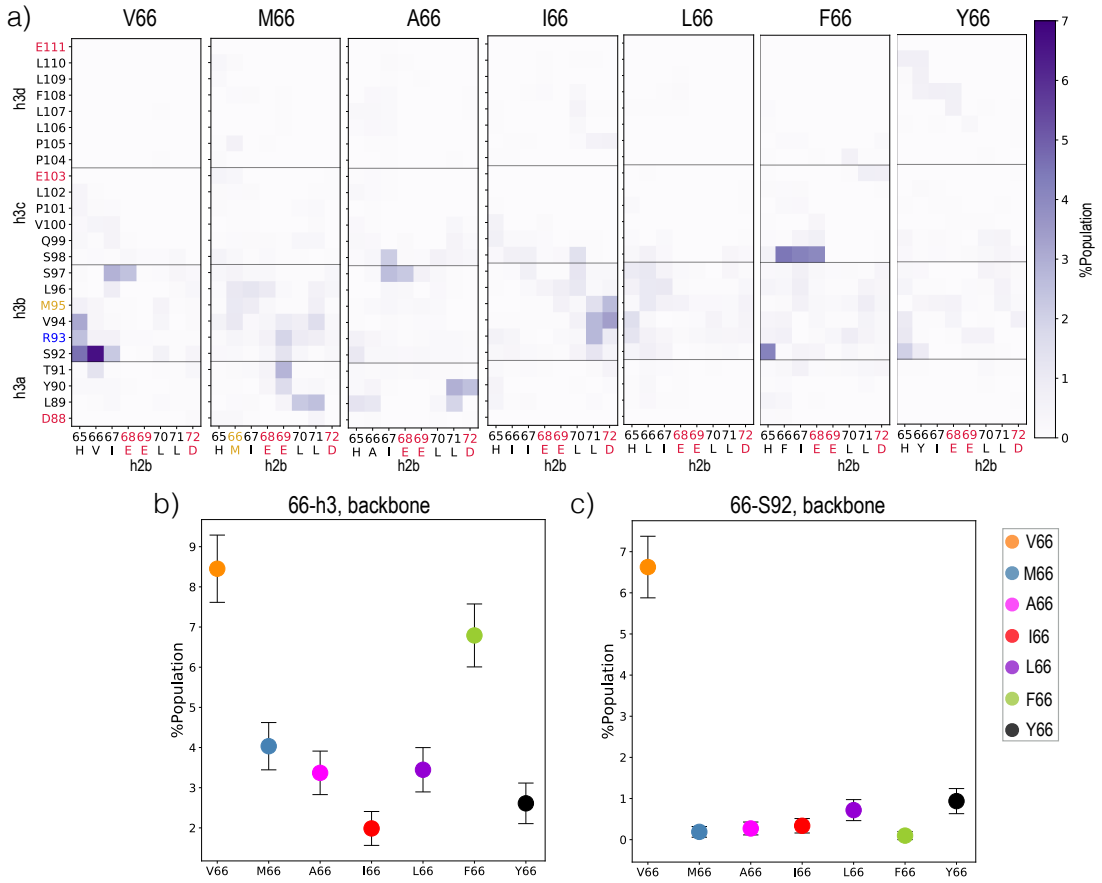

Figure 2.8: **Residues forming the cross-boundary h2b-h3 backbone contact.** Same as Fig 2.6 but only for backbone contacts.

indicating a role for side-chain specific interactions even within IDPs.

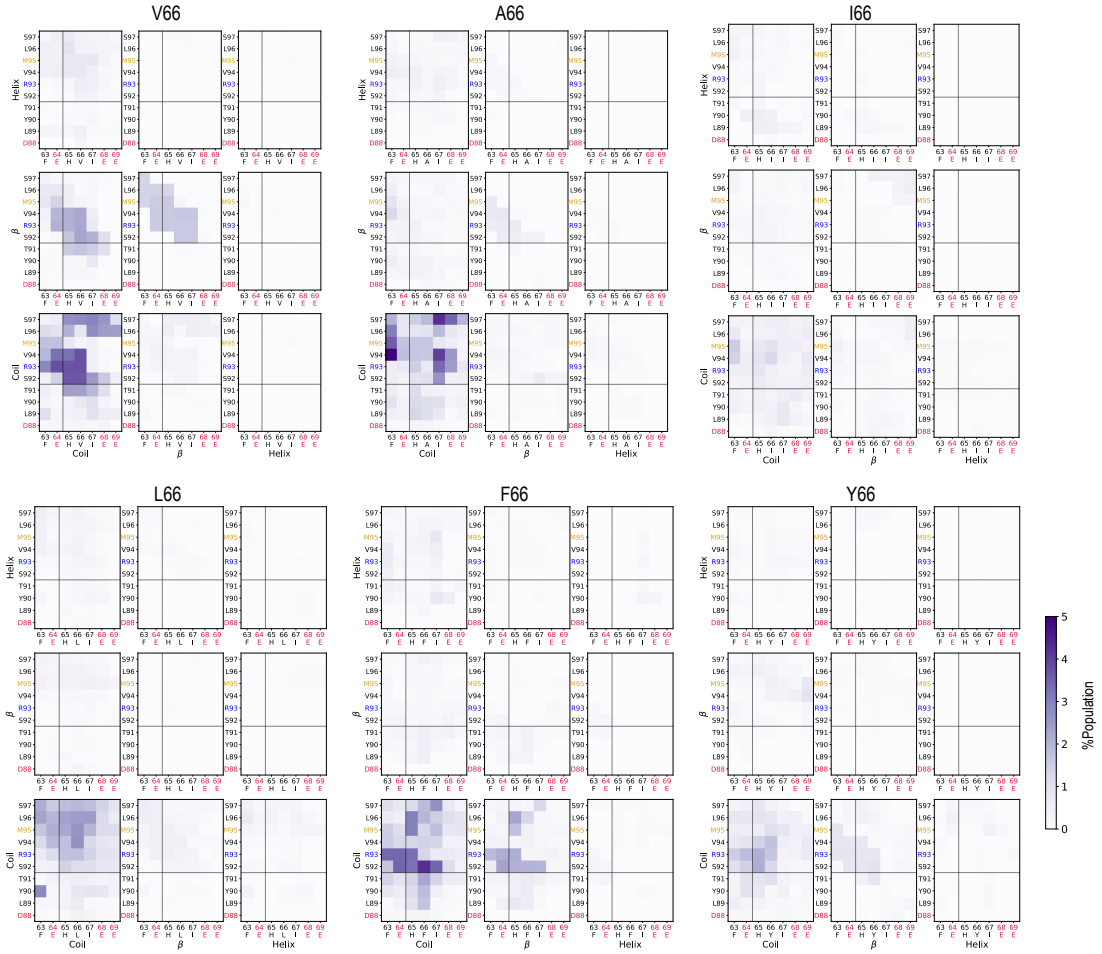

Figure 2.9: β-β **pairing at residue level.** Contact probability between residues 63-69 and each residue in h3ab, when respective secondary structure is formed at each residue, for all the sequences.

2.2.2 Comparing the effect of protonating histidine at residue 65 in V66 and M66 sequence on BDNF prodomain ensemble

His65$^+$ BDNF prodomain sequence phase annotation

Protonation at residue His65 moves the entire blob h2b from a strong polyelectrolyte region with low κ to a strong polyampholyte region with high κ (Fig 2.10, Table 2.1).

Table 2.1: **Comparing the sequence based properties of h2b^{65+} and h2b blobs identified in the BDNF prodomain, as shown in Fig 2.10.**

Blob	N[a]	NCPR[b]	$\langle H \rangle$[c]	FCR[d]	f-[e]	f+[f]	κ[g]	Sequence	R[h]	P[i]
h2b^{65+}(V66)	8	-0.25	0.54	0.50	0.38	0.13	0.6	H$^+$VIEELLD	3	0.00
h2b^{65+}(M66)	8	-0.25	0.50	0.50	0.38	0.13	0.6	H$^+$MIEELLD	3	0.00
h2b(V66)	8	-0.38	0.54	0.38	0.38	0.00	0.3	HVIEELLD	4	0.00
h2b(M66)	8	-0.38	0.50	0.38	0.38	0.00	0.3	HMIEELLD	4	0.00

[a] Number of residues in the blob

[b] Net charge per residue

[c] Mean hydrophobicity, average of Kyte-Dolittle (Kyte & Doolittle 1982) scores for each residue in the blob scaled to fit between 0 and 1

[d] Fraction of charged residues

[e] Fraction of positively charged residues

[f] Fraction of negatively charged residues

[g] Charge distribution parameter κ as defined by Das and Pappu (Das & Pappu 2013), calculated using CIDER (Holehouse et al. 2017)

[h] Region in phase diagram proposed by Das and Pappu (Das & Pappu 2013), (Fig 2.10a)

[i] Fraction of Proline residues

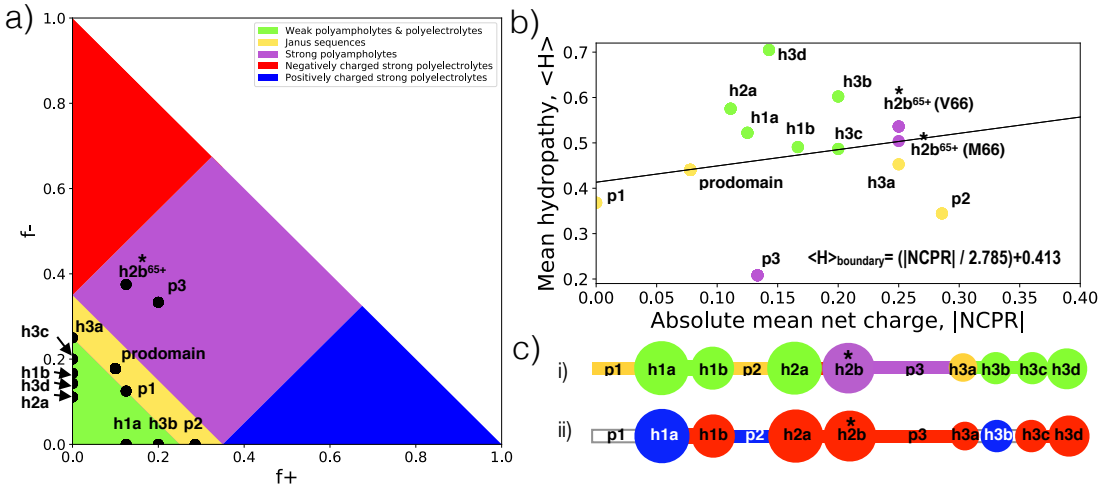

Figure 2.10: **Sequence-based decomposition of the BDNF prodomain with protonated His65.** All the panels are same as Fig 1.2 where blob h2b is replaced with protonated His65 containing blob h2b^{65+}. Additional properties of the h2b^{65+} blob can be found in Table 2.1.

Effect of His65 protonation on local and non-local secondary structure

When protonated at His65, the helix propensity in group h2 further increased only in M66^{65+}. M66^{65+} has 2 fold helix propensity within h2 group relative to M66. This is consistent with previous observations. Using a meta-structure analysis approach Geist et al. (Geist et al. 2013) showed that IDPs have increased tendency of forming α-helical secondary structure elements when protonated due to reduced electrostatic repulsion in the otherwise negatively charged protein. Helix propensity decreases at blob h1b, h3a and h3d in M66^{65+} relative to M66. Val66^{65+} does not show any significant change in helix propensity at any residue relative to V66.

The β propensity decreases for both V66^{65+} and M66^{65+} sequence when compared with their neutral H65 sequence. V66^{65+} still has increased β propensity in the h3ab blob when compared with M66^{65+} or M66.

Protonation increases the intra-blob contacts within h2 Fig 2.11b. When

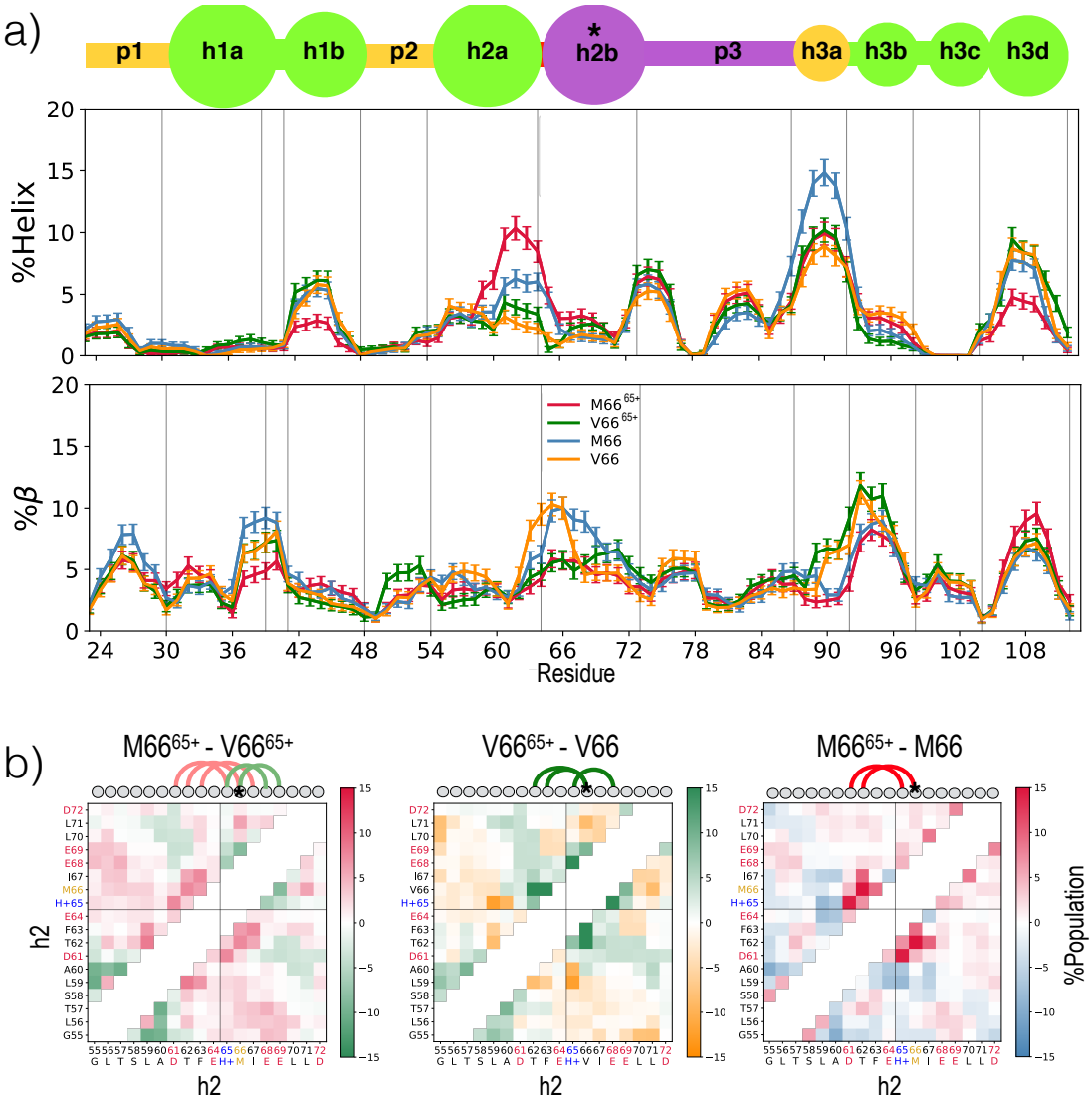

Figure 2.11: **Effect of histidine protonation on secondary structure.** Helix (top) or β (bottom) propensity for each simulated residue in V66, M66, V66^{65+} and M66^{65+} sequence at the 300K replica, defined as the probability of a given residue being part of a sequence of four or more consecutive residues whose dihedral angles place them in the helical (left) region or β (right) region of the Ramachandran map (further described in methods). b) Contact probability for each residue pair within the h2 group for each sequence.

protonated, both V66 and M66 sequences gain contacts with His65, Met/Val66, and Ile67. These increases in intra-blob contact frequency in h2b^{65+} are consistent

with expectations, since protonation at H65 reduces the net charge and increases κ of the otherwise negatively charged polyelectrolyte blob with low κ (Table 2.1).

Regions of tertiary enrichment

In Chapter 1 we proposed the tertiary enrichment test to measure changes in tertiary contacts in IDPs. We find that intermonomer contact frequency in a SAHP was a useful reference for detecting specific tertiary interactions, as long as the monomers mimic the blobs of the real protein (RP).

Fig 2.12 shows the probability of blob-blob contacts for $\mathrm{Val66}^{65+}$ and $\mathrm{Met66}^{65+}$ sequences of the RP, calculated analogously to those in the SAHP.

We find consistent segmentation of contact maps into regions at p3 boundary for the protonated sequence simulation as well. The frequencies of contacts within Region I and within Region III were quantitatively consistent with the SAHP predictions.

Contacts between pre-linker Region I and post-linker Region III are about three times as common in the RP as in the SAHP, indicating specific tertiary interactions beyond those expected for a polymer undergoing a random-walk. We find consistently that in every other sequence of RP simulated the interactions between region I-III is consistently enriched relative to SAHP. However, these interactions are no longer very sensitive to residue at 66. We find no significant differences in the enrichment of these contacts for the two protonated sequences.

We further looked at the differences between blob contact maps of the two protonated sequences. We find that blob h2b forms frequent contact with blob h3ab and h3cd in $\mathrm{V66}^{65+}$ and $\mathrm{M66}^{65+}$ respectively.

When we compared the differences between the protonated and neutral His65 sequences for V66 and M66, we find that due to protonation both the sequences lose contacts between blob h2b and h3a.

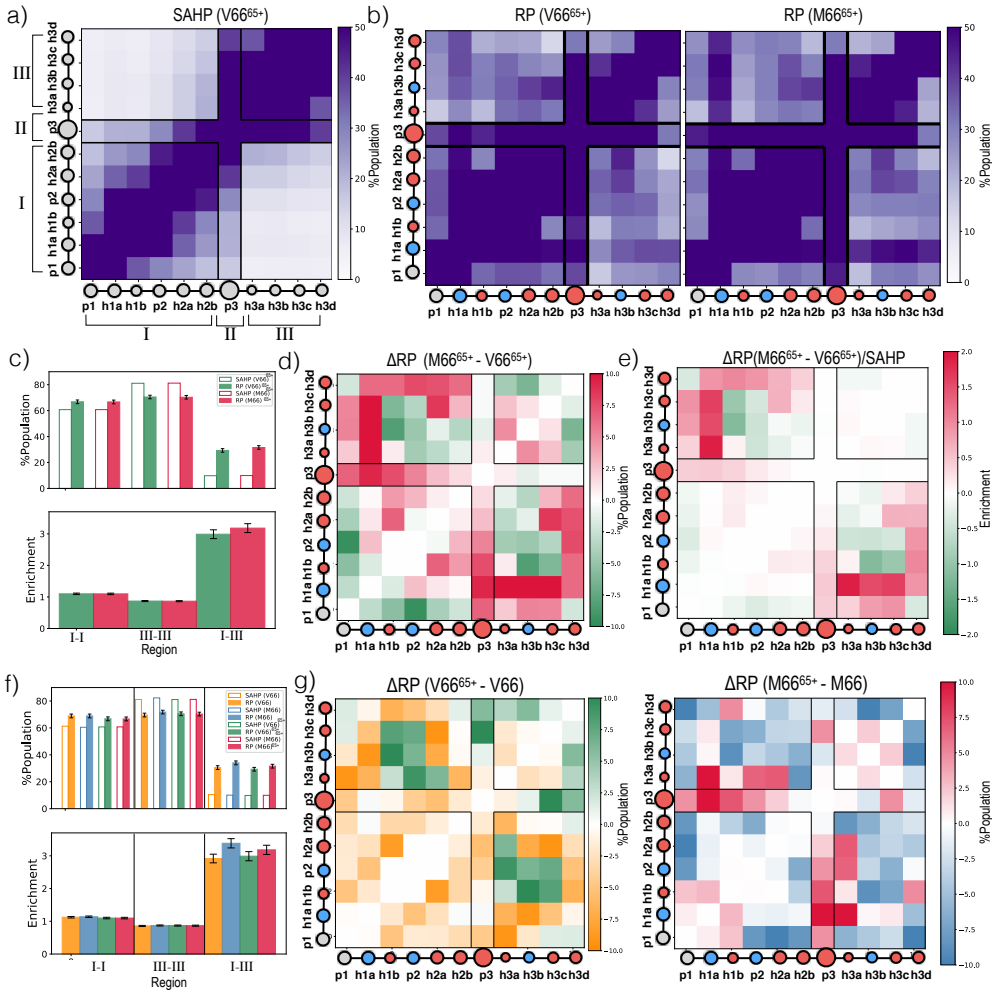

Figure 2.12: **Effect of His65 protonation on contacts between blobs.** a) Blob-blob contact probability for the V66^{65+} self-avoiding heteropolymer (SAHP). The black boxes mark the regions identified. b) Blob-blob contact probability shown in a) for the real protein (RP); V66^{65+} (left) and M66^{65+} (right) sequences. c) %Population of contacts (top) and enrichment in RP contacts with respect to SAHP (bottom) for each region. d) Difference between the contact probabilities shown in b). e) Differences shown in d) with respect to SAHP. f) Same as c) but also includes V66 and M66 sequence. g) Difference between the contact probabilities of protonated and neutral His65 sequence for V66 (left) and M66 (right).

Residue-specific interactions at residue 66 and group h3

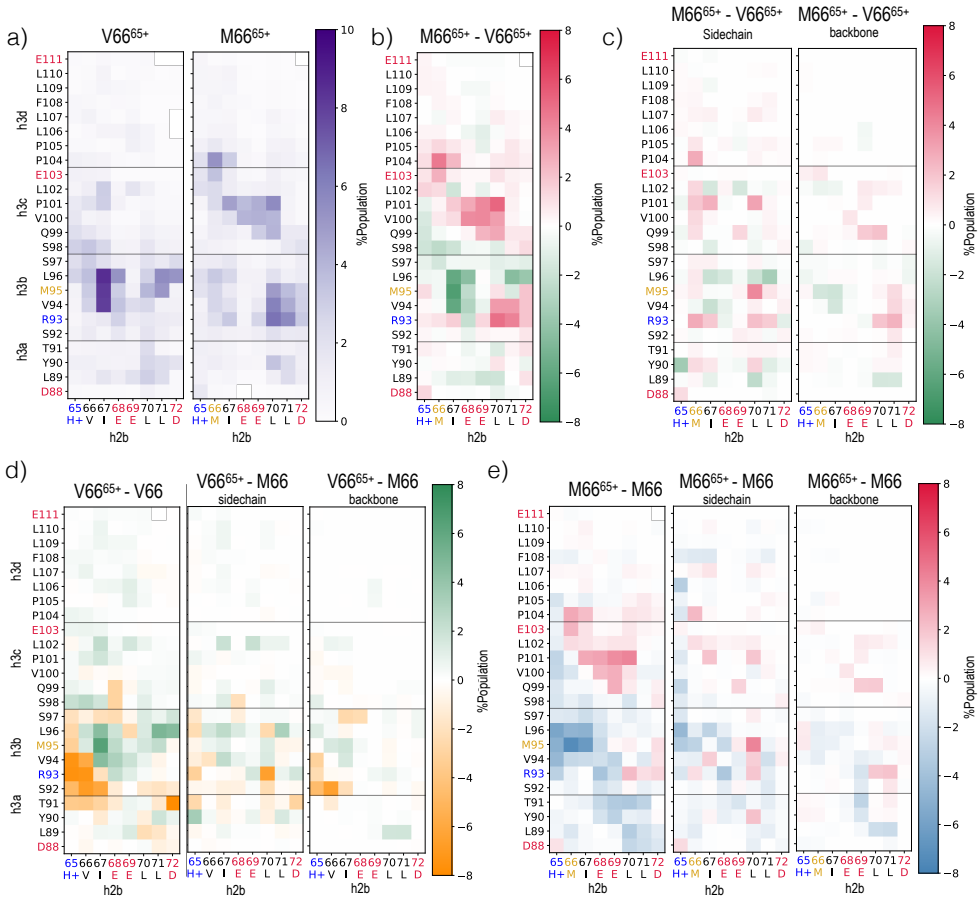

Figure 2.13: **Residues forming the cross-boundary h2b-h3 contact.** a) Contact probability at each residue in h2b with each residue in h3 for $V66^{65+}$ (left) and $M66^{65+}$ (right). b) Difference between the contact probability shown in a). c) Same as b) but only for sidechain (left) and backbone (right) contacts. Difference between the protonated and neutral His65 sequences for V66 (d) and M66 (e)

We consider here the effects of the substitution and protonation on the h2b-h3 blob contacts at the residue level. $V66^{65+}$ forms frequent contacts at residue Ile67 and hydrophobic residues Val94, Met95 and Leu66. $M66^{65+}$ does not form any preferred residue specific contacts, instead, it forms several contacts between blob h2b and group h3.

We further zoomed into the contact maps at backbone and sidechain contacts at the residue level. We find that Met66 forms several non-specific sidechain contacts with several residues in group h3 when compared to Val66. Val66 instead forms weak backbone contacts with Met95.

Next, we compared the residue level contacts between the protonated and neutral His65 sequences. We find that the h2b^{65+} blob loses several residue level contacts with h3b blob and instead gains few contacts with h3c blob in both sequences. This change due to protonation can be explained with electrostatics. The newly positively charged residue His65 loses contacts with the positively charged blob h3b and instead shifts contacts to the negatively charged blob h3c. Val66 loses the frequent β pairing with Ser92. As we found in Chapter 1, these pairings were also stabilized with Glu64-Arg93 salt bridge. The presence of now protonated His65 breaks these β pairs. Instead, Val66 forms frequent backbone contacts with Met95.

2.2.3 Comparing blob and chain properties from all 9 simulations

While determining intermolecular interactions that cause compaction in the disordered domain of poly(A)-binding protein using small-angle X-ray scattering (SAXS), Riback et al. (Riback et al. 2017) reported that increases in net hydrophobicity resulted in more compact proteins. This trend held across large increases in hydrophobicity but was not predictive for more subtle changes, such as a single residue replacement of Valine with Methionine. We looked at the R_g for each sequence simulated vs the Kyte-Dolittle (Kyte & Doolittle 1982) hydropathy score. We find no obvious trends with a change in hydropathy. Instead, we find that M66, F66, and V66 form slightly compact ensembles whereas the remaining sequences forms relatively expanded ensembles. The compact conformations in

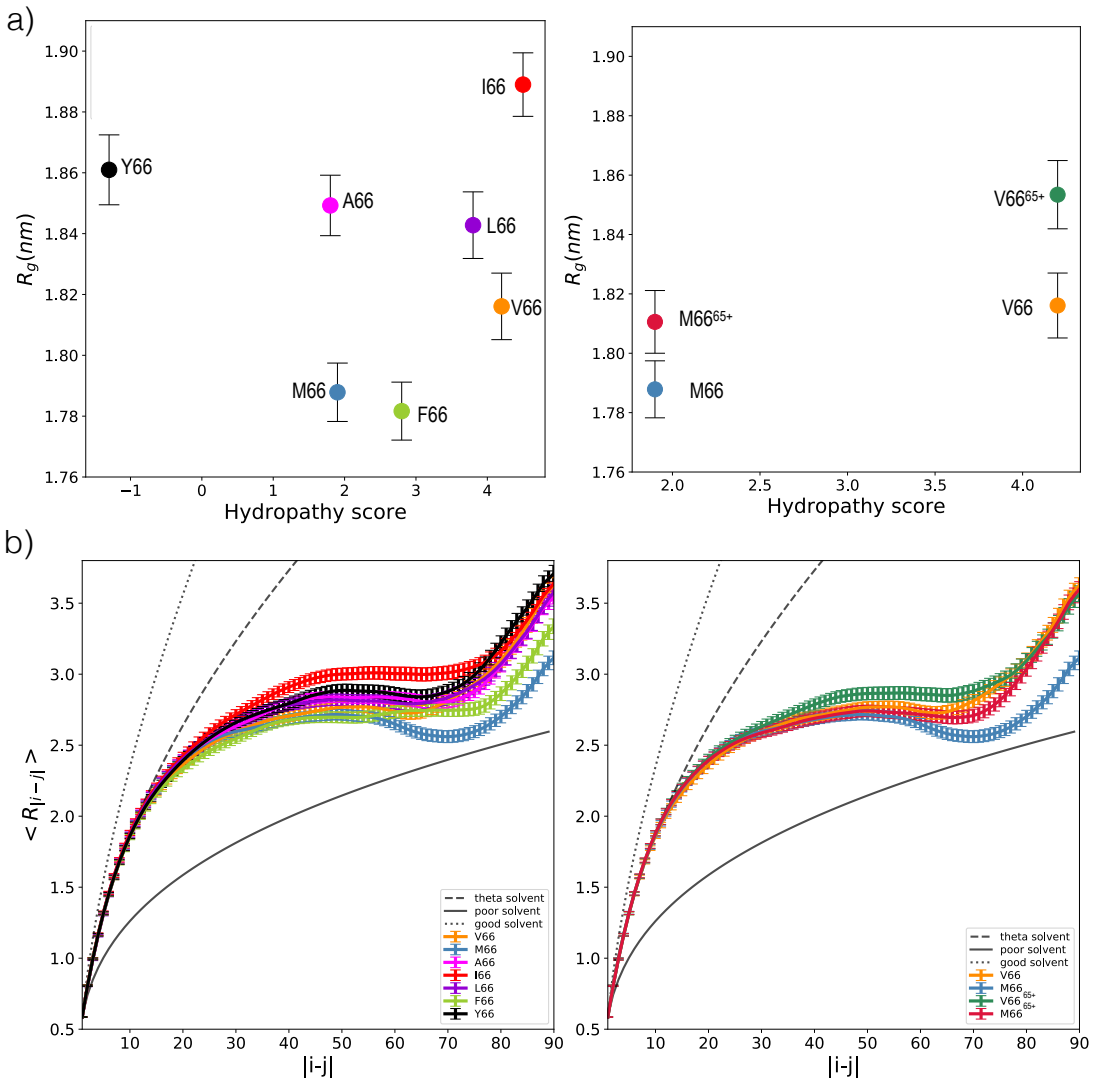

Figure 2.14: **Average R_g and scaling behavior of each simulated sequence.** a) Average R_g for all 7 neutral His65 sequence simulated (left) and protonated His65 sequences (right). b) Ensemble averaged interchain distance profiles for all 7 neutral His65 sequence simulated (left) and protonated His65 sequences (right). Theoretical polymer scaling limits are shown with grey lines (prefactor A = 0.59 nm)

M66, F66, and V66 are consistent with the preferred residue specific interactions formed in these sequences at residue 66.

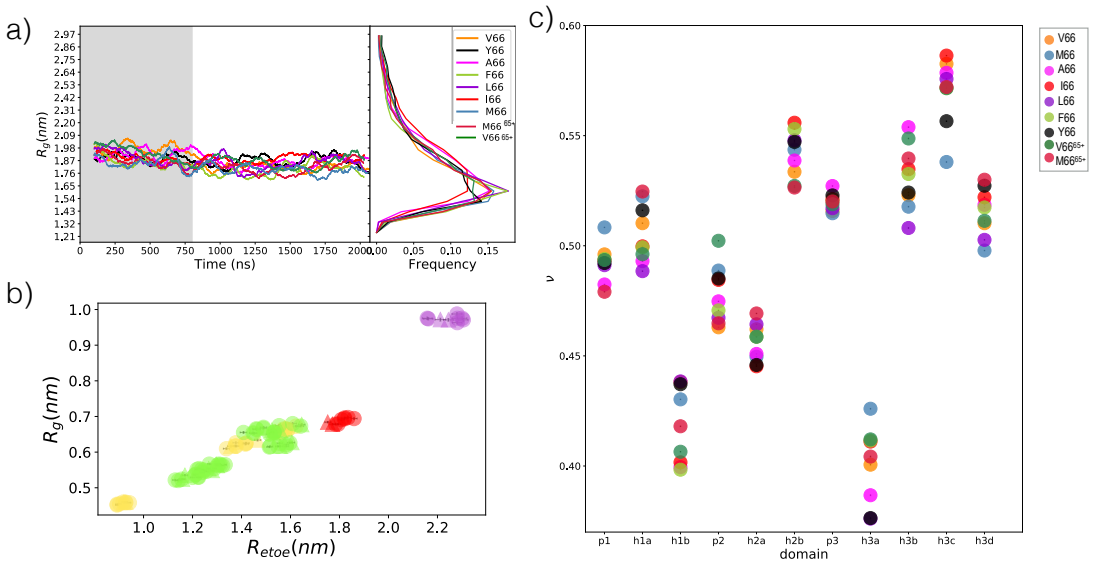

Figure 2.15: **Simulation convergence and scaling behavior of each blob in each sequence simulated.** a) R_g at 300K vs the simulation time (left), using a 100 ns moving window and distribution of R_g (right) for all 9 sequences simulated. b) Average end to end distance (R_{etoe}) vs Average R_g of each blob for all 9 sequences simulated. Flory scaling exponents (ν) of each blob for all 9 sequences simulated.

We further looked at the effect of protonation at His65. Protonation consistently increases the R_g for both V66 and M66 sequence. This can be explained due to the loss of contacts between His65^{+} and the h3b blob.

We further compared the blob shape, size and ν for each blob in all 9 simulations. Each blob populates similar R_{etoe} vs R_g across all 9 simulations. Within each blob, the protein obeys Flory polymer scaling laws. However, the scaling exponent ν varies slightly within each blob across all 9 simulations. The blob h2b^{65+} is slightly more compact when compared with neutral His65 sequence simulations, thus the observed ν is also smaller for these blobs relative to h2b.

2.3 Discussion

IDPs are extremely sensitive to changes in their environment. Post-translational modifications (PTMs), changes in temperature, pH, ion concentration presence of binding partners can modify the sequence ensemble relationship of IDPs (Uversky 2009; Darling & Uversky 2018).

Additionally, the Val66Met substitution is present in a region with a high density of negatively charged residues (D61, E64, E68, E69). In this scenario, residue H65 can exist in protonated or neutral charge state in vivo due to its moderate pKa.

Studying the effect of $Val66^{65+}$ and $M66^{65+}$ sequence, gives us further insight into the sensitivity of IDPs to their environment. Furthermore, comparison of the protonated vs neutral His65 sequences for both V66 and M66 sequence, allows us to compare the effects of reduced pH and view them in the context of charge-neutral mutations.

Both the neutral and charged His65 prodomain lie at the boundary region of the Das and Pappu diagram. However, $His65^{+}$ shifts the h2b blob from a strong polyelectrolyte region of the Das and Pappu phase diagram to a strong polyampholyte region. Since the prodomian is negatively charged, one would predict that adding a positive charge would collapse the sequence. However, we find that $His65^{+}$ only collapses the h2b blob, and it further expands the protein due to loss of h2b-h3b contacts. This emphasizes that the blob within which the charge is added or removed matters and our blob topology framework is very sensitive to a single charge residue addition or deletion.

Analyzing and comparing the residue level insight from all 9 simulations helped us in further establishing the significance of the tertiary enrichment test in IDPs. We find that all the substitutions exhibit nearly identical polymer properties; effective real protein (RP) segmentation was observed in all the simulations,

but the effective enrichment in contacts across the segmentation is specific to the amino acid at residue 66. Furthermore, we find that for the interactions between region I and region III, the SNP-containing blob (h2b in region I) always forms the most frequent contact with the only positively charged blob h3b (in region III). Protonation at His65 weakens the interaction between these two blobs.

Consistent with our predictions in Chapter 1, we find that the Met66 sequence has increased intra-protein contacts due to preferred Met-Met interactions. Protonation increases intra-blob contacts across the protein but reduces inter-blob contacts, thus, slightly lowering pH can mediate the interaction between the SNP blob h2b and SorCS2 surface.

Chapter 3

Disease associated mutations in intrinsically disordered proteins: evidence of genome-wide enrichment in hydrophobic domains

Abstract

The consequence of an amino acid substitution in disordered proteins is difficult to predict. In structured proteins, a single amino acid substitution can affect its function by causing a change in protein structure. In disordered proteins, how mutations disrupt protein function or ensemble is not well understood. In the previous study, we developed a novel hierarchical sequence-based framework for analysis and conceptualization of long intrinsically disordered proteins (IDPs). This sequence-based blob decomposition framework was able to predict enrichment of higher-order (tertiary) structure in a disordered protein due to a disease-causing SNP. In the current study, we extend the sequence classification analysis to 11,752 proteins and 65,291 SNPs, testing for enrichment of disease-associated SNPs within IDP subdomains. We find several properties of the blobs which are enriched in disease-associated SNPs relative to non disease-associated SNPs. Finally, we developed a web tool for interactive identification of sequence topology within disordered regions for any given protein sequence. This work represents the first systematic, bottom-up, attempt to both identify and annotate subdomains within disordered proteins that are enriched for functional effects.

3.1 Introduction

With the era of cost-effective genome sequencing, a plethora of information regarding the complete genome has been available (Glusman 2013; Metzker 2010). Millions of SNPs from the general population is deposited at Exome Aggregation Consortium (ExAC) (Lek et al. 2016) and in the Genome Aggregation Database (gnomAD) (Karczewski et al. 2019). Clinical relevance information for a large number of these SNPs is also available in several databases including ClinVar (Landrum et al. 2018) and OMIM (McKusick 2007) among others. A missense SNP changes an amino-acid in the protein sequence. These SNPs could be either associated with disease (DA), non disease-associated (NDA), or remain unclassified depending on whether they are implicated in diseases or not according to literature reports in UniProtKB (Yip et al. 2008). NDA SNP is also used to describe rare SNPs as well as polymorphisms that have an effect on protein function, but with no resulting clinical phenotype (Yip et al. 2008). A large number of SNPs are still unclassified since their effect on the protein function is not yet known. Mapping genetic variation within protein coding sequences to effects on protein function remains an active area of genomics and proteomics research.

Several features of SNPs have been identified which can predict its likelihood to be associated with diseases or vice-versa. A large number of these features rely on the solved protein structures (Iqbal et al. 2019) and evolutionary conservation of residues (Ng & Henikoff 2003; Kircher et al. 2014). It has been found that DA SNPs frequently disrupt the structural property of a protein and thereby affect its function. However, these structural and sequence conservation features do not work well for IDPs, which lack a unique 3D structure and have low sequence conservation. More than 25% of DA SNPs are found in IDPs. They are correlated with specific diseases, for example, more than 79% of cancer-related

proteins contain disordered regions (Iakoucheva et al. 2002). Due to their signaling and regulatory roles, IDPs tend to be tightly regulated, and disruptions in their regulation have been linked to disease (Darling & Uversky 2018).

Quantification of the occurrence of SNPs and post translational modifications (PTMs) in IDPs is a relatively new field (Vacic et al. 2012; Lu et al. 2015). Few studies have been done to characterize the effect of SNPs or PTMs on the disordered protein ensemble (Firman & Ghosh 2018). Even though disordered proteins are depleted in hydrophobic residues, it has been observed that small hydrophobic motifs in disordered regions are involved in the binding of these proteins with their partners (Mohan et al. 2006). PTMs such as Ser/Thr phosphorylation can change the FCR and NCPR properties of IDPs and can thus tune the sequence ensemble relationships of IDPs due to their polyampholytic nature (Firman & Ghosh 2018; Das & Pappu 2013). SNPs in IDPs can cause disorder $->$ order transitions (Vacic et al. 2012; Uversky et al. 2014). Moreover, DA SNPs can disrupt molecular recognition features (MoRFs) in disordered proteins. Because MoRFs mediate protein-protein interactions, it follows that protein-protein interaction networks may also be disrupted (Uversky et al. 2014).

In the current study, we investigate the correlation between DA SNPs and IDP topology, where the topology of subdomains is predicated based on sequence hydrophobicity. Using our hierarchical sequence-based framework, we identified "blobs" representing local globular regions (h) or linkers (p) within the sequence of a given protein. The h blobs are stretches of four or more residues with comparatively higher hydrophobicity. The remaining residues are classified as p blobs. We compare several blob properties of 29,230 disease-associated SNPs with 36,060 non disease-associated SNPs.

3.2 Results

3.2.1 Disease associated SNPs are enriched in h blobs and depleted in p blobs

A total of 65,291 SNPs were analyzed for disease associated enrichment from 11,752 genes. Unless otherwise noted, DA SNPs are tested for enrichment relative to the expectation set by NDA SNPs. For example, the phrase "DA SNPs are enriched in region X" means that DA SNPs are found at a higher rate in X regions than are NDA SNPs. We find that about 50% of the DA SNPs are found in h blobs and the remaining in p blobs (Fig 3.8a). We further compared the enrichment of DA SNPs in h or p blobs. We find that DA SNPs are 1.15 fold enriched in h blobs. In Fig 3.8b we looked at the hydrophobicity distribution of DA and NDA SNP's neighboring residue in its sequence. We find that in general a SNP is more likely to be associated with disease if it has hydrophobic neighbors.

We further divided the SNP to be in ordered or disordered regions as described in methods. As observed previously by Vacic et al (Vacic et al. 2012), we find that DA SNPs are enriched in ordered regions (Fig 3.8c). Although DA SNPs are depleted in disordered regions, more than 40% of DA SNPs are still found in disordered regions, due to the high frequency of disordered SNP population. Since IDPs are depleted in hydrophobic residues, a higher proportion of SNPs in p blobs are identified in the disordered region when compared to ordered regions. We find that for both ordered and disordered regions, DA SNPs are enriched in h blobs and depleted in p blobs (Fig 3.8d). This was not surprising, the significance of hydrophobic regions in structured proteins are well established. The hydrophobic effect is considered to be the major driving force for the folding of globular proteins(Dill 1990). Even though disordered proteins are depleted in hydrophobic residues, it has been frequently observed that small hydrophobic motifs in disordered regions are involved in the binding of these proteins with

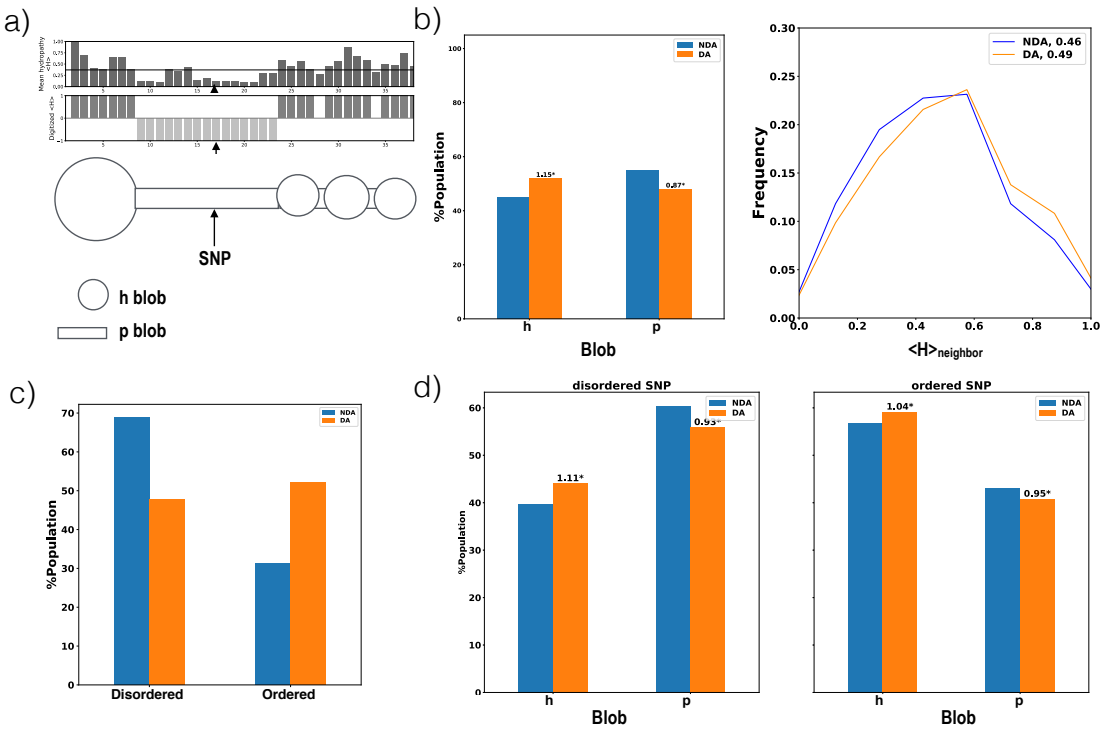

Figure 3.1: **Disease-associated SNPs are more enriched in hydrophobic (h) blobs.** a) Cartoon representation of identified h and p blobs for a given protein sequence. b) Disease-associated (DA) SNPs and non disease-associated (NDA) SNPs proportion in h and p blobs (left) and Kyte-Dolittle (Kyte & Doolittle 1982) hydrophobicity distribution of a SNP's neighboring residue in its protein sequence (right) in all of the 65,291 SNPs dataset. $H_{neighbour} = \frac{H_{i-1}+H_{i+1}}{2}$, where i is the residue index of SNP in its sequence. The mean of each histogram distribution is also reported in the caption. c) The proportion of SNP that fall in ordered and disordered regions. A large number of NDA SNPs are found in the disordered region. d) Same as b) when the dataset was divided into the ordered and disordered region. If enrichment or depletion in DA SNPs is significant (p-value $< 5 \times 10^{-5}$) it is annotated with the observed fold enrichment and a star.

their partners (Mohan et al. 2006). Interestingly, the enrichment of DA SNPs is higher in disordered regions when compared with ordered regions, 1.04 fold and 1.11 fold enrichment respectively. Many of the disordered predictors predict the

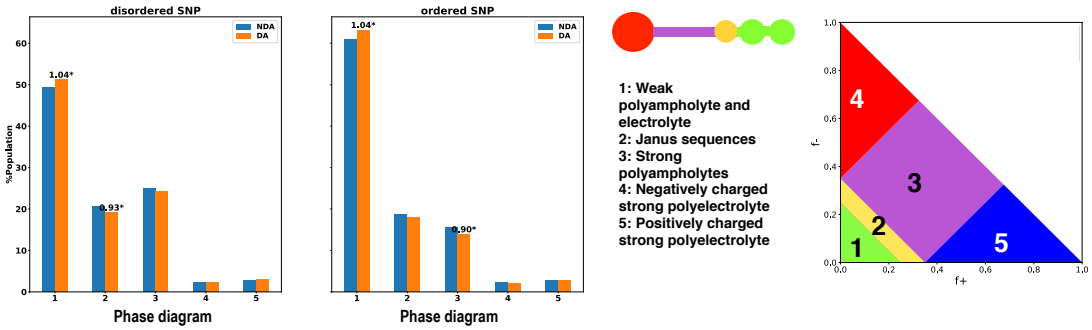

Figure 3.2: **Distribution of blob containing SNP in various regions of Das and Pappu (Das & Pappu 2013) phase diagram.** The proportion of each SNP in various regions of Das and Pappu (Das & Pappu 2013) phase diagram in ordered and disordered SNPs (left). Cartoon representation of identified h and p blobs for a given protein sequence further colored according to its phase annotation (right). If enrichment or depletion in DA SNPs is significant (p-value $< 5 \times 10^{-5}$) it is annotated with the observed fold enrichment and a star.

h blobs in proteins as a part of the ordered region. Therefore, the enrichment observed in h or p blobs is dependent on the disorder predictor used.

We further looked at the proportions of SNPs in all the 5 regions of Das and Pappu phase behavior (Fig 3.2). More than 60% of SNPs in ordered regions and 50% of SNPs in disordered regions lie in the globular region (region 1) of phase diagram. SNPs in disordered regions have a higher proportion (10% more) of polyampholytes blobs. This is consistent with the higher frequency of h blobs and p blobs in ordered and disordered regions respectively. Since h blobs are enriched in DA SNPs (Fig 3.8d), we also find slight enrichment in globular blobs (region 1) in both ordered and disordered regions (Fig 3.2).

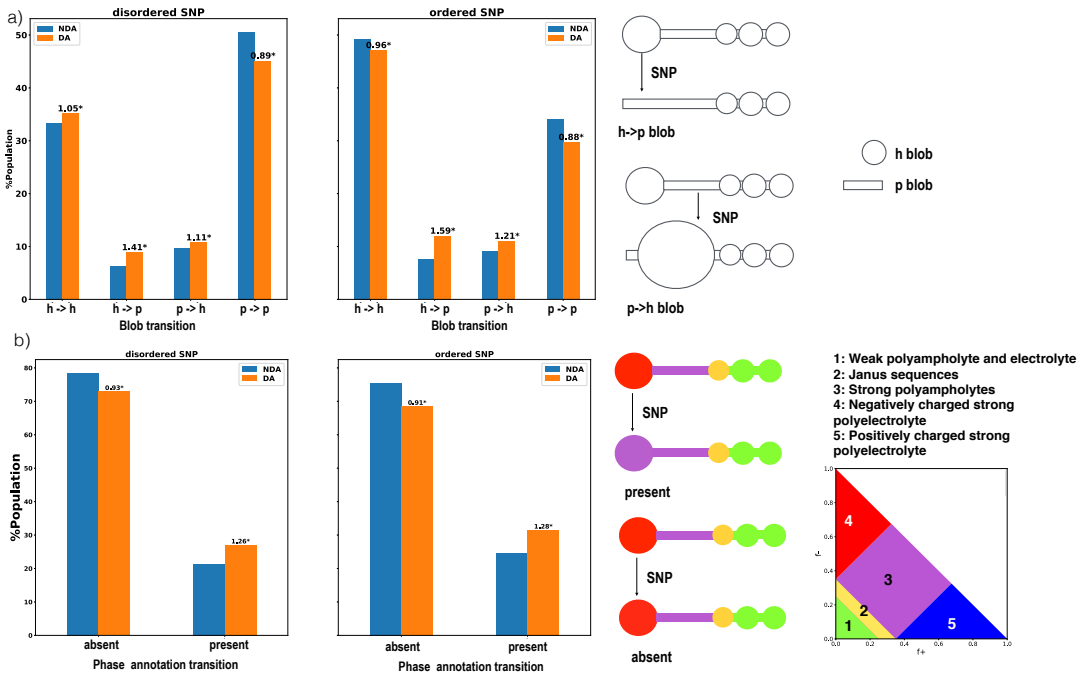

Figure 3.3: **Disease-associated SNPs are enriched in blob transitions.** Disease-associated (DA) SNPs and non disease-associated (NDA) SNPs distribution in blob transitions (a) and phase annotation transitions (b). Depending on whether a SNP changes phase annotation or not it is grouped as phase annotation transition present or absent respectively. Transitions of individual phases are shown in Fig 3.4. If enrichment or depletion in DA SNPs is significant (p-value $< 5 \times 10^{-5}$) it is annotated with the observed fold enrichment and a star.

3.2.2 Disease associated mutations could cause transitions in blob and phase annotations

To better understand how DA SNPs effect blob properties, we examined whether a SNP could change blob properties such as switching the blob from h to p and vice versa. When analyzed for blob transitions, the given mutation in a SNP can have two outcomes: (i) it can change, reduce, or increase the hydrophobicity to cross the blob cutoff score, resulting in a h $->$p blob or p $->$ h blob. ii) it does not change the blob assignment, resulting in a h $->$ h or p $->$ p blob.

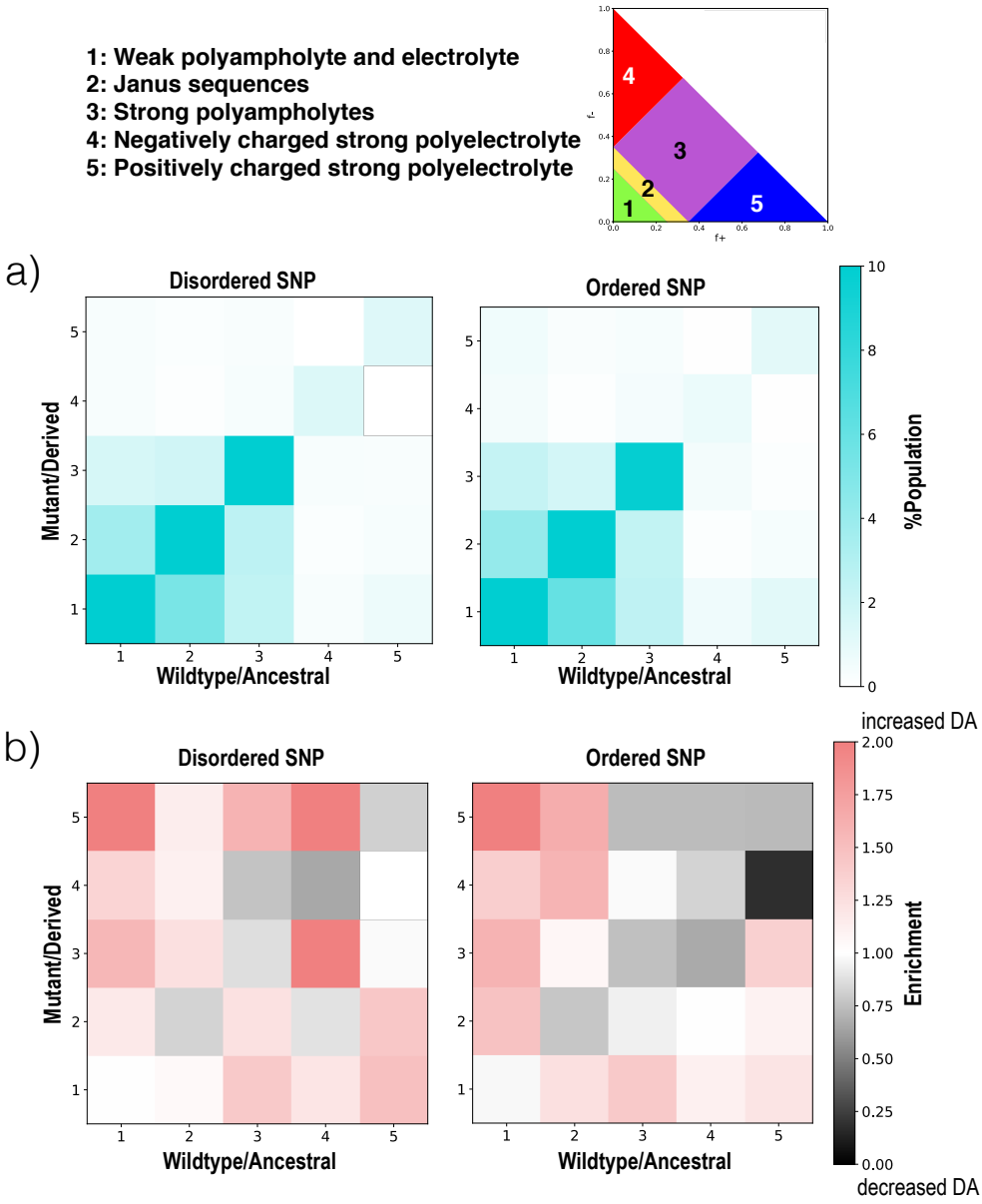

Figure 3.4: **Disease-associated SNPs are enriched in phase annotation transitions.** a) Enrichment or depletion observed in a phase annotation changes in ordered and disordered SNPs. b) Population of SNPs undergoing a given phase annotation change in ordered and disordered SNPs.

We find that irrespective of whether the SNP is in ordered or disordered regions, DA SNPs are enriched in changing h blobs to p blobs and vice versa (Fig 3.3a). In structured proteins, enrichment or depletion of disease mutations

in h blobs can disrupt the folding core or the kinetics of the binding core as well as specific interactions with other molecules. In disordered proteins, analogous to blob transitions, disorder $->$ order or order $->$ disorder transitions due to DA SNP has been observed (Vacic et al. 2012; Uversky et al. 2014). It has been found that these transitions frequently disrupt MoRFs in disordered proteins. Because MoRFs mediate protein-protein interactions, it follows that protein interaction networks may also be disrupted (Uversky et al. 2014).

We further examined whether a SNP could change blob phase annotation such as switching the blob from one region into another region in Das and Pappu phase diagram (Das & Pappu 2013) (Fig 3.3b). When analyzed for blob phase annotation transitions, the given mutation in a SNP can have two outcomes: (i) it can increase or decrease the FCR in a blob and thus move the blob into a new region in the Das and Pappu phase diagram (Das & Pappu 2013) or it can transition the blob itself which might change its phase assignment, resulting in a change in the phase annotation. ii) it does not change the blob phase annotation, resulting in no change (Fig 3.3a). We find that irrespective of whether the SNP is in ordered or disordered region DA SNPs are enriched in changing phase annotation of blobs.

We find that more than 25% of DA SNPs in both ordered and disordered region can change phase annotation. The phase annotation and κ values predict if a blob is collapsed and expanded (Das & Pappu 2013). It has been frequently observed that PTMs such as Ser/Thr phosphorylation can change the FCR and NCPR properties of IDPs and can thus tune the sequence ensemble relationships of IDPs due to their polyampholytic nature (Firman & Ghosh 2018; Das & Pappu 2013). We further examined the specific phase annotation transitions observed (Fig 3.4). We find that DA SNPs frequently transitions a strong polyampholyte blob to weak polyampholyte blob and vice versa in both ordered and disordered SNPs.

3.2.3 Disease associated mutations are enriched at the boundary of p blobs and larger h blobs

We further analyzed the likelihood of a mutation and its proximity at the boundary of h and p blobs (Fig 3.5). Interestingly, we find that DA SNPs are frequently found at the boundary of p blobs. We further tested the relationship between the length of a blob and its likelihood to be associated with diseases (Fig 3.6). We find that larger h blobs are more enriched in DA SNPs. This enrichment is higher in the disordered region SNPs when compared to the ordered region SNPs.

In p blobs we find the reverse to be true. We find that smaller p blobs are more enriched in DA SNPs. This is consistent with our above observation(Fig 3.5) that DA SNPs are enriched at the p boundary. Long linker (>10 residues) even in the disordered region are less likely to have DA SNPs.

3.2.4 Visualization of blob topology for disordered proteins

Conceptualization of long structured proteins relies heavily on the consecutive secondary structure elements that form the protein's topology, allowing for a coarse cartoon-style representation. No such approach for constructing an IDP topology has been available. In the current work we have also identified sequence properties of IDPs that are informative for whether mutations are likely to impact function. We have created a web application tool that visualizes this information for user-input protein sequences.

We thus present this conceptual tool, that will allow others to easily impose this sequence representation for their protein of choice and visualize the location of disease-associated SNPs within this representation (Fig 3.7). It also lets the user visualize the blobs at any cutoff in the range of 0-1.

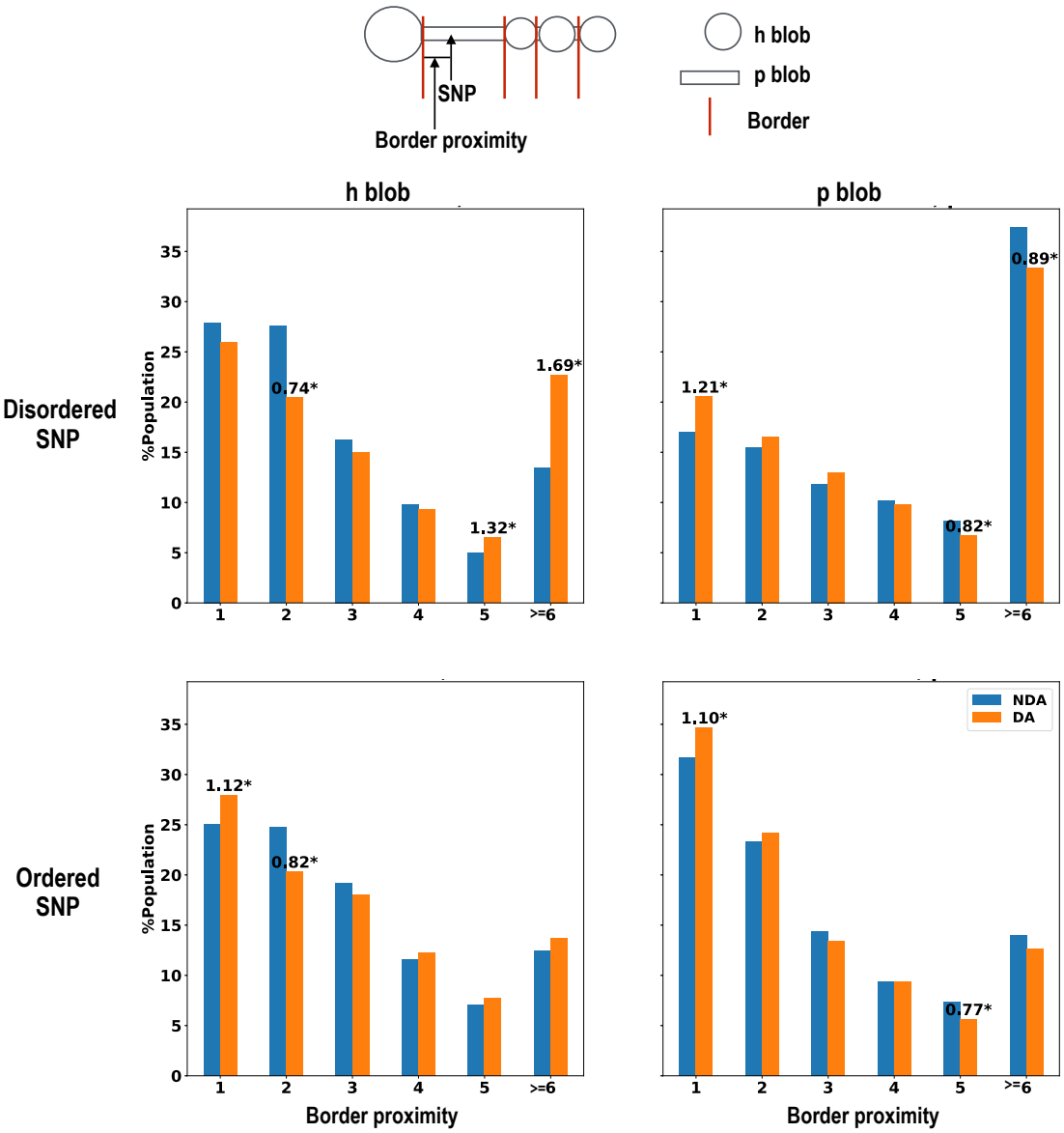

Figure 3.5: **Disease-associated SNPs are enriched at the boundary of p blobs.** Frequency of SNP residue locations within h (left) or p (right) blobs in disordered (top) and ordered (bottom) region. If a SNP lies at the border its proximity is labeled as 1, if its 1 residue away from the border its proximity is labeled as 2 and so on. If enrichment or depletion in DA SNPs is significant (p-value $< 5 \times 10^{-5}$), it is annotated with the observed fold enrichment and a star.

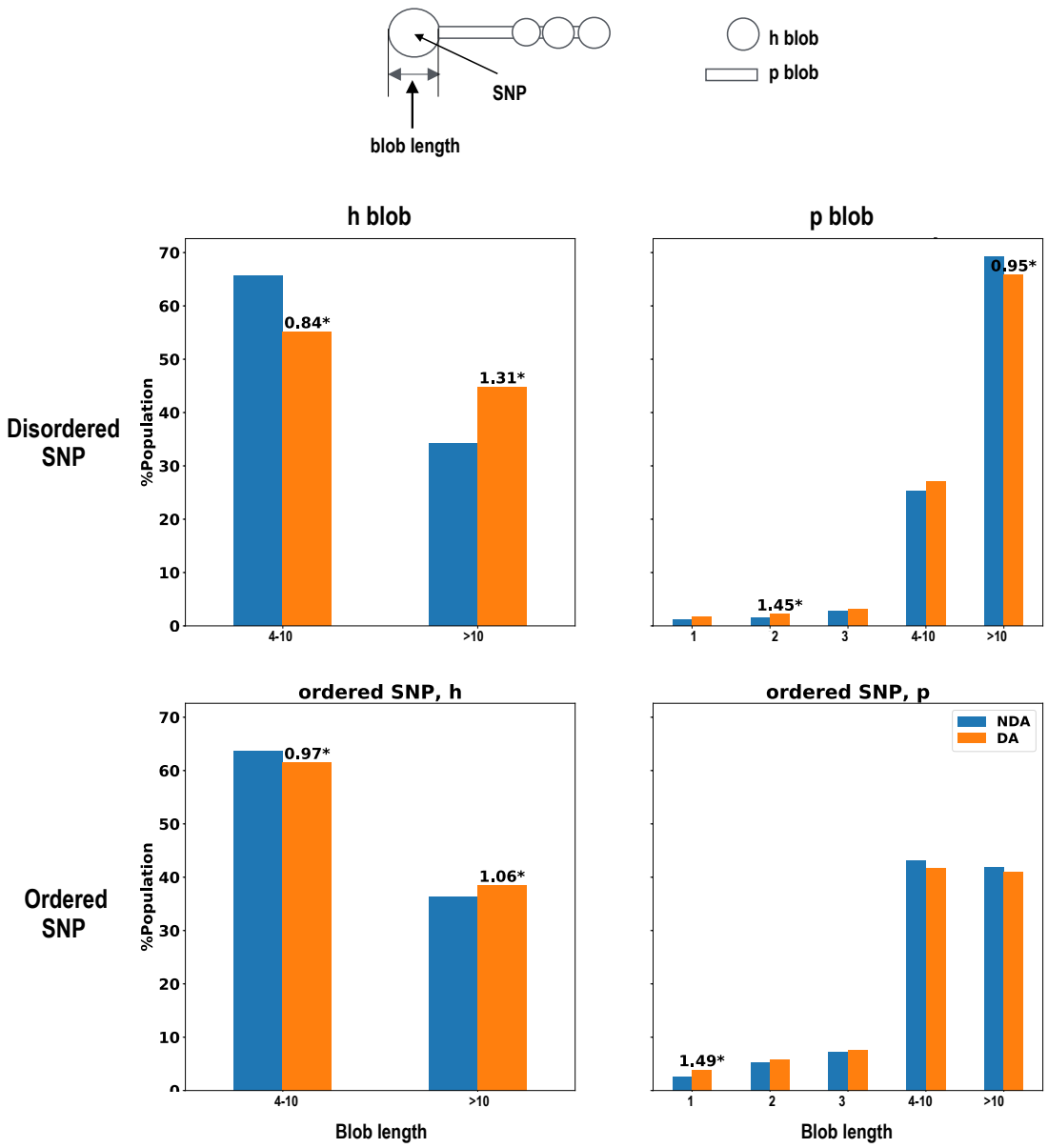

Figure 3.6: **Disease-associated SNPs are enriched in long h blobs and short p blobs.** a) Distribution of blob length of SNP containing blob in h (left) or p (right) blobs in disordered (top) and ordered (bottom) region SNPs. If enrichment or depletion in DA SNPs is significant (p-value $< 5 \times 10^{-5}$) it is annotated with the observed fold enrichment and a star.

3.3 Discussion

To summarize, our results shows that DA SNPs in IDPs frequently 1) causes blob transitions; h to p or p to h transitions, 2) causes phase annotations transitions;

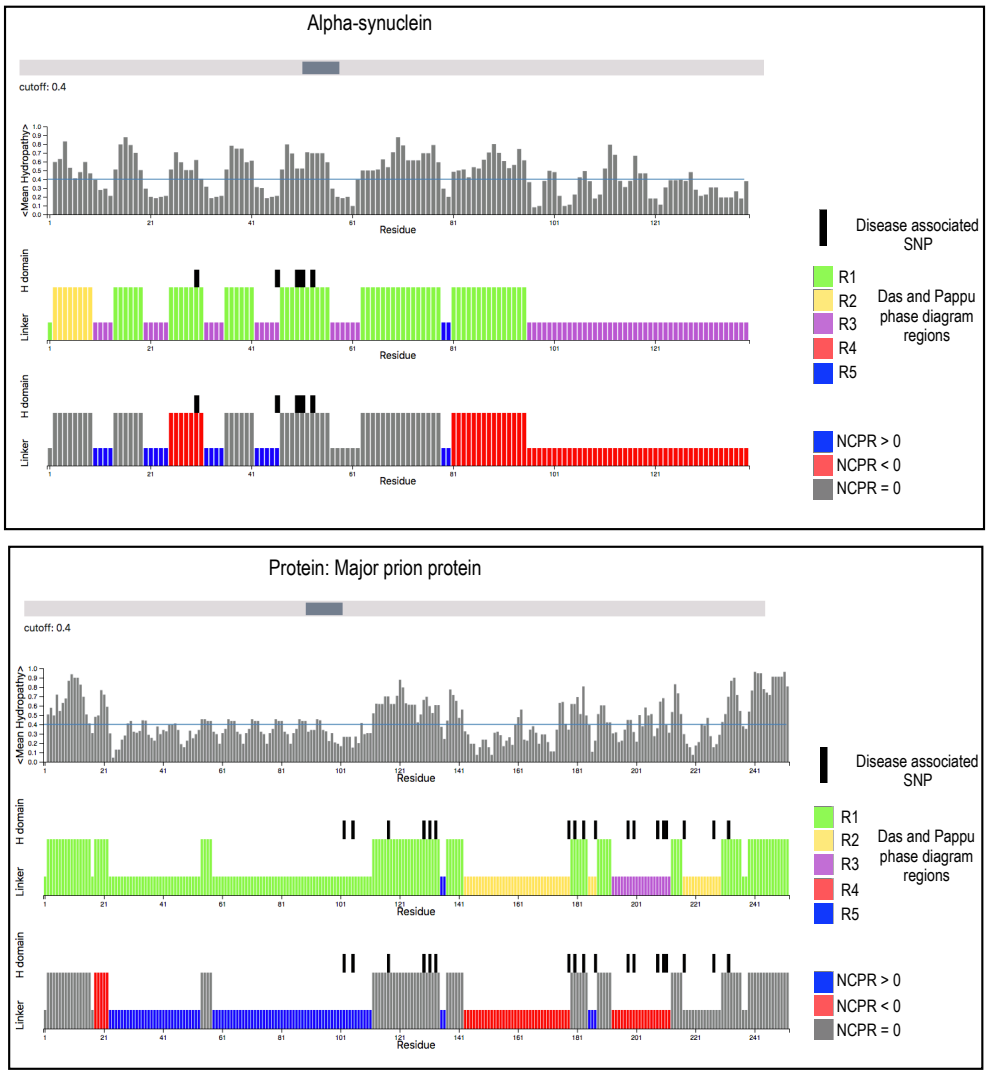

Figure 3.7: **Example of sequence decomposition approach applied to disordered proteins using the web tool.** Sequence decomposition approach applied to α-synuclein (top) and major prion protein identified by (bottom). The first row shows the mean hydrophobicity per residue (top), the middle row shows the digitized hydrophobicity colored by phase diagram region, and the bottom row is colored by net charge per residue. Within the digitized hydrophobicity, the high regions correspond to hydrophobic blobs, while the low regions correspond to non-hydrophobic blobs. DA SNP location is annotated with black bars

could change the blob to stronger or weaker polyampholytes, 3) occur in h blobs and the frequency further increases as the length of the blob increases 4) are depleted in p blobs but the depletion decreases as the length of the blob decreases.

The current phase diagram of IDPs are insensitive to single residue substitutions for long proteins. For example in an IDP of more than 30 residues, a single charge addition or deletion would probably not change the predicted phase behavior, because the number of changed residues is very small compared to the total number of residues. Firman et al (Firman & Ghosh 2018) showed that the certain regions within the Das and Pappu phase diagram are more sensitive to very small changes in FCR. With our current sequence decomposition, identified blobs are much more sensitive to a single residue mutation including charge neutral mutations.

BDNF Val66Met SNP which motivated this study seems to be typical of SNPs overall and thus is a useful model system. We observed various blob properties in the MD simulations of BDNF Val66Met SNP which supports our findings in this Chapter.

- Disease-associated SNPs are enriched in h blobs. We find that the Val66Met is found in a h blob and is associated with various disorders.

- Consistent with the observation that disease-associated SNPs are enriched in changing phase annotation. Protonation at residue His65 in both V66 and M66 changes the h2b blob phase annotation, and causes loss in specific interaction for this blob. It reduces its interaction with h3 group, including the loss of β coupling and Met-Met interaction in V66 and M66 sequence respectively.

- We find that a SNP at the boundary of p blobs is enriched to be associated with diseases. We observed for BDNF prodomain that some of the boundary residues play a critical role. We find that Val66 forms specific preferred

interaction with boundary residue Ser92. Another boundary residue Glu64 forms salt bridge with Arg93, to help stabilize β β coupling in V66 sequence.

- The residues in the long 15 residues p3 linker blob doesn't form any preferred interactions with the rest of the sequence and forms a visible boundary in the contact maps of all the BDNF prodomain sequence simulated. This is consistent with the observation that disease-associated SNPs are depleted in long linker regions probably because they do not form any specific contacts with the rest of the sequence.

3.4 Methods

Datasets

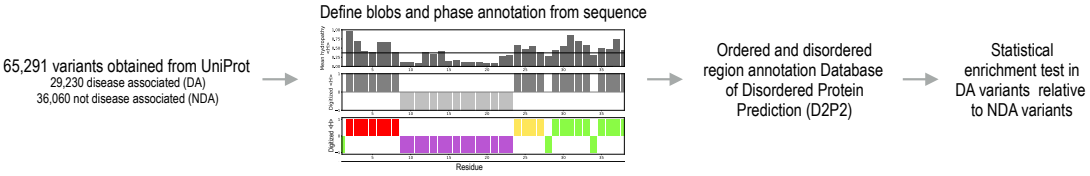

Figure 3.8: **Flowchart for the study.**.

The list of all missense SNPs annotated in human UniProtKB/Swiss-Prot entries was obtained from http://beta.uniprot.org/docs/humsavar (last Release: 8th May 2019) (Yip et al. 2008). This manually curated catalog contains missense mutations on the most common isoform of the given protein and does not contain frameshift and nonsense mutations. A SNP is annotated as 'Disease-associated (DA)' or 'Non Disease-associated (NDA)' depending on if it is implicated in disease or not according to literature reports. NDA SNP is also used to describe rare SNPs as well as polymorphisms that have an effect on protein function, but with no resulting clinical phenotype (functional polymorphisms) (Yip et al. 2008). A total of 65,291 SNPs were analyzed from 11,752 genes. Among the total missense

mutations found in the database, 29,230 (44.7%) are DA while the remaining 36,060 (55%) are NDA.

Disorder identification

Protein disorder for wild type sequences was obtained from the Database of Disordered Protein Prediction (D2P2) (http://d2p2.pro) (Oates et al. 2012). D2P2 has disorder predicted from nine disorder predictor including PONDR VL-XT (Romero et al. 2001; Li et al. 1999), PONDR VSL2b (Peng et al. 2006), PrDOS (Ishida & Kinoshita 2007), PV2 (Ghalwash et al. 2012), Espritz (all variants) (Walsh et al. 2012) and IUPred (all variants) (Dosztányi et al. 2005). We annotated any residue as disordered if any of the disorder predictors predicts it to be disordered.

Selecting the appropriate disorder predictor

We note that depending on the disorder predictors used the enrichment observed in DA SNPs could change. Most of the disorder predictors predict hydrophobic regions within disordered proteins as ordered. However, our enrichment analysis holds well even if no disorder predictor is used. Most of the enrichments observed in disordered regions are found in ordered regions as well and is also seen in the combined set. Every disorder predictor has its own flavor and determining which predictor to use for a given protein is a difficult task. For example, no two predictors predicted the same regions of disorder for proBDNF sequence (Fig 3.9).

Blob identification

Mean hydrophobicity ($\langle H \rangle$) at each residue is defined as the average Kyte-Dolittle(Kyte & Doolittle 1982) score with a window size of 3 residues, scaled to fit between 0

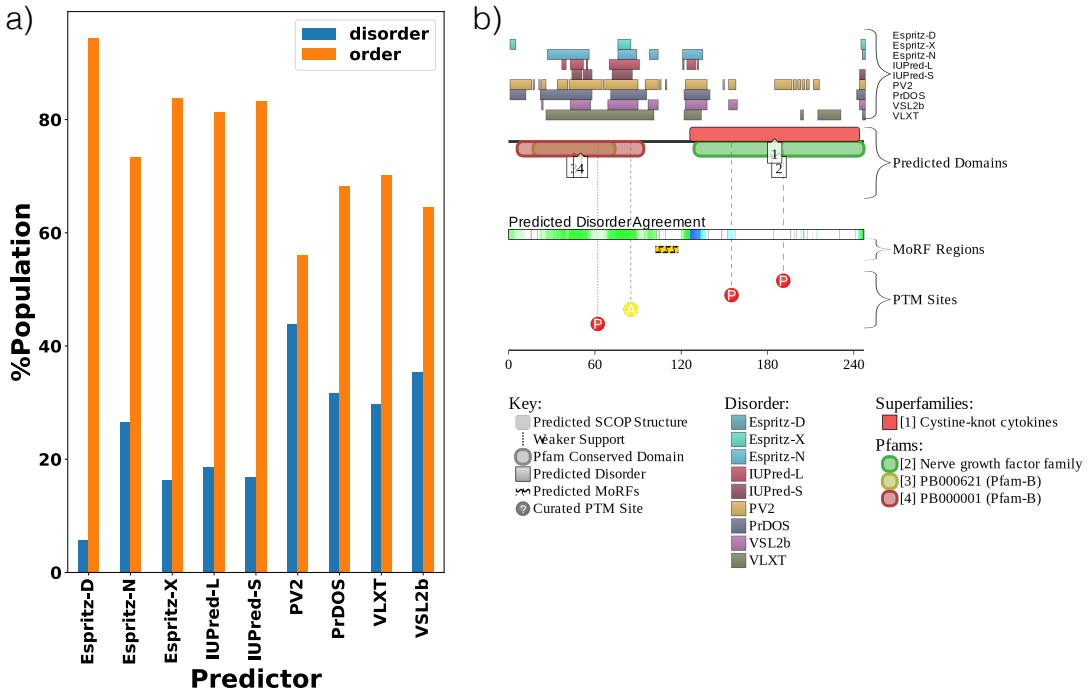

Figure 3.9: **Selecting the appropriate disorder predictor.**a) ordered and disordered proportions predicted by various disorder predictor. b) Predicted disorder for proBDNF from 9 different disorder predictor generated using Database of Disordered Protein Prediction (D2P2) (http://d2p2.pro) (Oates et al. 2012)

and 1. Any stretch of four or more residues with $\langle H \rangle > 0.4$ is classified as h blob and the remaining residues is classified as p blob.

Optimizing the cutoff for blobs identification

Optimizing the cutoff is a complex task. Increasing the boundary between h or b blobs increases enrichment in h blobs but reduces coverage. A cutoff in the range of 0.38 to 0.4 depending on whether the sequence is disordered or ordered respectively seems to be a good choice (Fig 3.10).

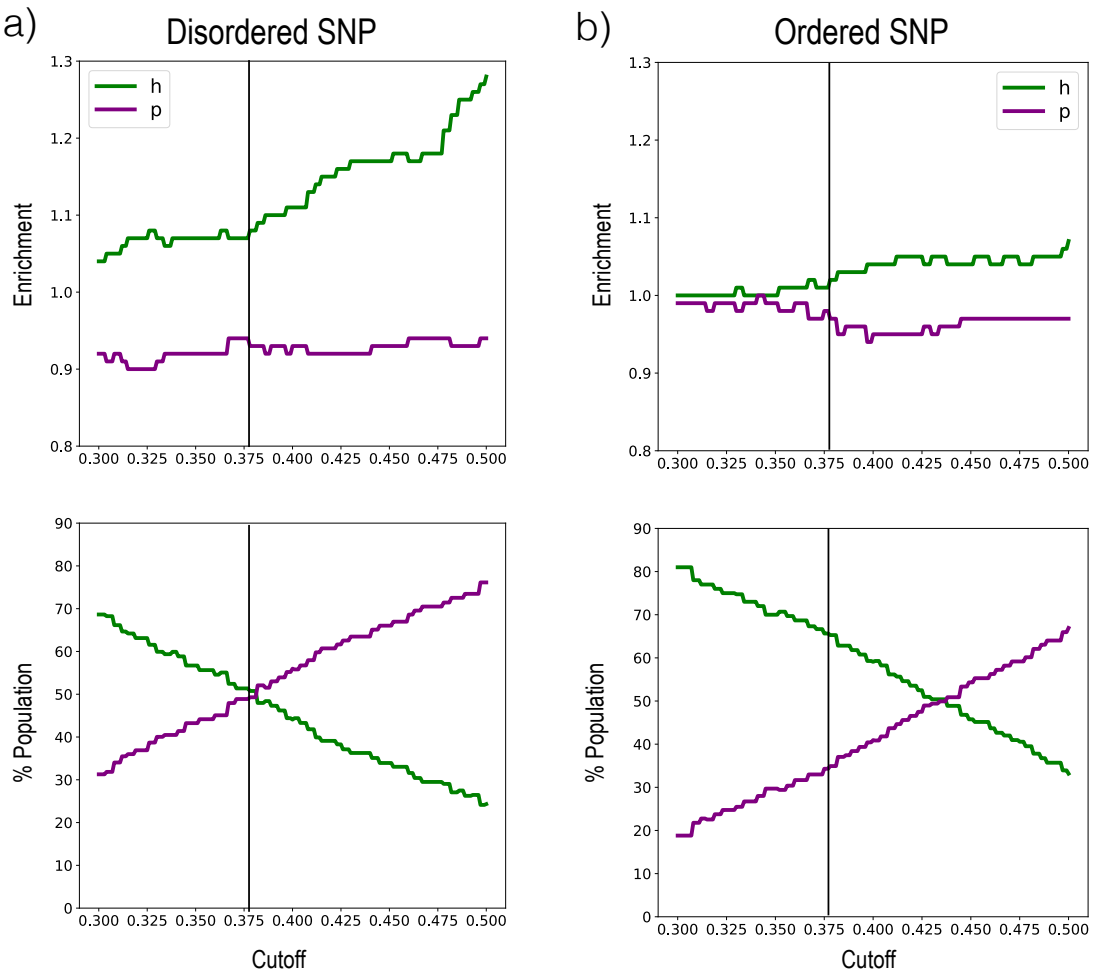

Figure 3.10: **Cutoff selection for blob identification.** Enrichment observed in DA SNPs for h and p blobs in ordered (a) and disordered (b) region for a given cutoff. The proportion of SNPs in h or p blobs for a given cutoff (bottom). For disordered SNPs (b) enrichment in h blob increases almost linearly with an increase in the cutoff.

Statistical analysis

Binomial test was used for calculating the fold enrichment. Any enrichment or depletion in DA SNPs is significant if p- value $<$ pcut-off = 5.0e-05.

Acknowledgments

Dr. Matt Hansen is the senior author in this project and has assisted with the analysis presented here. Kaitlin Bassi, an undergraduate researcher in the lab has helped with the web tool development.

1 Appendix A: β-pairing for blob pairs in V66 and M66 sequences

Frames were first clustered by whether the X-Y contact was formed (purple) or broken (green), and then by whether β structure was present in X (solid) or absent (dashed). The dark-gray window indicates the contacting blob that is constrained to have high or vanishing values by construction of the cluster, while the white window indicates the contacting blob without constrained secondary structure. If the contact is coupled to simultaneous β-strand formation, the peak within the white window for the solid purple curve should be significantly higher than other curves. Errors represent standard error of a Bernoulli trial with n number of samples, where n is the product of total number of unique replicas in a given cluster and average number of roundtrips per replica (17).

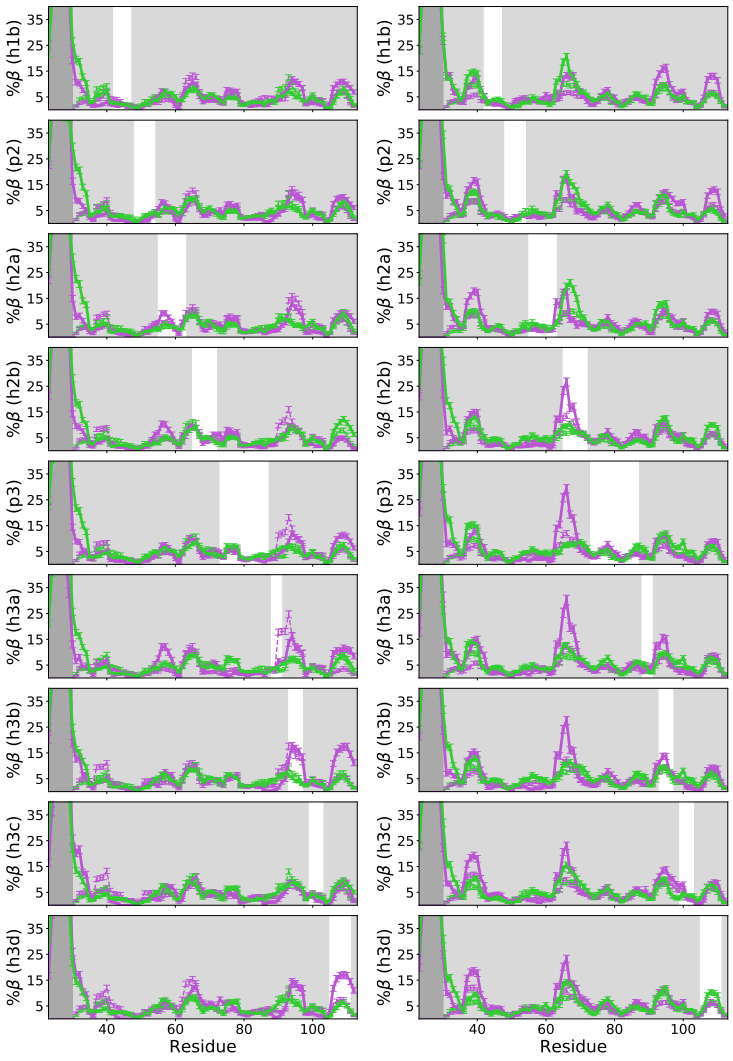

Figure 1: β**-pairing between blob p1 and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

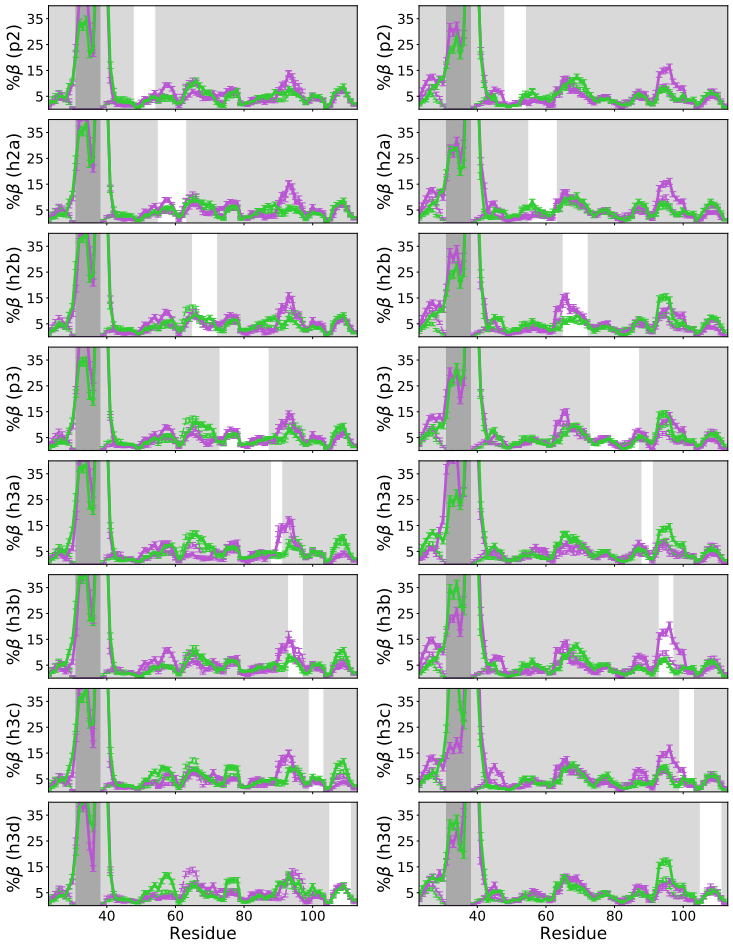

Figure 2: β**-pairing between blob h1a and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

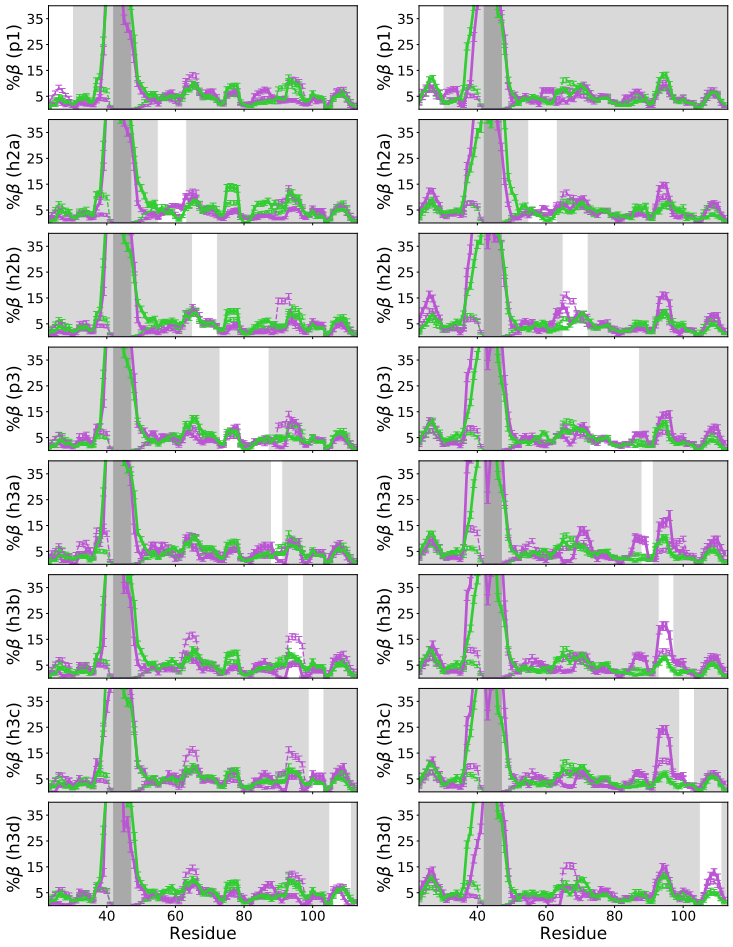

Figure 3: β**-pairing between blob h1b and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

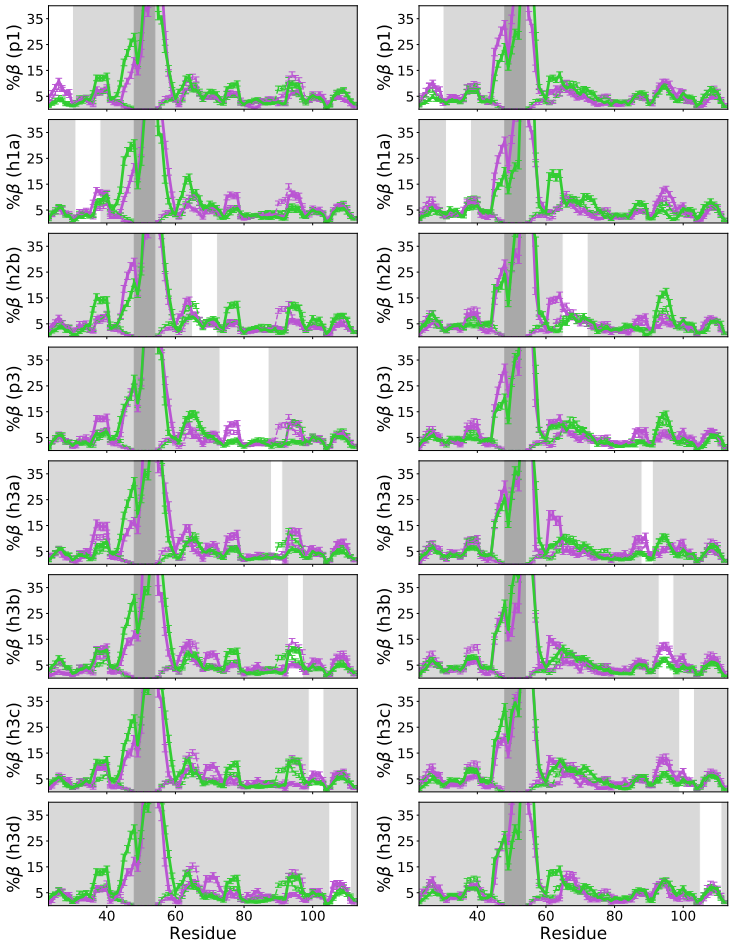

Figure 4: β**-pairing between blob p2 and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

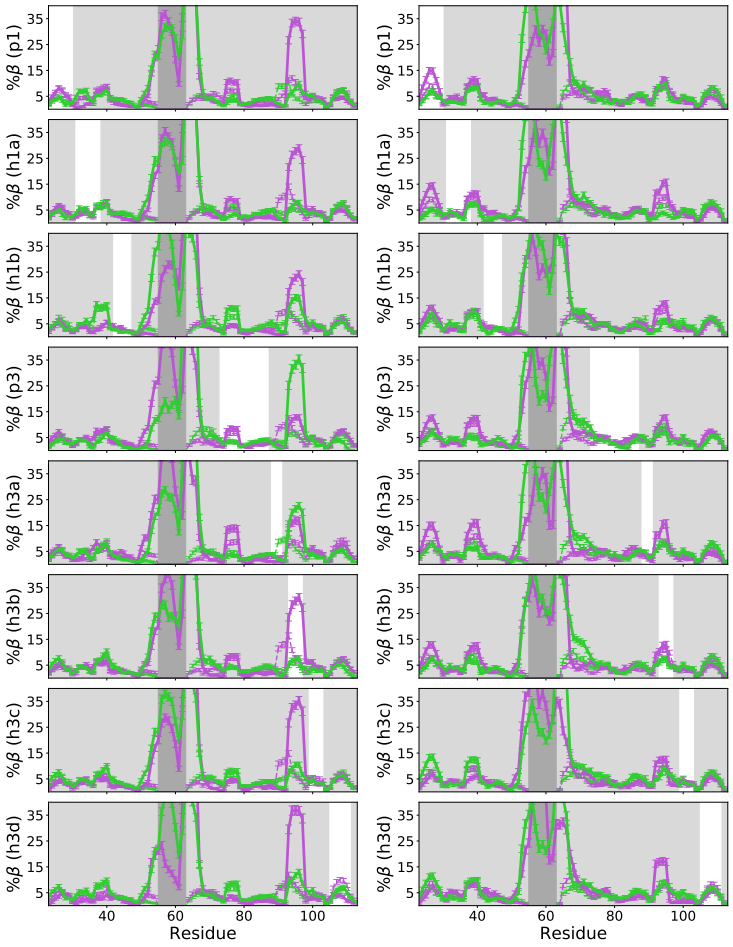

Figure 5: β**-pairing between blob h2a and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

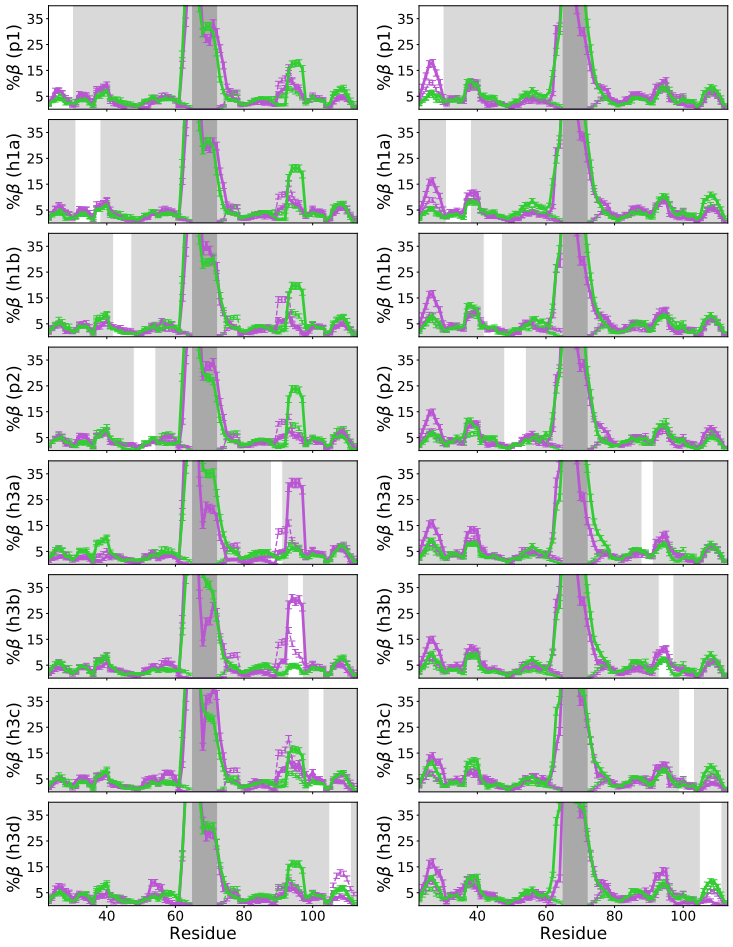

Figure 6: β**-pairing between blob h2b and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

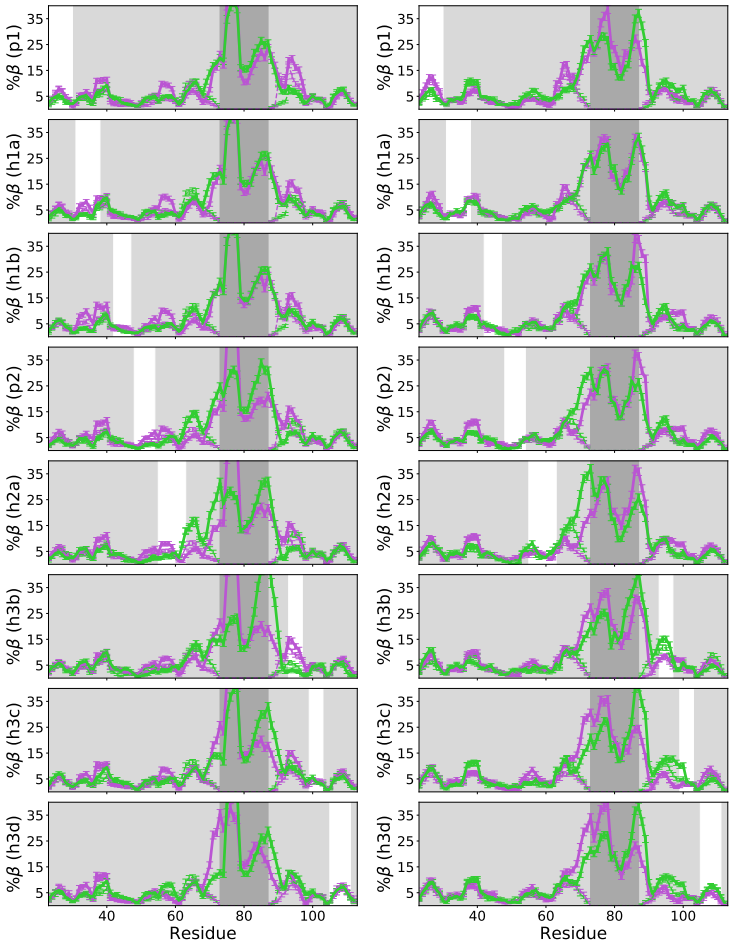

Figure 7: β**-pairing between blob p3 and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

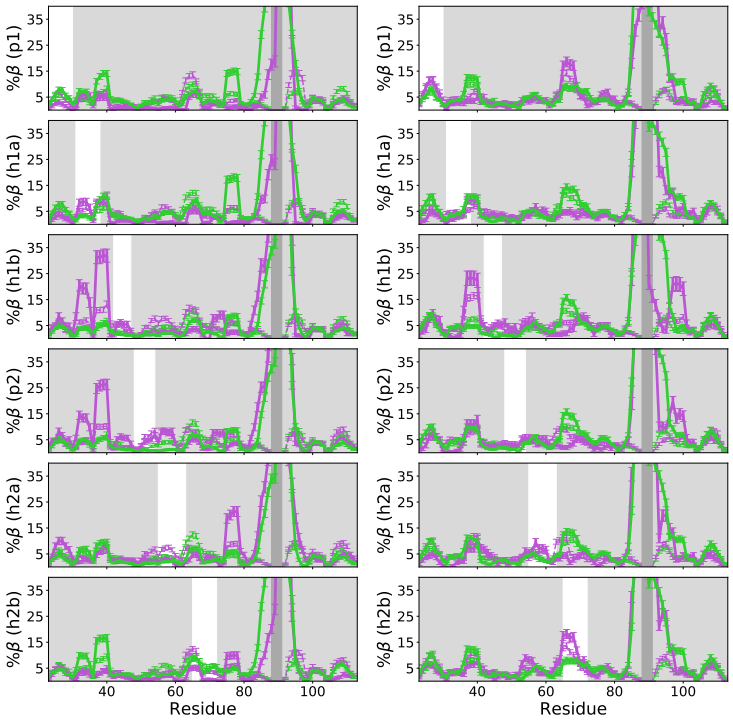

Figure 8: β**-pairing between blob h3a and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

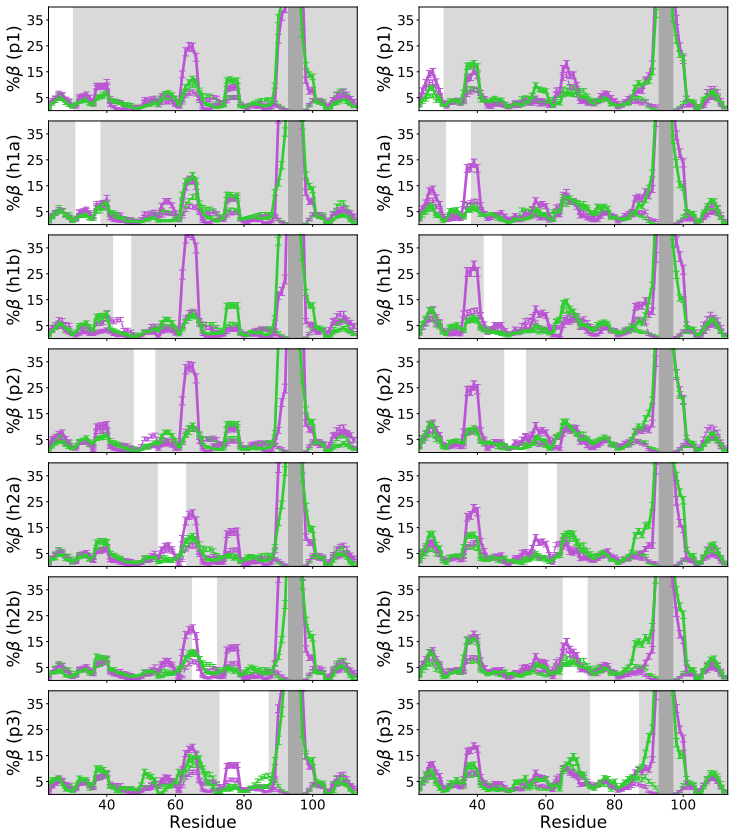

Figure 9: β**-pairing between blob h3b and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

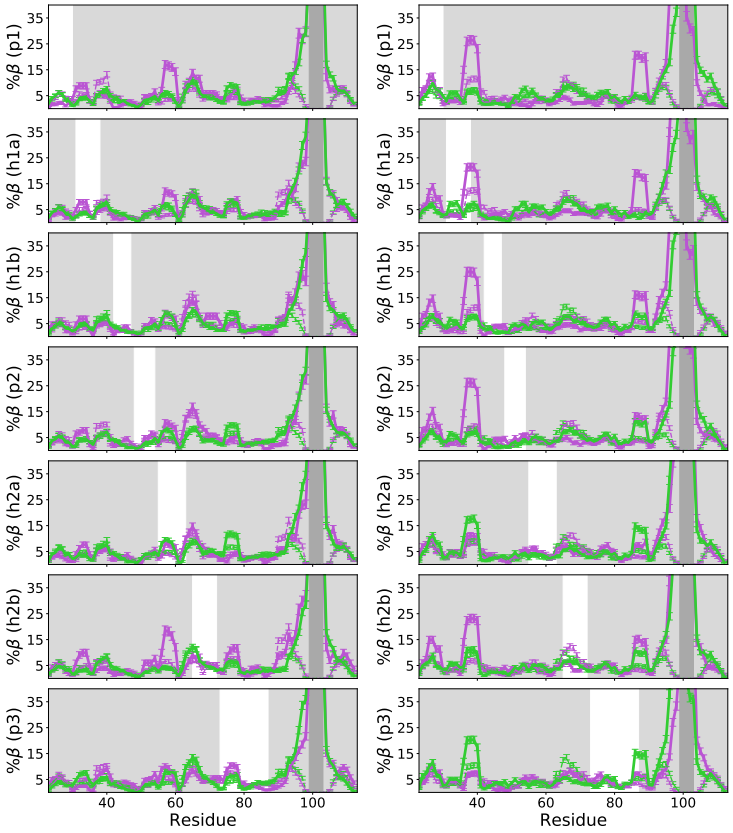

Figure 10: β**-pairing between blob h3c and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

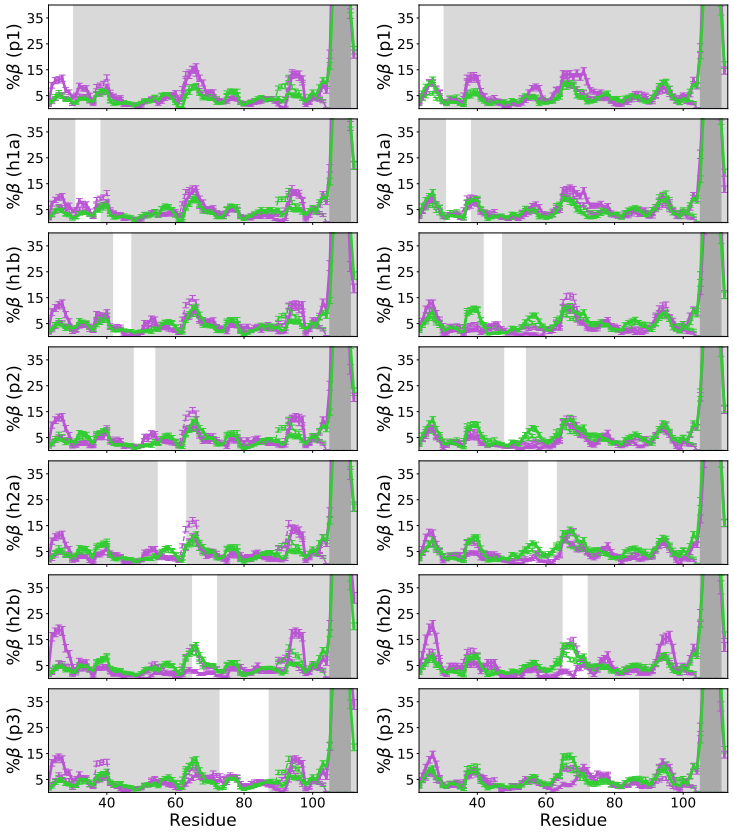

Figure 11: β**-pairing between blob h3d and each remaining blob, excluding adjacent or intra group pairs in the V66 (left) and M66 (right) sequences.**

Bibliography

Abascal, J. L. F., & Vega, C. 2005, J. Chem. Phys., 123, 234505

Abeln, S., & Frenkel, D. 2008, PLoS Comput. Biol., 4, e1000241

Abraham, M. J., Murtola, T., Schulz, R., et al. 2015, SoftwareX, 1-2, 19

Ahlstrom, L. S., Baker, J. L., Ehrlich, K., et al. 2013, J. Mol. Graph. Model., 46, 140

Akimoto, M., Selvaratnam, R., McNicholl, E. T., et al. 2013, Proc. Natl. Acad. Sci. U. S. A., 110, 14231

Anastasia, A., Deinhardt, K., Chao, M. V., et al. 2013, Nat. Commun., 4, 2490

Autry, A. E., Adachi, M., Nosyreva, E., et al. 2011, Nature, 475, 91

Autry, A. E., & Monteggia, L. M. 2012, Pharmacol. Rev., 64, 238

Bah, A., & Forman-Kay, J. D. 2016, J. Biol. Chem., 291, 6696

Ball, K. A., Wemmer, D. E., & Head-Gordon, T. 2014, J. Phys. Chem. B, 118, 6405

Benjamin, S., McQuoid, D. R., Potter, G. G., et al. 2010, Am. J. Geriatr. Psychiatry, 18, 323

Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. 1995, Comput. Phys. Commun., 91, 43

Best, R. B., & Hummer, G. 2009, J. Phys. Chem. B, 113, 9004

Best, R. B., Zheng, W., & Mittal, J. 2014, J. Chem. Theory Comput., 10, 5113

Bhattacharya, S., & Lin, X. 2019, Biomolecules, 9, doi:10.3390/biom9040146

Björkholm, C., & Monteggia, L. M. 2016, Neuropharmacology, 102, 72

Boehr, D. D., Nussinov, R., & Wright, P. E. 2009, Nat. Chem. Biol., 5, 789

Brzovic, P. S., Heikaus, C. C., Kisselev, L., et al. 2011, Mol. Cell, 44, 942

Buée, L., Bussière, T., Buée-Scherrer, V., Delacourte, A., & Hof, P. R. 2000, Brain Res. Rev., 33, 95

Burley, S., & Petsko, G. 1985, Science (80-.)., 229, 23

Canales, Á., Rösinger, M., Sastre, J., et al. 2017, PLoS One, 12, e0189171

Chen, Z.-Y., Bath, K., McEwen, B., Hempstead, B., & Lee, F. 2008, Novartis Found. Symp., 289, 180

Chen, Z.-Y., Ieraci, A., Teng, H., et al. 2005, J. Neurosci., 25, 6156

Conicella, A. E., Zerze, G. H., Mittal, J., & Fawzi, N. L. 2016, Structure, 24, 1537

Creamer, T. P., & Rose, G. D. 1992, Proc. Natl. Acad. Sci. U. S. A., 89, 5937

Darling, A. L., & Uversky, V. N. 2018, Front. Genet., 9, 158

Das, R. K., & Pappu, R. V. 2013, Proc. Natl. Acad. Sci., 110, 13392

Das, R. K., Ruff, K. M., & Pappu, R. V. 2015, Curr. Opin. Struct. Biol., 32, 102

Davies, A. M. 2003, EMBO J., 22, 2537

DeForte, S., & Uversky, V. N. 2016a, RSC Adv., 6, 11513

—. 2016b, Protein Sci., 25, 676

Dill, K. A. 1990, Biochemistry, 29, 7133

Dosztányi, Z., Csizmók, V., Tompa, P., & Simon, I. 2005, J. Mol. Biol., 347, 827

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., & Uversky, V. N. 2005, FEBS J., 272, 5129

Dyson, H. J., & Wright, P. E. 1998, Nat. Struct. Biol., 5, 499

—. 2005, Nat. Rev. Mol. Cell Biol., 6, 197

Erler, J., Zhang, R., Petridis, L., et al. 2014, Biophys. J., 107, 2911

Essmann, U., Perera, L., Berkowitz, M. L., et al. 1995, J. Chem. Phys., 103, 8577

Faure, G., Bornot, A., & de Brevern, A. G. 2008, Biochimie, 90, 626

Feng, D., Kim, T., Özkan, E., et al. 2010, J. Mol. Biol., 396, 967

Feuerstein, S., Solyom, Z., Aladag, A., et al. 2012, J. Mol. Biol., 420, 310

Firman, T., & Ghosh, K. 2018, J. Chem. Phys., 148, 123305

Flory, P. J. 1949, J. Chem. Phys., 17, 303

Fuxreiter, M., Tompa, P., Simon, I., et al. 2008, Nat. Chem. Biol., 4, 728

Ganguly, D., & Chen, J. 2015, PLOS Comput. Biol., 11, e1004247

Gao, M., Maynard, K. R., Chokshi, V., et al. 2014, J. Neurosci., 34, 10770

García, A. E., & Sanbonmatsu, K. Y. 2002, Proc. Natl. Acad. Sci. U. S. A., 99, 2782

Geist, L., Henen, M. A., Haiderer, S., et al. 2013, Protein Sci., 22, 1196

Ghalwash, M. F., Dunker, A. K., & Obradović, Z. 2012, Mol. BioSyst., 8, 381

Giza, J. I., Kim, J., Meyer, H. C., et al. 2018, Neuron, 99, 163

Glusman, G. 2013, Genome Biol., 14, 303

Gómez-Tamayo, J. C., Cordomí, A., Olivella, M., et al. 2016, Protein Sci., 25, 1517

Habchi, J., Tompa, P., Longhi, S., & Uversky, V. N. 2014, Chem. Rev., 114, 6561

He, Y., Chen, Y., Mooney, S. M., et al. 2015, J. Biol. Chem., 290, 25090

Hofmann, H., Soranno, A., Borgia, A., et al. 2012, Proc. Natl. Acad. Sci., 109, 16155

Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O., & Pappu, R. V. 2017, Biophys. J., 112, 16

Holehouse, A. S., & Pappu, R. V. 2018, Biochemistry, 57, 2415

Hornak, V., Abel, R., Okur, A., et al. 2006, Proteins Struct. Funct. Bioinforma., 65, 712

Huang, J., Rauscher, S., Nawrocki, G., et al. 2017, Nat. Methods, 14, 71

Huang, Z., Kirkwood, A., Pizzorusso, T., et al. 1999, Cell, 98, 739

Iakoucheva, L. M., Brown, C. J., Lawson, J., Obradović, Z., & Dunker, A. 2002, J. Mol. Biol., 323, 573

Iešmantavičius, V., Jensen, M. R., Ozenne, V., et al. 2013, J. Am. Chem. Soc., 135, 10155

Iglesias, J., Sanchez-Martínez, M., & Crehuet, R. 2013, Intrinsically Disord. Proteins, 1, e25323

Invernizzi, G., Lambrughi, M., Regonesi, M. E., Tortora, P., & Papaleo, E. 2013, Biochim. Biophys. Acta, 1830, 5236

Iqbal, S., Jespersen, J. B., Perez-Palma, E., et al. 2019, bioRxiv, 693259

Ishida, T., & Kinoshita, K. 2007, Nucleic Acids Res., 35, W460

Ithuralde, R. E., Roitberg, A. E., & Turjanski, A. G. 2016, J. Am. Chem. Soc., 138, 8742

Jin, F., Yu, C., Lai, L., & Liu, Z. 2013, PLoS Comput. Biol., 9, e1003249

Jorda, J., Xue, B., Uversky, V. N., & Kajava, A. V. 2010, FEBS J., 277, 2673

Jorgensen, W. L. 1981, J. Am. Chem. Soc., 103, 335

Karczewski, K. J., Francioli, L. C., Tiao, G., et al. 2019, bioRxiv, 531210

Kendrew, J. C., Bodo, G., Dintzis, H. M., et al. 1958, Nature, 181, 662

Kim, D. E., Chivian, D., & Baker, D. 2004, Nucleic Acids Res., 32, W526

Kircher, M., Witten, D. M., Jain, P., et al. 2014, Nat. Genet., 46, 310

Knott, M., Best, R. B., Hummer, G., de Bakker, P., & Word, J. 2012, PLoS Comput. Biol., 8, e1002605

Korte, M., Carroll, P., Wolf, E., et al. 1995, Proc. Natl. Acad. Sci., 92, 8856

Kovalskyy, D. B., & Ivanov, D. N. 2014, Biochemistry, 53, 1466

Kurcinski, M., Kolinski, A., & Kmiecik, S. 2014, J. Chem. Theory Comput., 10, 2224

Kyte, J., & Doolittle, R. F. 1982, J. Mol. Biol., 157, 105

Landrum, M. J., Lee, J. M., Benson, M., et al. 2018, Nucleic Acids Res., 46, D1062

Larini, L., Gessel, M. M., LaPointe, N. E., et al. 2013, Phys. Chem. Chem. Phys., 15, 8916

Lek, M., Karczewski, K. J., Minikel, E. V., et al. 2016, Nature, 536, 285

Levine, Z. A., Larini, L., LaPointe, N. E., Feinstein, S. C., & Shea, J.-E. 2015, Proc. Natl. Acad. Sci. U. S. A., 112, 2758

Levine, Z. A., & Shea, J.-E. 2017, Curr. Opin. Struct. Biol., 43, 95

Li, Romero, Rani, Dunker, & Obradovic. 1999, Genome Inform. Ser. Workshop Genome Inform., 10, 30

Li, P., Banjade, S., Cheng, H.-C., et al. 2012, Nature, 483, 336

Lim, J. M., Kim, G., & Levine, R. L. 2019, Neurochem. Res., 44, 247

Lindorff-Larsen, K., Piana, S., Palmo, K., et al. 2010, Proteins, 78, 1950

Liu, B.-h., Li, Y.-t., Ma, W.-p., et al. 2011, Neuron, 71, 542

Lu, H.-C., Chung, S. S., Fornili, A., & Fraternali, F. 2015, Front. Mol. Biosci., 2, 47

Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., & Pappu, R. V. 2010, Proc. Natl. Acad. Sci. U. S. A., 107, 8183

Marsh, J. A., Singh, V. K., Jia, Z., & Forman-Kay, J. D. 2006, Protein Sci., 15, 2795

McGregor, C. E., & English, A. W. 2019, Front. Cell. Neurosci., 12, 522

McKusick, V. A. 2007, Am. J. Hum. Genet., 80, 588

Meng, F., Bellaiche, M. M., Kim, J.-Y., et al. 2018, Biophys. J., 114, 870

Mercadante, D., Milles, S., Fuertes, G., et al. 2015, J. Phys. Chem. B, 119, 7975

Metzker, M. L. 2010, Nat. Rev. Genet., 11, 31

Minezaki, Y., Homma, K., Kinjo, A. R., & Nishikawa, K. 2006, J. Mol. Biol., 359, 1137

Mittag, T., & Forman-Kay, J. D. 2007, Curr. Opin. Struct. Biol., 17, 3

Mohan, A., Oldfield, C. J., Radivojac, P., et al. 2006, J. Mol. Biol., 362, 1043

Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S., & Deniz, A. A. 2007, Proc. Natl. Acad. Sci., 104, 2649

Ng, P. C., & Henikoff, S. 2003, Nucleic Acids Res., 31, 3812

Nielsen, J. T., & Mulder, F. A. A. 2018, J. Biomol. NMR, 70, 141

Nodet, G., Salmon, L., Ozenne, V., et al. 2009, J. Am. Chem. Soc., 131, 17908

Notaras, M., Hill, R., & van den Buuse, M. 2015, Mol. Psychiatry, 20, 916

Oates, M. E., Romero, P., Ishida, T., et al. 2012, Nucleic Acids Res., 41, D508

Ojeda-May, P., & Pu, J. 2013, Biophys. Chem., 184, 17

Oldfield, C. J., Cheng, Y., Cortese, M. S., et al. 2005, Biochemistry, 44, 1989

Oldfield, C. J., Meng, J., Yang, J. Y., et al. 2008, BMC Genomics, 9 Suppl 1, S1

Ortega, A., Amorós, D., & García de la Torre, J. 2011, Biophys. J., 101, 892

Panchenko, A. R., & Babu, M. M. 2015, Curr. Opin. Struct. Biol., 32, viii

Pappu, R. V., Wang, X., Vitalis, A., & Crick, S. L. 2008, Arch. Biochem. Biophys., 469, 132

Parashar, M., Brennan-Tonetta, M., Rodero, I., & Villalobos, J. J. 2018, doi:10.13140/RG.2.2.11579.87846

Patel, S., Ramanujam, V., Srivastava, A. K., & Chary, K. V. R. 2014a, Phys. Chem. Chem. Phys., 16, 12703

Patel, S., Vierling, E., & Tama, F. 2014b, Biophys. J., 106, 2644

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., & Obradovic, Z. 2006, BMC Bioinformatics, 7, 208

Peng, Z., Mizianty, M. J., Xue, B., Kurgan, L., & Uversky, V. N. 2012, Mol. Biosyst., 8, 1886

Pezawas, L., Verchinski, B. A., Mattay, V. S., et al. 2004, J. Neurosci., 24, 10099

Piana, S., Donchev, A. G., Robustelli, P., & Shaw, D. E. 2015, J. Phys. Chem. B, 119, 5113

Radhakrishnan, I., Pérez-Alvarado, G. C., Parker, D., et al. 1997, Cell, 91, 741

Rauscher, S., Gapsys, V., Gajda, M. J., et al. 2015, J. Chem. Theory Comput., 11, 5513

Riback, J. A., Katanski, C. D., Kear-Scott, J. L., et al. 2017, Cell, 168, 1028

Robustelli, P., Piana, S., & Shaw, D. E. 2018, Proc. Natl. Acad. Sci. U. S. A., 115, E4758

Romero, P., Obradovic, Z., Li, X., et al. 2001, Proteins Struct. Funct. Genet., 42, 38

Roy, A., Kucukural, A., & Zhang, Y. 2010, Nat. Protoc., 5, 725

Rubinstein, M., & Colby, R. H. 2003, Polymer physics (Oxford University Press), 440

Šali, A., & Blundell, T. L. 1993, J. Mol. Biol., 234, 779

Sawle, L., & Ghosh, K. 2015, J. Chem. Phys., 143, 085101

Shen, Y., & Bax, A. 2010, J. Biomol. NMR, 48, 13

Shen, Y., Delaglio, F., Cornilescu, G., & Bax, A. 2009, J. Biomol. NMR, 44, 213

Shoemaker, B. a., Portman, J. J., & Wolynes, P. G. 2000, Proc. Natl. Acad. Sci. U. S. A., 97, 8868

Sickmeier, M., Hamilton, J. A., LeGall, T., et al. 2007, Nucleic Acids Res., 35, D786

Soliman, F., Glatt, C. E., Bath, K. G., et al. 2010, Science, 327, 863

Staller, M. V., Holehouse, A. S., Swain-Lenz, D., et al. 2018, Cell Syst., 6, 444

Staneva, I., Huang, Y., Liu, Z., & Wallin, S. 2012, PLoS Comput. Biol., 8, e1002682

Stanley, N., Esteban-Martín, S., & De Fabritiis, G. 2015, Prog. Biophys. Mol. Biol., 119, 47

Sugita, Y., & Okamoto, Y. 1999, Chem. Phys. Lett., 314, 141

Sundaralingam, M., Bergstrom, R., Strasburg, G., et al. 1985, Science, 227, 945

Tompa, P., & Fuxreiter, M. 2008, Trends Biochem. Sci., 33, 2

Truong, P. M., Viet, M. H., Nguyen, P. H., Hu, C.-K., & Li, M. S. 2014, J. Phys. Chem. B, 118, 8972

Uesugi, M., Nyanguile, O., Lu, H., Levine, A. J., & Verdine, G. L. 1997, Science, 277, 1310

Uversky, V. N. 2009, Protein J., 28, 305

—. 2013, Biochim. Biophys. Acta, 1834, 932

—. 2019, Front. Phys., 7, 10

Uversky, V. N., Davé, V., Iakoucheva, L. M., et al. 2014, Chem. Rev., 114, 6844

Uversky, V. N., & Dunker, A. K. 2010, Biochim. Biophys. Acta - Proteins Proteomics, 1804, 1231

Uversky, V. N., Gillespie, J. R., & Fink, A. L. 2000, Proteins, 41, 415

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. 2008, Annu. Rev. Biophys., 37, 215

Vacic, V., & Iakoucheva, L. M. 2012, Mol. Biosyst., 8, 27

Vacic, V., Markwick, P. R. L., Oldfield, C. J., et al. 2012, PLoS Comput. Biol., 8, e1002709

Valley, C. C., Cembran, A., Perlmutter, J. D., et al. 2012, J. Biol. Chem., 287, 34979

van der Lee, R., Buljan, M., Lang, B., et al. 2014, Chem. Rev., 114, 6589

Verhagen, M., van der Meij, A., van Deurzen, P. A. M., et al. 2010, Mol. Psychiatry, 15, 260

Viet, M. H., Nguyen, P. H., Derreumaux, P., & Li, M. S. 2014, ACS Chem. Neurosci., 5, 646

Viet, M. H., Nguyen, P. H., Ngo, S. T., Li, M. S., & Derreumaux, P. 2013, ACS Chem. Neurosci., 4, 1446

Viguera, A. R., & Serrano, L. 1995, Biochemistry, 34, 8771

Vucetic, S., Xie, H., Iakoucheva, L. M., et al. 2007, J. Proteome Res., 6, 1899

Walsh, I., Martin, A. J. M., Di Domenico, T., & Tosatto, S. C. E. 2012, Bioinformatics, 28, 503

Ward, J., Sodhi, J., McGuffin, L., Buxton, B., & Jones, D. 2004, J. Mol. Biol., 337, 635

Weathers, E. A., Paulaitis, M. E., Woolf, T. B., & Hoh, J. H. 2006, Proteins Struct. Funct. Bioinforma., 66, 16

Wright, P. E., & Dyson, H. 1999, J. Mol. Biol., 293, 321

Wright, P. E., & Dyson, H. J. 2009, Curr. Opin. Struct. Biol., 19, 31

—. 2015, Nat. Rev. Mol. Cell Biol., 16, 18

Xu, L., Shan, S., & Wang, X. 2013, J. Phys. Chem. B, 117, 6206

Yang, J., Yan, R., Roy, A., et al. 2014, Nat. Methods, 12, 7

Yedvabny, E., Nerenberg, P. S., So, C., & Head-Gordon, T. 2015, J. Phys. Chem. B, 119, 896

Yip, Y. L., Famiglietti, M., Gos, A., et al. 2008, Hum. Mutat., 29, 361

Zerze, G. H., Best, R. B., & Mittal, J. 2015, J. Phys. Chem. B, 119, 14622

Zhan, Y. A., Wu, H., Powell, A. T., Daughdrill, G. W., & Ytreberg, F. M. 2013, Proteins Struct. Funct. Bioinforma., 81, 1738

Zhang, Y. 2008, BMC Bioinformatics, 9, 40