

https://dspace.org/

Investigating FOXP2 dynamics, stability, and DNA-binding capabilities

Perumal, Cardon

2024

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/41921

Perumal, Cardon. (2024). Investigating FOXP2 dynamics, stability, and DNA-binding capabilities [PhD thesis, University of the Witwatersrand, Johannesburg]. WireDSpace. https://hdl.handle.net/10539/41921

Downloaded from a DSpace repository. https://dspace.org/



Investigating FOXP2 dynamics, stability, and DNA-binding capabilities

by

Cardon Perumal

(1112043)

Thesis

Submitted in fulfilment of the requirements for the degree

Philosophiae Doctor

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Sylvia Fanucchi

October 2023

Declaration

I, Cardon Maria Perumal (1112043), am a student registered for the degree of Philosophiae Doctor (PhD) at the University of the Witwatersrand. I hereby declare the following:

- I am aware that plagiarism is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other university.
- I have followed the required conventions in referencing the work of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is evidence that this is not my own unaided work or that I have failed to acknowledge the sources of the ideas in my writing.



Ms C.M Perumal

2023/10/12

Abstract

All forkhead box (FOX) transcription factors share a conserved DNA-binding domain called the forkhead domain. They regulate gene expression in different organisms and have widespread biological roles, ranging from embryogenesis to immune regulation. The FOXP subfamily, unlike other FOX transcription factors, have the capacity for dimerisation, an evolutionary fate yet to be fully understood. In particular, the FOXP2 forkhead domain (FHD) has been shown *in vitro* to form domain-swapped dimers. Additionally, the FOXP2 leucine zipper domain forms heterotypic associations with FOXP1, 2, and 4. These somewhat isolated structural characterisations of FOXP2 have informed the domain-specific functionality of the DNA-binding forkhead domain but the leucine zipper domain has been characterized to a lesser extent, although it is thought to be implicated in FOXP2 DNA-binding as well. Elucidating the structural and functional impact of the leucine zipper on FOXP2 DNA-binding remains challenging, as it is unclear how these motifs work together to achieve binding and whether complexation is a requirement for DNA-binding. Owing to this, the cooperative structural contributions of both the leucine zipper and forkhead domain and the effect of DNA on the structure and stability of these domains have not been considered. Consequently, the aim of this study was to gain a comprehensive understanding of the FOXP2 DNA-binding mechanism by comparing the structure, stability, and dynamics of both the leucine zipper domain and the forkhead domain in the presence and absence of DNA. Assessed here, for the first time, is the conformational dynamics of the FOXP2 leucine zipper domain and FHD flanking disordered regions using hydrogen-deuterium exchange mass spectrometry. The results confirm the binding of DNA to the FHD recognition helix as well as the change in dynamics of the interlinking loop region. Additionally, the FOXP2 folding mechanisms and stability for each domain was characterised, revealing a 2-state unfolding mechanism for the forkhead domain and a 3-state mechanism for the longer LeuZip variant. Fluorescence anisotropy studies revealed that the LeuZip variant bound DNA with a higher affinity than the forkhead domain. The findings of this study highlight the structural significance of the leucine zipper domain and unstructured regions and the functional cooperativity of the FOXP2 domains investigated.

Contents

Investigating FOXP2 dynamics, stability, and DNA-binding capabilities i
Declaration i
Abstract ii
List of abbreviations v
List of figures vii
List of tables viii
Chapter 1: Introduction 1
1.1 Transcription factors 2
1.1.2 Winged-helix transcription factors 6
1.2 Protein folding and stability 9
1.3 Oligomerisation 11
1.3.1 Oligomerisation in transcription factors 12
1.4 FOX Transcription Factors 15
1.4.1 FOXP Transcription Factors 16
1.4.2 FOXP Oligomerisation 18
1.5 FOXP2 Transcription Factors 19
1.5.1 FOXP2 Forkhead Domain 19
1.5.2 FOXP2 oligomerisation 20
1.5.2.1 Forkhead domain oligomerisation 20
1.5.2.2 Leucine zipper oligomerisation 21
1.6 DNA and FOXP2 22
1.6.1 DNA-binding associated with FOXP2 22
1.7 Rationale 26
Chapter 2: Aim and Objectives 28
2.1 Aim 28
2.2 Objectives 28
Chapter 3: Materials and Methods 29
3.1 Experimental constructs 29
3.1.1 Protein constructs 29
3.1.2 Oligonucleotides 31
3.2 Transformation 31
3.3 Protein expression and cell harvesting 31
3.4 Purification 33
3.5 Protein molecular weight and homogeneity 35
3.6 Protein purity and concentration 36

3.7 Structural characterisation 37
3.7.1 Secondary structure characterisation 37
3.7.2 Tertiary structure characterisation 38
3.8 Equilibrium unfolding experiments 39
3.8.1 Urea Induced Equilibrium Unfolding 39
3.9.1 Electrophoretic mobility-shift assays 43
3.9.2 Fluorescence Anisotropy 44
3.9.3 Hydrogen deuterium exchange mass spectrometry 45
Chapter 4: Results 49
4.1 Expression and purification of the LZ-END and FHD constructs 50
4.2 Protein homogeneity and molecular weight confirmation 53
4.3 Assessing secondary structure 55
4.4 Investigating the tertiary structure of the LZ-END and FHD constructs 57
4.5 Assessing LZ-END and FHD functionality and oligomeric state via electrophoretic mobility shift assays 60
4.6 Investigating FHD and LZ-END stability via fluorescence anisotropy 62
4.7 Investigating the conformational dynamics of FOXP2 by hydrogen-deuterium exchange mass spectrometry 64
4.7.1 FOXP2 leucine zipper conformational dynamics 66
4.7.2 FOXP2 forkhead domain conformational dynamics 68
4.7.3 FOXP2 unstructured regions' conformational dynamics 72
4.8 Investigating equilibrium unfolding of the LZ-END construct and FHD 75
4.8.1 Establishing Reversibility 76
4.8.2 Urea-induced equilibrium unfolding assessed at the secondary and tertiary structure level 78
5. Discussion 84
6. Appendix 94
7. References 119

List of abbreviations

aa	amino acid
ACN	acetonitrile
AMAS	N-(α-maleimidoacetoxy)-succinimide ester
APS	ammonium persulfate
ART	acid-rich tail
bHLH	basic helix-turn helix
bp	base pairs
bZIP	basic leucine zipper
CLMS	cross-linking mass spectrometry
CRR	cis-regulatory regions
DTT	dithiothreitol
E. coli	Escherichia coli
FHD	forkhead domain
FOX	forkhead box
H1/H2/H3	helix 1/helix 2/helix 3
HDX MS	hydrogen deuterium exchange mass spectrometry
His-tag	poly-histidine tag
HD ZIP	homeodomain-leucine zipper
HTH	helix-turn-helix
ILR	interlinking loop region
IMAC	immobilised metal affinity chromatography
IPTG	isopropyl-1-thio-β- D-galactopyranoside
*k*d/ *k*a	association/dissociation rate constant
K_D	binding constant
LC	liquid chromatography
LeuZip	leucine zipper
LZ-END	leucine zipper domain to the end of FOXP2
MADS-box	mini-chromosome maintenance 1, agamous, deficiens and serum response factor
PAS	Per-Arnt-Sim
PMT	photon multiplier tube
Poly-Q	polyglutamine region
PTM	post-translational modification
SDS-PAGE	sodium dodecyl sulphate-polyacrylamide gel

	electrophoresis
STAT	signal transducer and activator of transcription
TEMED	tetramethylethylenediamine
TF	transcription factor
Tregs	regulatory T cells
v/v	volume/ volume
w/v	weight/ volume
W1/W2	wing 1/wing 2 (of the forkhead domain)
wt	wildtype
ZNF	zinc finger

List of figures

Figure 1. Schematic illustrating the process of transcription. 2
Figure 2. Schematic illustrating transcription factor-DNA recognition mechanisms....4
Figure 3. Schematic depicting examples of crystalised DNA-binding domains. 6
Figure 4. Diagrammatic representation of the FOXA1 winged-helix protein. 8
Figure 5. Protein folding funnel representing energy landscapes of the protein folding pathway....11
Figure 6. Illustration of the monomer and dimer DNA-binding pathways in transcription factors. 13
Figure 7. Schematic depicting the different oligomerisation domains found in transcription factors....15
Figure 8. Schematic illustrating the FOXP subfamily domains....17
Figure 9. Schematic illustrating the FOXP2 domains for the full-length FOXP2 protein (1 to 715 amino acids). 19
Figure 10. Illustration of the FOXP2 forkhead domain monomer and domain-swapped dimer. 21
Figure 11. Structure of the FOXP2 FHD bound to DNA. 24
Figure 12. Superimposition of FOX transcription factors to highlight structural differences and similarities upon DNA-binding. 25
Figure 13. Amino acid sequences of the FOXP2 LZ-END construct and FHD construct, inclusive of poly-histidine tags and thrombin cleavage sites....30
Figure 14. Structural representation of the non-covalent bond formation between histidine and Ni^{2+}. ...34
Figure 15. Schematic illustrating the experimental workflow of a typical bottom-up HDX-MS setup....46
Figure 16. Illustration of the experimental progression and datasets presented in the study. 50
Figure 17. Elution profiles of the LZ-END and FHD construct, purified via IMAC 52
Figure 18. SDS-PAGE and calibration curve analyses of the purified LZ-END and FHD constructs. 54
Figure 19. Circular dichroism spectropolarimetry profiles of the LZ-END and FHD constructs in the absence and presence of DNA. 56
Figure 20. Intrinsic tryptophan fluorescence emission spectra of the LZ-END and FHD constructs in the absence and presence of DNA and 8 M Urea. 59
Figure 21. Native electrophoretic mobility shift assays for the LZ-END and FHD constructs. 61
Figure 22. DNA-binding isotherm illustrating the independent binding interaction of the LZ-END and FHD constructs with DNA. 63
Figure 23. FOXP2 AlphaFold model with predicted local distance difference test score. 65
Figure 24. Comparison of the time-lapsed conformational dynamics of the leucine zipper domain in the absence and presence of DNA. 68
Figure 25. Comparison of the time-lapsed conformational dynamics of the FOXP2 forkhead domain of the LZ-END construct, in the absence and presence of DNA. 71
Figure 26. Comparison of the apo time-lapsed conformational dynamics of the full FOXP2 LZ-END. 72
Figure 27. Comparison of the DNA-complexed, time-lapsed conformational dynamics of the full FOXP2 LZ-END construct. 73
Figure 28. Native, unfolded and refolded secondary and tertiary profiles of the LZ-END and FHD construct. 77
Figure 29. The equilibrium unfolding curves of the LZ-END and FHD at the tertiary and secondary structural level. 80
Figure 30. Comparative illustration of the isolated forkhead domain and full length FOXP2 structure. 85
Figure 31. Comparing the conformational dynamics of the isolated forkhead domain (by Morris *et al.,* 2018) and the LZ-END construct in the presence of DNA. 90
Figure 32. Heat map illustrating the backbone dynamics of the apo LZ-END protein....94
Figure 33. Heat map illustrating the backbone dynamics of the LZ-END protein in complex with DNA. 95
Figure 34. Difference map illustrating the backbone dynamics of the LZ-END protein in the absence and presence of DNA. 96

Figure 35. Comparison of the time-lapsed conformational dynamics of the full FOXP2 LZ-END construct in the absence and presence of DNA.....97
Figure 36. Deuterium uptake plots for the unstructured interlinking loop containing the *PTTPTAPVTP* sequence.....98

List of tables

Table 1. IMAC purification wash steps for the LZ-END and FHD constructs:.....34
Table 2. MS1 and MS2 analysis settings.....47
Table 3. Comparison of the thermodynamic parameters obtained from equilibrium unfolding curves of the LZ-END and FHD construct. 81

Chapter 1: Introduction

All living organisms function according to a set of biological instructions, which are predetermined by hereditary units called genes (Wilhelm Johannsen; 1905; Lewin and Dover, 1994). Genes encompass all the information necessary for determining the traits of an organism. They comprise DNA, a stable molecular material which contains vital information required to maintain and continue the life of an organism (Watson and Crick; 1953; Cooper *et al*., 2000). In biology, the central dogma is used to describe the biological architecture that facilitates the processing of DNA contained in genes, to protein products. It communicates that DNA is transcribed into RNA transcripts, which are subsequently translated into proteins (Crick, 1970).

Transcription is a fundamental process that transcribes nucleotides into amino acids (Kornberg *et al*., 2000). Transcription in bacteria, such as *Escherichia coli*, is comprised of four general steps. These include (i) identifying the promoter region of the template DNA strand, (ii) transcription initiation, (iii) elongation, and (iv) termination (Kornberg, 2007; Sanders and Bowman, 2019). The process employs a single DNA strand (template) to construct an RNA transcript. The holoenzyme, RNA polymerase, initiates this process by identifying specific recognition sequences, 5' upstream of the template DNA strand (Zawel and Reinberg, 1993). Free RNA nucleotides suspended in the nucleus then base pair with nucleotides found on the DNA template strand, a process facilitated by RNA polymerase and other enzymes (Topal and Fresco, 1976). RNA bases are added to the lengthening RNA strand until RNA polymerase recognises a termination sequence, located at the 3' end of the template strand (Birnstiel *et al*., 1985). Subsequently, the process of transcription is discontinued, and a mature mRNA transcript is produced as the final product (Erkmann and Kutay, 2004). An overview of the transcription process is shown in Fig.1 below. The mRNA transcript is then ready to be transported out of the nucleus, to the cytoplasm, where the process of translation can take place (Erkmann and Kutay, 2004). Briefly, the mRNA strand is used to produce a string of amino acids, specified by codons in the mRNA strand (McKee and McKee, 2003). This string of amino acids forms the basis for protein synthesis (McKee and McKee, 2003).

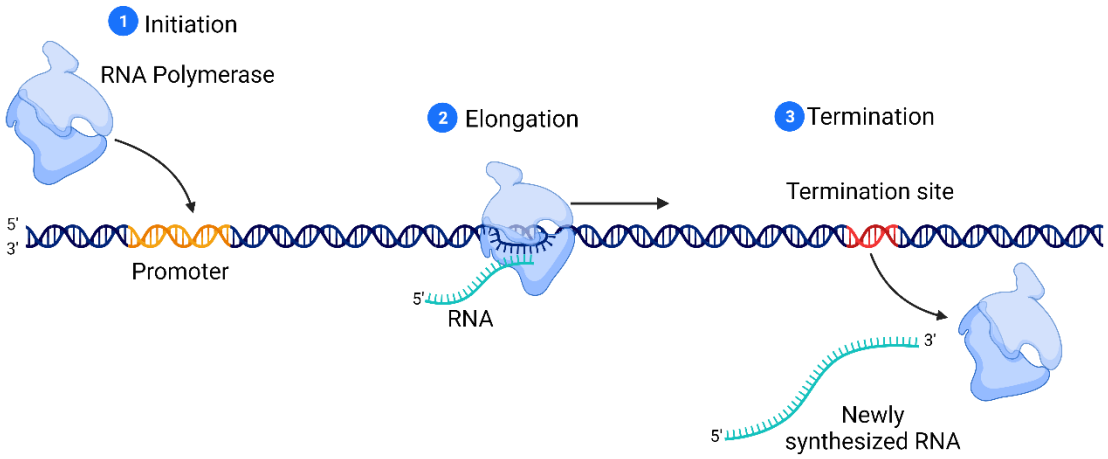


Figure 1. Schematic illustrating the process of transcription.

The process of transcription is initiated by RNA polymerase which recognising a promoter sequence on double-stranded DNA. During elongation, the double-stranded DNA unwinds, and the RNA polymerase projects along the strand, adding free nucleotides to the 3' prime end. The process is terminated when RNA polymerase encounters the termination site and the newly synthesised RNA strand, along with the RNA polymerase molecule is released [Image created originally using BioRender].

The process of transcription is achieved through the complexation of many proteins, not just RNA polymerase, that work together to ensure that the right genes are being expressed at the right time and in adequate amounts (Alberts *et al*., 2013). Thus, the process is highly regulated and requires a group of specialised proteins, other than RNA polymerase, to guide efficient and error-free transcription.

1.1 Transcription factors

Transcription is regulated by two main classes of proteins, that is, gene-nonspecific factors (like RNA polymerase II) and gene-specific factors, called transcription factors (TFs) (Suter, 2020). Transcription factors are proteins that regulate gene expression and monitor and control transcriptional events within cells (Latchman, 1993). They usually work in conjunction with other proteins and co-factors to maintain homeostasis and optimal biological functioning and are important in elucidating disease response (Latchman, 1990; Chen and Rajewsky, 2007). Transcription factors are characterised by their ability to bind target DNA strands by recognising short DNA regulatory regions

on chromatin, which initiates either the downregulation or upregulation of certain genes (Grunstein, 1997). In eukaryotes, these sequences are usually short degenerate sequences (6 - 12 bp long). The recognition sequences alone do not provide the effectual specificity required to ensure successful transcription factor binding/ recognition. This low sequence specificity suggests that more complex mechanisms dictate the functional outcome of TF binding and indeed, there are several contributing factors. The selectivity is challenged by complex cellular networks, crowding of TFs at several regulator sites on a particular sequence, as well as dynamic conformational changes that can occur in the protein-DNA complex (Todeschini *et al.,* 2014). In some cases, transcription factors need to be flexible at the binding sites or work with cofactors or other transcription factors to achieve binding and execute regulatory control (Todeschini *et al.,* 2014; Reiter *et al.,* 2017; Yesudhas *et al.,* 2017).

The specificity of transcription regulation is governed by the cooperativity of the expression systems' trans elements with the cis-regulatory regions (CRRs) of DNA (Spitz & Furlong, 2012). Together, these elements elicit a transcriptional response. Enhancer regions mostly comprise a cluster of regulatory sites, and with the expression of TFs that share overlapping spatial domains, the resultant modes of binding may produce distinct expression frameworks that direct specific TF binding events.

Deconvoluting the mechanisms used by TFs to bind DNA remains elusive. However, a few mechanisms have been proposed, as shown in Fig. 2. The predominant argument describes a ‘sliding’ mechanism, where the TF, initially located on a non-target site, progresses along the DNA until it arrives at the target binding site (Suter, 2020; Yesudhas *et al.,* 2017). Given the propelling nature of this type of mechanism, the sliding rate is therefore dependent on the hydrodynamic radius of the TF; indeed, larger TFs would require a greater rotational force to be propagated along the DNA and would do so at a slower rate (Muers, 2012; Yesudhas *et al.,* 2017; Marklund *et al.,* 2017).

Another intriguing mechanism called the ‘hopping’ mechanism, involves the displacement or dissociation of a TF, in intranuclear space, from a non-specific site to the target site (Yesudhas *et al.,* 2017). The jump need not be as exaggerated as the name suggests, as the displacement can occur by electrostatic interactions within the

same chain and binding may be proximal to the original dissociation site. It is suggested that the hopping mechanism would occur at a slower rate than the sliding mechanism because the energy required for dissociation and re-association is greater (Suter, 2020). The hopping mechanism illustrates the electrostatic and non-electrostatic interactions that govern precise binding in TFs that employ this strategy (van Zon, 2006; Swift & Coruzzi, 2017).

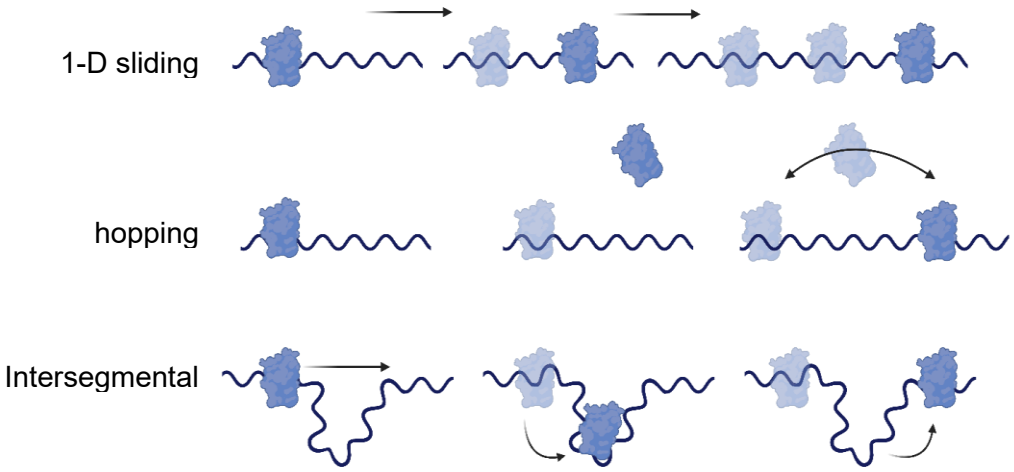


Figure 2. Schematic illustrating transcription factor-DNA recognition mechanisms.

The 1-D sliding mechanism occurs by propagation of the TF along the DNA strand towards the binding-site by electrostatic-assisted diffusion. The hopping mechanism is a result of dissociation and re-association to the binding site due to electrostatic and non-electrostatic interactions between the TF and the DNA. Intersegmental transfer occurs mainly for TFs with two binding sites that flank a segment of DNA. Here the TF transiently binds both sites due to the formation of a DNA loop that brings both sites closer together (Adapted from: Yesudhas *et al.,* 2017; Image created in BioRender).

In intersegmental transfer, the TF transiently binds two sites concurrently, due to a looping or bended conformation in the DNA strand (refer to Fig. 2) (van Zon, 2006). Naturally, this mechanism is employed by TFs that have two binding sites, for instance the *Lac* repressor. These TFs are capable of binding to both distant target sites non-specifically until they are brought into closer proximity through the formation of the DNA loop (Yesudhas *et al.,* 2017; Suter, 2020). Interestingly, the TF is able to remain transiently bound to the DNA at the point of the intersegmental transfer.

Notably, some transcription factors may employ only one of these DNA-binding mechanisms, while others may employ a combination of all three. For instance, the transcription factor TATA-binding protein (TBP) primarily uses sliding to bind to DNA (Coleman and Pugh 1995), while other transcription factors, such as the glucocorticoid receptor, use a combination of sliding, hopping, and intersegmental transfer (Oakley, and Cidlowski, 2013). Therefore, the modes of DNA binding used by transcription factors are specific to each transcription factor and can vary depending on the context of the DNA sequence they are binding to (Suter, 2020).

DNA-binding domains

DNA-binding domains (DBDs) are functional protein motifs that form part of the larger TF protein, enabling the binding of the factor to single or double-stranded DNA (Busch and Sassone-Corsi, 1990). The most prevalent types of DBDs include the following: helix-turn-helix (HTH) group, winged helix-turn-helix group, zinc coordinating group, zipper type group, other alpha-helix groups, β-sheet group, β-hairpin-ribbon group, signal transducer and activator of transcription (STAT) proteins, enzyme group, and an unclassified structure group. Differences in the structural folds of the different DBDs enable each type to recognise and interact with target DNA sequences, in distinct orientations (Atchley and Fitch, 1997).

DNA-binding mechanisms are characterised by specific interactions of the nitrogenous bases that form the DNA strand. They mostly occur at the major groove since the backbones are further apart and nucleotides are more exposed (Hamilton and Arya, 2012), with the secondary structural elements (usually α-helices) comprising the DBD (Harrison, 1991; Pabo and Sauer, 1992). The intermolecular forces that exist between the phosphate sugar backbone of the target DNA and the DBD include van der Waals contacts and hydrogen bonds (Luscombe *et al.*, 2001). These interactions stabilise the DBD-DNA interface (Pabo and Sauer, 1992).

To understand the function of specific transcription factors (TFs), it is important to first understand the structure of their DNA-binding domains. Many TFs contain helix-loop-helix and leucine zipper domains that allow them to interact with DNA through their N-terminal α-helix (Busch and Sassone-Corsi, 1990). Fig. 3 below illustrates the secondary structural features of a few DBDs that contact DNA.

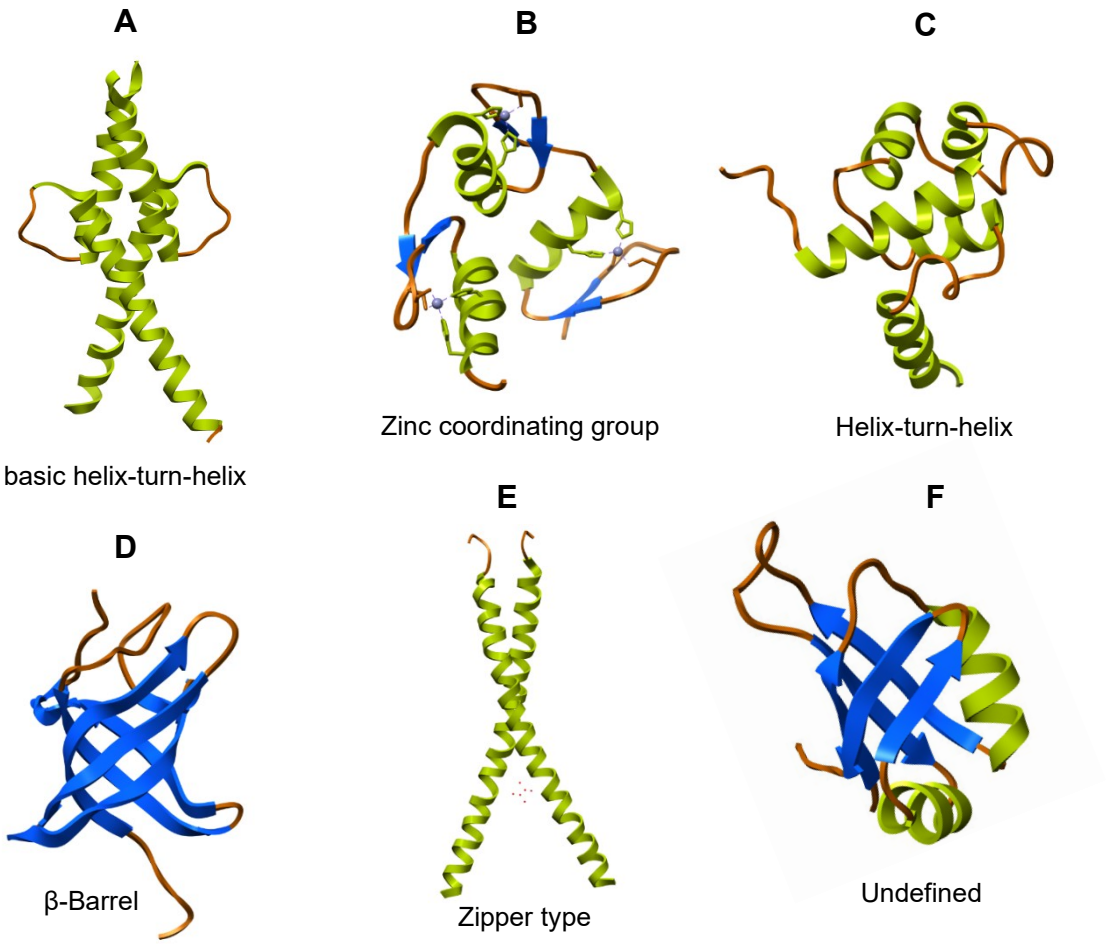


Figure 3. Schematic depicting examples of crystalised DNA-binding domains.

(**A**) basic helix-loop-helix (bHLH) (PDB: 1mdy, Ma *et al*., 1994), (**B**) zinc coordinating group (ββα-zinc finger) (PDB: 1aay, Elrod-Erickson *et al*., 1996), (**C**) helix-turn-helix (PDB: 1lmb, Beamer and Pabo, 1992)**,** (**D**) β-Barrel (PDB: 3ulj, Mayr *et al*., 2012), (**E**) zipper type (PDB: 1dh3, Schumacher *et al.,* 2000), (**F**) Undefined (PDB: 1x0f, Enokizono *et al.,* 2005). Figure adapted from: Keller *et al*., 1995**.**

1.1.2 Winged-helix transcription factors

The helix-turn-helix (HTH) family of TFs is mostly helical, with the simplest form comprising two α-helices that are connected by a short stretch of amino acids where helix 2 is regarded as the DNA recognition helix (Brennan and Matthews, 1989). The winged-helix proteins are a subfamily of the HTH transcription factors. They deviate from the larger family of HTH TFs as they include two signature 'wings', which comprise three α-helices and three β-strands and two β-sheets (Gajiwala and Burley, 2000; Kaestner *et al*., 2002).

Upon investigation of liver-specific TFs, particularly the hepatocyte nuclear factor-3 (HNF-3) family, the characterisation of winged-helix proteins emerged (Costa *et al*.,

1989). Crystallisation studies performed by Clark and team in 1993, revealed the DNA recognition mechanism of a winged-helix protein in the DBD of an HNF-3γ–DNA complex. The crystal structure of the HNF-3γ–DNA complex, now classified as FOXA1, annotated in Fig. 4., demonstrates that helix H3, which is the DNA-recognition helix, makes contact with the major groove of the DNA. The study revealed that 14 protein–DNA contacts were distributed throughout the length of the protein. Six of the observed interactions, including all the specificity determining contacts, map to the recognition helix within the major groove, with another four contacts involving wing 2 and the minor groove. Four electrostatic contacts occur at positively charged amino acids (Lys20 and Arg253) and the DNA backbone. Most of these interactions are mediated by polar sidechains, either directly or through bridging water molecules. Binding of HNF-3γ deforms the DNA by inducing a 13° bend and narrowing the major groove where it embraces helix-3.

Many winged-helix transcription factors have a patch of hydrophobic residues on their surfaces, presumed to facilitate protein-protein interactions (Clark *et al*., 1993). The site-directed mutagenesis of the winged-helix, T4 transcription factor, MotA, demonstrated that the exposure of the mutated pair of acidic/ hydrophobic residues (aspartic acid and phenylalanine) compromise its functionality. As a result, T4 growth was supressed as was transcriptional activation (Finnin *et al.,* 2003).

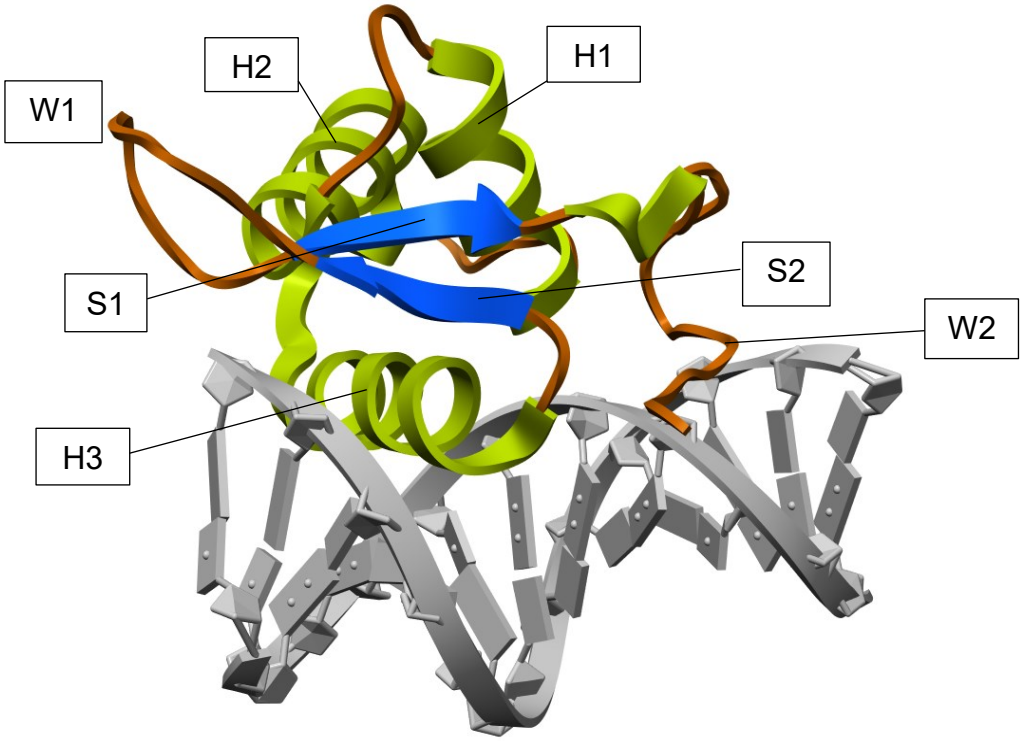


Figure 4. Diagrammatic representation of the FOXA1 winged-helix protein.

Winged-helix proteins consist of the following structural motifs: wing 1, wing 2, two or three beta-strands, and three alpha-helices. Most of the N terminal region is helical, and the C terminal region consists of the S1 and S2 beta strands (Gajiwala and Burley, 2000). These wrap to form an antiparallel beta-sheet. Wings 1 and 2 form a loop-like structure. The S1 and S2 beta strands are connected by W1; this connection extends to H3. Both W1 and W2 flank H3. Winged helix proteins are arranged as follows: H1-S1-H2-H3-W1-S2-W2. Adapted from Gajiwala and Burley, 2000 and Caputo *et al*., 2006. (PDB: 1VTN; Clark *et al.,* 1993).

As seen in Fig. 3 and 4, DNA-binding domains take on specific secondary structure conformations. How these motifs fold, directly impacts their function, i.e., transcriptional activity. The folded structure of all proteins is dictated by their amino acid sequence (Anfinsen, 1973), however, how a protein folds is also dependent on its environment. An understanding of how transcription factors fold, will reveal the stability of their DNA-binding interactions and how they occur mechanistically.

1.2 Protein folding and stability

The native form of a protein is achieved through protein folding events or mechanisms that render it active or functional and in a thermodynamically stable state. This is often the most thermodynamically favourable state it can assume (Anand and Mukherjee, 2011). The active function of a protein hinges on an accurate 3-dimensional fold or conformation, with misfolding (commonly due to sequence mutations) resulting in disease, such as the amyloid fibrils in Alzheimer's Disease (Selkoe, 2003). For transcription factors, attaining the native state means the TF will have the ability to perform routine regulation of genes, DNA-binding, co-factor recruitment, protein-protein interactions, and oligomerisation (Turjanski *et al.*, 2008).

There are four levels to protein folding: primary, secondary, tertiary, and quaternary, with each stage engaging in different levels of intermolecular forces that help to maintain the respective level of structure. At the primary level, amino acids are held together by peptide bonds (Jaenicke, 1993; Pelton and McLean, 2000). The secondary structure, comprising different structural motifs like coils, helices, and beta-strands, marks the beginning of protein folding events and these backbone structures are held together by hydrogen bonds (Shakhnovich and Karplus, 1994). The folded state of a protein refers to the complete generation of a three-dimensional tertiary structure, which is held together through hydrophobic interactions, van der Waals forces, hydrogen bonding networks, electrostatic interactions, and the solvation environment of the protein. The quaternary structure relies on the same intermolecular forces for folded protein chains to associate with other folded subunits (Anand and Mukherjee, 2011).

In the cell, a polypeptide is folded at the endoplasmic reticulum once it exits the ribosomes (Jaenicke, 1993). The process is governed by several factors and particular cellular environments that favour protein folding. These include processes that promote enthalpic gains (towards folding) and entropic losses (refolding), stabilisation through intermolecular forces, and hydrophobic interactions. As an example, water molecules forming a hydration shell around a protein, favour folding by shielding the polypeptide from degradation (Christiansen *et al.*, 2013). Because protein folding occurs in the cytosol, and there is typically a host of other interacting polypeptides and

macromolecules suspended in the cytosol (Minton, 2000), it is conceivable that many proteins within this crowded environment may become degraded or form premature non-essential interactions (Minton, 1981; Minton, 2000).

Levinthal (1969) postulated that the number of possible folded states or conformations for, arguably a 100 amino acid long protein with a possibility of 3 folded states per residue, could assume $3^{100} = 5 \times 10^{47}$ different conformations. Furthermore, if a protein can explore different conformations in the amount of time it takes for bonds to reorient (10^{13} structures per second), it would take 1.6×10^{27} years for this random search to be completed and the native structure arrived at (Levinthal, 1969; Zwanzig and Bagchi, 1992). Levinthal concluded that indeed, protein folding is not a random process but rather follows discrete pathways with intrinsic native structure fold- potential (Levinthal, 1969).

Protein folding pathways can be explained in terms of energetics. When a protein refolds (i.e., approaches a thermodynamically favourable state) it observes a decrease in Gibbs free energy (ΔG), which is continually observed until arriving at its native fold and reaching the lowest energy minimum/ barrier (at equilibrium), which is specific to the folded state (Pelton and McLean, 2000). The folding pathway of a protein may comprise many intermediate states that are stable enough to exist in a particular conformation but progress towards the final folded/ native state (this is shown in Fig. 5 below). Protein folding generating a single intermediate means that the folding rate is rather slow, as compared to the folding rate of a two-state process that generates no intermediates. The intermediate folds in a low-energy conformation; this type of pathway represents a 3-step folding mechanism (unfolded→intermediate→folded). Dynamic unfolding does not necessarily have to overcome taxing free energy barriers; the rate towards unfolding happens rather quickly and displays a two-step folding mechanism (unfolded→folded).

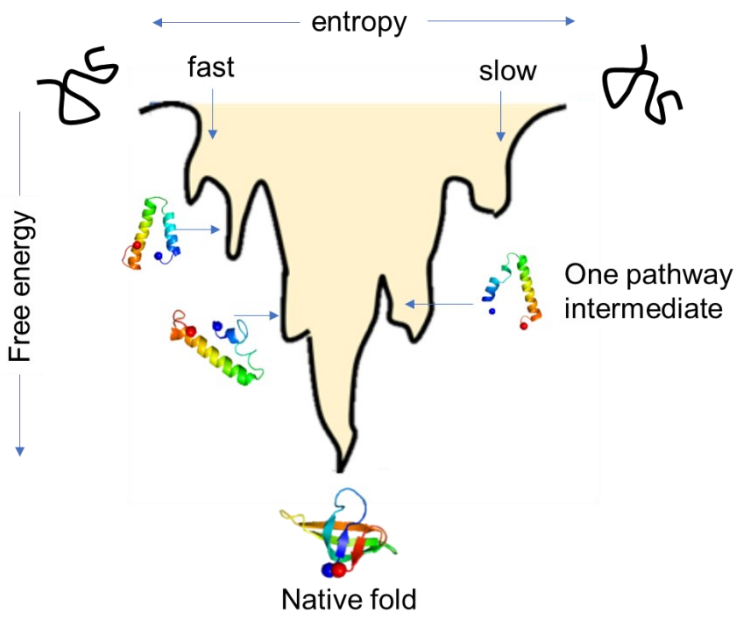


Figure 5. Protein folding funnel representing energy landscapes of the protein folding pathway.

The protein folding funnel depicts the progression from an unfolded state towards the folded state. The numerous troughs represent the folding intermediates that form along the way and illustrate that progression toward the folded state means a gradual decrease in free energy. Entropic losses occur due to the formation of secondary and tertiary structure. The folding intermediate must overcome energy barriers on its course, which affect the rate of folding [Schematic created originally in Canva].

1.3 Oligomerisation

Many research efforts, within the field of structural biology, are geared towards understanding protein-protein interactions, why they occur, the chemical specificity of such interactions, and their cellular functions. The effects of protein complexation range from the regulation of enzymatic activity to the regulation of ion channels, receptor functioning, and the regulation of gene expression (Hashimoto and Panchenko, 2010).

Why would two or more proteins associate at all? Research suggests protein oligomerisation stems from evolutionary events that may have selected for (directly or indirectly) protein associations (Pereira-Leal *et al.*, 2007). Gene duplication events may have led to the formation of oligomeric paralogs, subsequently giving rise to the formation of new and unique protein complexes (Dayhoff *et al.*, 2010). At the genomic level, insertions, deletions, and substitutions play an important evolutionary role in selecting for protein complexation (Pereira-Leal and Teichmann, 2005). The removal

or addition of certain amino acids in different monomeric protein paralogs, could, for instance, promote protein-protein association if new hydrophobic patches are introduced- thus creating a potential interface for oligomerisation to occur (Pereira-Leal *et al*., 2007). Similarly, recombination events could have caused a monomeric protein, once lacking an oligomerisation domain, to fuse with one that does possess an oligomerisation domain and in doing so, creating new oligomerisation potential (Nussinov *et al*., 1998; Levy *et al*., 2008). Protein oligomerisation, viewed in light of protein evolution, is advantageous in that it creates new prospects for allosteric and functional regulation (Ali and Imperiali, 2005).

There are numerous additional advantages for protein oligomerisation. Larger proteins for instance are less prone to degradation, have increased stability and due to the reduced surface area of the monomer in the complex, they are protected (to a certain extent) against degradation (Pereira-Leal and Teichmann, 2005). Protein complexation may have the added benefit of incorporating additional active sites, which supports more efficient cellular processes and increased regulatory flexibility (Dayhoff *et al*., 2010). Discriminating between specific processes may require higher-order complexation, which can confer the regulation of energy expenditure and control of the conditions that govern oligomerisation (Ali and Imperiali, 2005). A great 'gain' in protein oligomerisation is the fact that proteins can assemble to form large complexes, without the genome size having to expand. Furthermore, a large protein comprising several, short subunits is less likely to have errors within the protein sequence as compared to a large protein synthesised from a single polypeptide (Pereira-Leal *et al*., 2007). Thus, it seems biologically favourable that oligomeric proteins be formed. However, not all proteins oligomerise. Transcription factors in particular employ oligomerisation to execute transcriptional activity and DNA-binding.

1.3.1 Oligomerisation in transcription factors

Several families of transcription factors undergo homo- and hetero-oligomerisation to bind DNA. Depending on the cellular environment, these TF oligomers initiate regulatory events and determine cell fates (Amoutiaz *et al*., 2008). Cell proliferation, metabolic processes, and apoptosis are a few processes that are initiated by the oligomerisation of TFs (Dayhoff *et al*., 2010). TFs may exist as predominantly homo-oligomeric structures in nature (Pabo and Sauer, 1992), with homodimeric and

heterodimeric TFs accounting for more entries in protein databanks than TFs with higher stoichiometries (greater than 2). This is largely due to difficulties in experimental procedures, as trying to isolate higher-order oligomeric species tends towards aggregation (Hashimoto and Panchenko, 2010). There exist two pathways in which homotypic and heterotypic associations can occur. Two TF monomers may first have to associate to form a dimer, before they can bind DNA (called the dimer pathway, as seen in Fig. 6), alternatively, the TF monomers may bind DNA first and then associate to form a dimer (called the monomer pathway) (Kohler and Schepartz, 2001). In addition to oligomerisation-induced DNA-binding, transcription factors may also oligomerise to regulate nuclear trafficking, impose inhibitory and regulatory effects and accommodate for changes in DNA-binding specificities (Dayhoff *et al*., 2010).

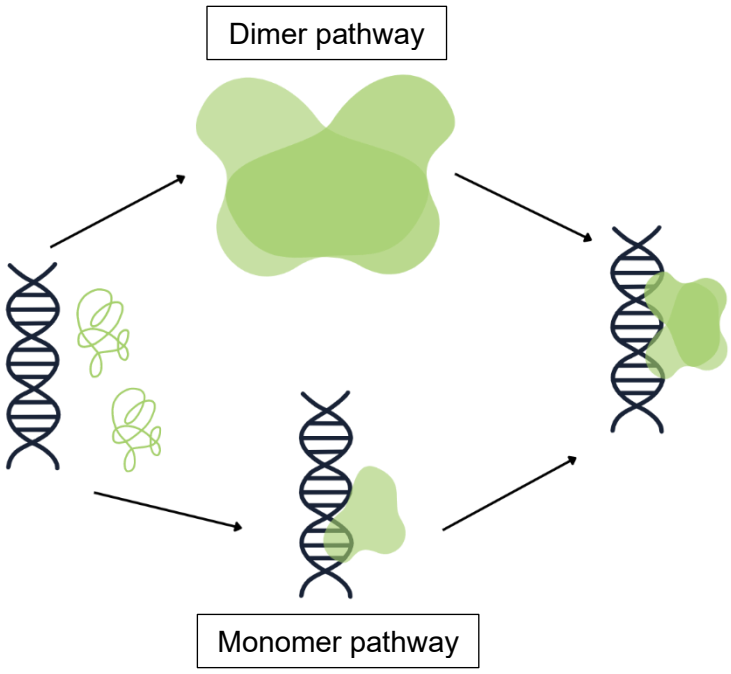


Figure 6. Illustration of the monomer and dimer DNA-binding pathways in transcription factors.

To favourably arrive at the ternary complex, two DNA binding pathways may be employed, that is, the monomer pathway or the dimer pathway. In some cases, as a prerequisite to DNA binding, dimeric proteins must form a ternary complex before binding DNA (monomer pathway), that is, one DNA molecule is bound by two protein molecules (a dimer - represented by the green globules). In the dimer pathway, the transcription factor must first dimerise and then associate with DNA. The monomer pathway involves the independent association of monomer transcription factors to the DNA first. This is the followed by dimerisation at the DNA binding site. Adapted from: Kohler and Schepartz, 2001.

Transcription factor families including the basic helix-turn-helix (bHLH), mini-chromosome maintenance 1, agamous, deficiens and serum response factor (MADS-box), homeodomain leucine zipper (HD-ZIP), and the basic leucine zipper (bZIP) are a few examples of TF families whose homotypic associations have been well studied (Amoutiaz *et al*., 2008). Homodimerisation of these fore-mentioned TF families regulates cell processes such as homeostasis, reproduction, the cell cycle, immunity, and apoptosis (Levy and Darnell, 2002; Desvergne *et al*., 2006). These transcription factors all contain specialised domains, called dimerisation domains, that ensure the formation of dimeric associations (Refer to Fig. 7). For instance, the bHLH TFs comprise alpha-helical basic regions that bind bHLH target DNA as a dimer (Amoutiaz *et al*., 2004). Near the C-terminal, the HLH region forms a dimer consisting of four, 12 bundled α-helices (Amoutiaz *et al*., 2004). Dimerisation domains such as the Per-Arnt-Sim (PAS) and leucine zipper domains may also be found in bHLH TFs and are responsible for conferring specificity of the associations (Amoutiaz *et al*., 2004). In the MADS-box family of proteins, dimerisation domains are also located at the C-terminal (adjacent to the MADS domain, which is a DNA-binding domain) (de Folter *et al*., 2005). In MIKC TFs (a subtype of MADS-box proteins) dimerisation is conferred via the I domain, whilst the K domain (also a dimerisation domain) confers specificity via coiled-coil formations (West and Sharrocks, 1999; de Folter *et al*., 2005). In the bZIP family of TFs, the α-helical regions associate to form a dimer, before recognising a hexanucleotide DNA core, whilst the leucine zipper domain directs dimerisation activity (Vinson *et al*., 2006). Thus, there are several well-characterised examples of transcription factors that use oligomerisation to perform DNA regulation and other activities.

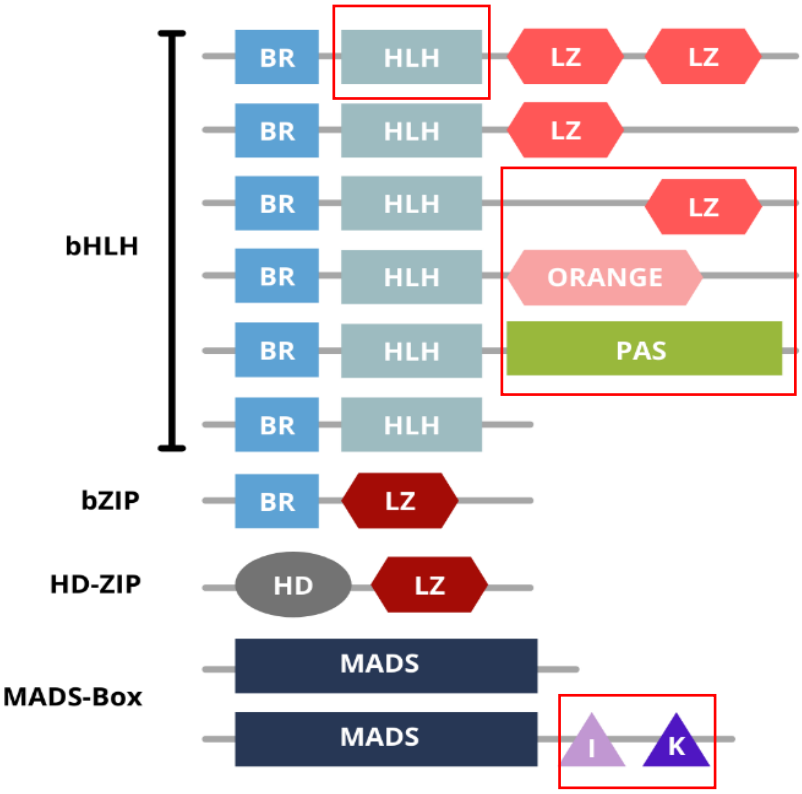


Figure 7. Schematic depicting the different oligomerisation domains found in transcription factors.

Dimerisation domains such as HLH, LZ, orange, PAS, I, and K (highlighted by red boxes) are shown for each TF superfamily depicted: bHLH, bZIP, HD-ZIP, and MADS-Box. The other domains represent the DNA-binding domains of these TF superfamilies, although the LZ domain does display DNA-binding functionality in most transcription factors. Adapted from: Amoutiaz *et al*., 2008.

1.4 FOX Transcription Factors

The forkhead box (FOX) proteins stem from the helix-turn-helix (HTH) family of transcription factors and each possesses a unique DNA-binding domain called the forkhead domain (FHD) (Weigel *et al*., 1989) These transcription factors are essential for routine cellular functioning, organogenesis, and determining cell fates (proliferation and apoptosis). They also have major implications in pathology and have been linked to metabolic processes and immune system defences (Katoh and Katoh, 2004; Jackson *et al*., 2010; Sin *et al*., 2015). Furthermore, FOXA, FOXM, FOXO, FOXC, and FOXP proteins are key players in tumour suppressive and oncogenic pathways (Lam *et al*., 2013).

There exist nineteen FOX subfamilies (FOXA – FOXS) and each subfamily shares a conserved forkhead domain (FHD) (Hackett, 2002). The FOX FHD stems from the

HTH family of DNA binding domains but deviates from the typical HTH DNA-binding domains in that the three α-helices comprising the FOX FHD are complexed with β-strands (Aravind *et al*., 2005). This forms the so-called “winged helix” fold which is a derivative of the HTH fold. The typical FOX FHD is 90 amino acids long and constitutes three α-helices, three β-strands, and two wings that flank the third β-strand (Lee and Frasch, 2004).

FOX TFs are reported to recognise the general DNA sequence motif, 5’- RYAAAYA -3’ (where R represents a purine and Y, a pyrimidine) (Chen *et al*., 2015). Within the FOX family of TFs, there seems to be great overlap concerning the DNA sequences that each member binds. This overlap was observed in DNA-binding studies involving FOXK3, FOXO3, and FOXJ3 (Chen *et al*., 2015). This led to the proposal of a DNA-binding model for FOX TFs, which hypothesises that dynamic equilibrium complexation may be a requirement for gene-specific regulation (Chen *et al*., 2015). However, this form of FOX TF complexation is yet to be pursued *in vivo*.

1.4.1 FOXP Transcription Factors

The FOXP subfamily (inclusive of FOXP1, FOXP2, FOXP3, and FOXP4) is involved in several regulatory pathways. Gene expression in the lung, brain, spleen, and gut is regulated by FOXP1, FOXP2, and FOXP4 (Lu *et al*., 2002). FOXP3 plays a role in autoimmunity and is primarily expressed in the thymus, where it regulates responses to regulatory T cells (Tregs) (Hori *et al*., 2003; Fontenot *et al*., 2003). FOXP4, expressed in thymocytes and peripheral CD4+ and CD8+ T cells, plays a role in T-cell development, and its deletion in the early stages of embryogenesis can be fatal (Li *et al*., 2004; Wiehagen *et al*., 2012). FOXP1 is a transcriptional repressor, involved in the development and function of several organs and tissues, and immune system regulation. It may also serve as a tumour suppressor as its expression has been observed in tumour cells (Banham *et al*., 2001, Shu *et al.,* 2001, Hu *et al.,* 2006).

All members of the FOXP subfamily contain a poly-glutamine (poly-Q) region (except for FOXP3) zinc finger (ZF) domain, leucine zipper (LeuZip) domain, a conserved forkhead domain (FHD) (Shu *et al*., 2001; Hackett, 2002; Häußermann *et al*., 2019) as shown in Fig. 8 below.

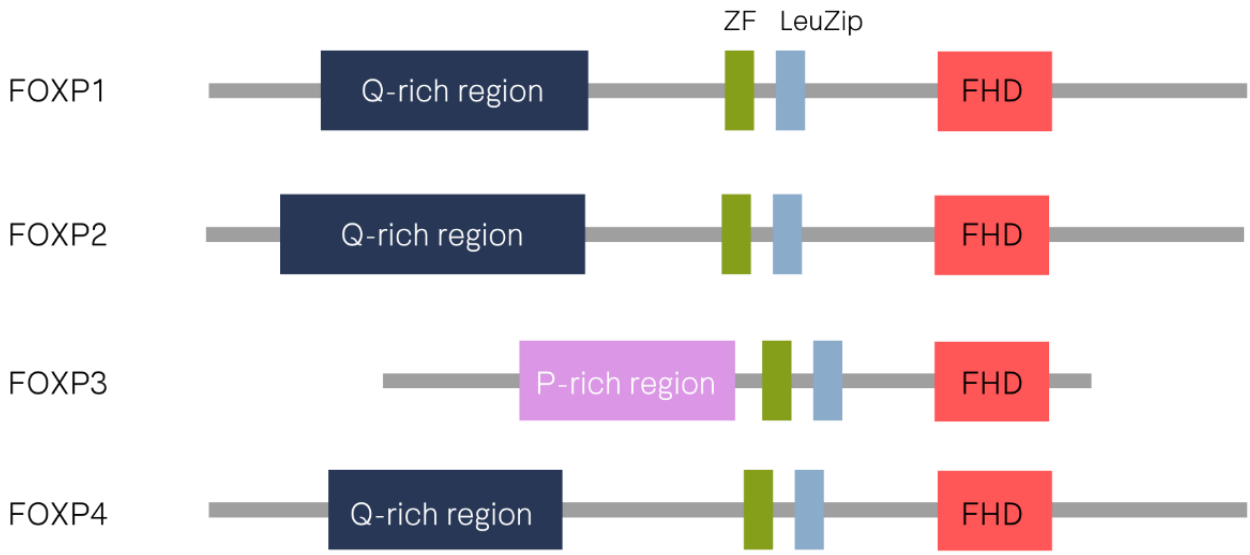


Figure 8. Schematic illustrating the FOXP subfamily domains.

The Q-rich region (dark blue) is prevalent in all subfamilies, excluding FOXP3, where it is replaced by a proline-rich region (purple). The zinc finger domain (green) and leucine zipper domain (light blue) precede the conserved DNA-binding, forkhead box domain (red).

Additionally, the FOXP subfamily FHD is located near the C terminus (other FOX proteins bear the FHD near the N terminus) (Sin *et al*., 2015). The poly-glutamine region is speculated to be involved in protein-protein interactions, which may be facilitated by its elongated helical structure; however, its precise role is still questionable within the FOXP subfamily (Lai *et al*., 2001; Schön *et al*., 2006 and Häußermann *et al*., 2019). Not much is known about the role of the zinc finger domain in the FOXP subfamily; however, these motifs generally facilitate DNA-binding in many transcription factors and have also been implicated in stabilising protein folds (Brown, 2005; Miller *et al*., 2010). The LeuZip domain has received a lot of attention in recent years concerning its function within the FOXP subfamily. It is hypothesised that the LeuZip domain could bind DNA, but much work is still to be done concerning its affiliation to DNA in the FOXP subfamily (Li *et al*., 2004). It has, however, been implicated in FOXP1/2/4 homo- and heterodimerisation (Li *et al*., 2004; Chae *et al*., 2006; Thulo *et al*., 2021). Extensive research has shown that the FHD is accredited with the specific binding of DNA (Wang *et al*., 2003; Stroud *et al*., 2006; Webb *et al*., 2017).

1.4.2 FOXP Oligomerisation

It is important to note that only the FOXP subfamily of FOX TFs can dimerise. The significance of this divergence from other FOX transcription factors is yet to be fully elucidated. There are two dimerisation interfaces in the FOXP subfamily, the leucine zipper domain and the forkhead domain. Studies by Li and company in 2004, describe a unique dimerisation mechanism of the FOXP subfamily (more specifically FOXP1, FOXP2, and FOXP4). These studies suggest that FOXP1/2/4 exhibits dimerisation (both homo and hetero) via the LeuZip domain (Li *et al*., 2004). The deletion of the LeuZip results in the inability of FOXP1/2/4 to bind DNA, however, with the addition of a dimer forming glutathione S-transferase tag, the DNA-binding properties were restored (Li *et al*., 2004). Studies by Chae and company in 2006 showed that the FOXP3 protein could also dimerise via the LeuZip and the necessity for FOXP3 dimerisation was further confirmed in studies that observed reduced DNA-binding activity in the FOXP3 protein when dimerisation had been stifled (Song *et al*., 2012). Such experiments suggest that FOXP dimerisation could exist physiologically and that effective DNA-binding could be dependent on FOXP oligomerisation. Furthermore, higher-order complexation was observed for murine FOXP3 (Song *et al*., 2012). Here, the LeuZip was implicated in transient dimer formation, which resulted in a tetrameric complex and higher-order complexes (Song *et al*., 2012).

The forkhead domain in the FOXP subfamily displays distinct dimerisation capabilities that are absent in other FOX proteins, in that it forms a domain-swapped dimer *in vitro.* This is achieved through the exchange of helix 1 and helix 2 of the FHD monomers and is only possible through the flexibility of the hinge-loop region (Stroud *et al.,* 2006, Fanucchi & Morris, 2016). Notably, FOXP1 and FOXP2 may exist as both a monomer and a dimer, although FOXP3 exists solely as a dimer (Chu *et al.,* 2011). The formation of the domain-swapped dimer has only really been observed between homo-oligomeric monomers, whilst the leucine zipper domain has heterotypic oligomerisation potential (Li *et al.,* 2004; Stroud *et al.,* 2006)

1.5 FOXP2 Transcription Factors

The FOXP2 transcription factor is essential for early embryonic development of the brain (Lai *et al*., 2001). The FOXP2 gene is associated with the acquisition of language and language-related disorders such as developmental verbal dyspraxia (DVD) and autism and has even been associated with cancer (Vernes and Fisher, 2013). The FOXP2 gene was first discovered in a British family (known as the KE family) whose members suffered from hereditary DVD (Belton *et al*., 2003). A single missense mutation that resulted in the substitution of an arginine residue by a histidine residue, was perpetuated across many generations of the KE family (Belton *et al*., 2003; Lai *et al*., 2003). The mutation had altered the functioning of FOXP2 and as a result, the family endured the effects of DVD (severe speech impairment as a consequence of orofacial myofunctionality and flawed comprehension) (Belton *et al*., 2003; Feuk *et al*., 2006). As previously mentioned, all FOXP members contain a poly-Q domain, ZNF domain, LeuZip domain, and FHD. The domains comprising FOXP2 are shown in Fig. 9 below.

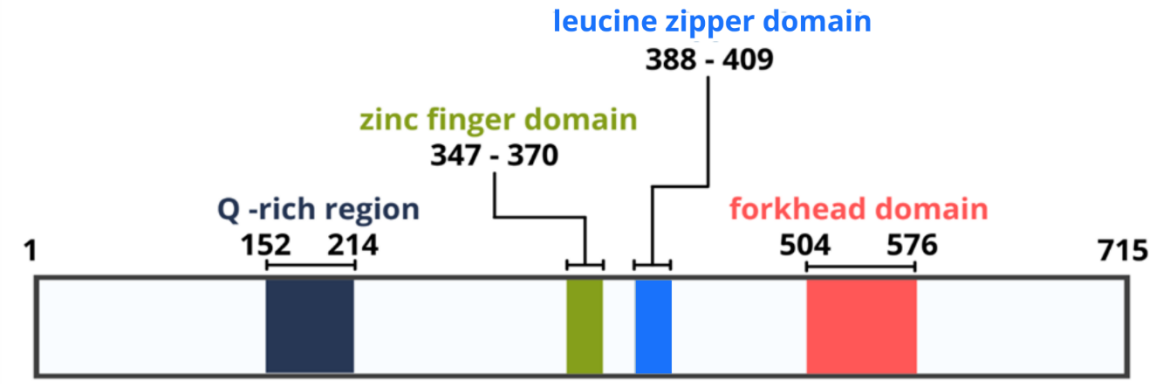


Figure 9. Schematic illustrating the FOXP2 domains for the full-length FOXP2 protein (1 to 715 amino acids).

The polyglutamine region (navy), the zinc finger domain (green), the leucine zipper domain (blue), and the forkhead domain (orange) are shown from N- to C terminals.

1.5.1 FOXP2 Forkhead Domain

The DNA binding domain, the forkhead domain of FOXP2 is a highly conserved winged-helix-turn-helix structural motif that is approximately 91 amino acids long and is the only region of the FOXP2 structure that has a solved crystal structure (Stroud *et al*., 2006). It contains three α-helices and two conserved loop-like structures or flanking wings. The FHD core comprises of three α-helices (H1, H2, and H3) bordered off by

an antiparallel β-sheet, made up of three strands (S1, S2, and S3). The recognition helix, H3, interacts specifically with the major groove of DNA. A fourth helix (H4) is found between helices 2 and 3 (Clark *et al*., 1993; Weigelt *et al*., 2001). A 'wing' motif resides between strands 2 and 3 in all FOX proteins except for FOXP2 which, has a shortened motif that forms a simple turn (instead of a loop) that connects strands 2 and 3 (Stroud *et al*., 2006). Additionally, in the FOXP2 FHD, a fifth helix (H5) serves as a 'wing' motif (seen in other FOX TF's) that spans the top of helix 1 and the DNA phosphate backbone (Jin *et al*., 1999).

1.5.2 FOXP2 oligomerisation

While most transcription factors display oligomerisation to execute their regulatory functions, the FOXP subfamily is unique amongst the FOX proteins in their ability to form oligomers, whereas other FOX proteins tend to be monomeric. However, oligomerisation events surrounding the FOXP2 transcription factor have not been distinctively established. Research on the potential oligomerisation activities of FOXP2 is always centred around the FHD and LeuZip domain.

1.5.2.1 Forkhead domain oligomerisation

The FHD displays a rather complex mechanism of associating two FOXP2 monomer interfaces, *in vitro*. The mechanism of FHD dimerisation is known as domain swapping (Stroud *et al*., 2006). According to this mechanism, two FOXP FHD monomers can form an intertwined FHD dimer by exchanging H3 and strands 2 and 3 (Stroud *et al*., 2006; Chen *et al*., 2015). Within the FOX family of TFs, the Pro539 residue is conserved, with the FOXP subfamily being the only exception. Here, proline is replaced with an alanine (Stroud *et al*., 2006; Chen *et al*., 2015). This substitution confers flexibility to the loop that connects helices 2 and 3 of the FHD (Stroud *et al*., 2006; Chen *et al*., 2015). Had this substitution not emerged, the occurrence of the FOXP2 FHD domain-swapped dimer (see Fig. 10) may not have been possible. Whilst dimerisation has been observed, at least *in vitro*, for the FOXP2 FHD, no other studies have probed or observed higher order formation (beyond dimer formation) for the FHD in isolation or complex with DNA. Newer developments suggest the acid-rich tail to interact specifically with the forkhead domain (Thulo *et al.,* 2019). This interaction may affect oligomerisation and DNA-binding.

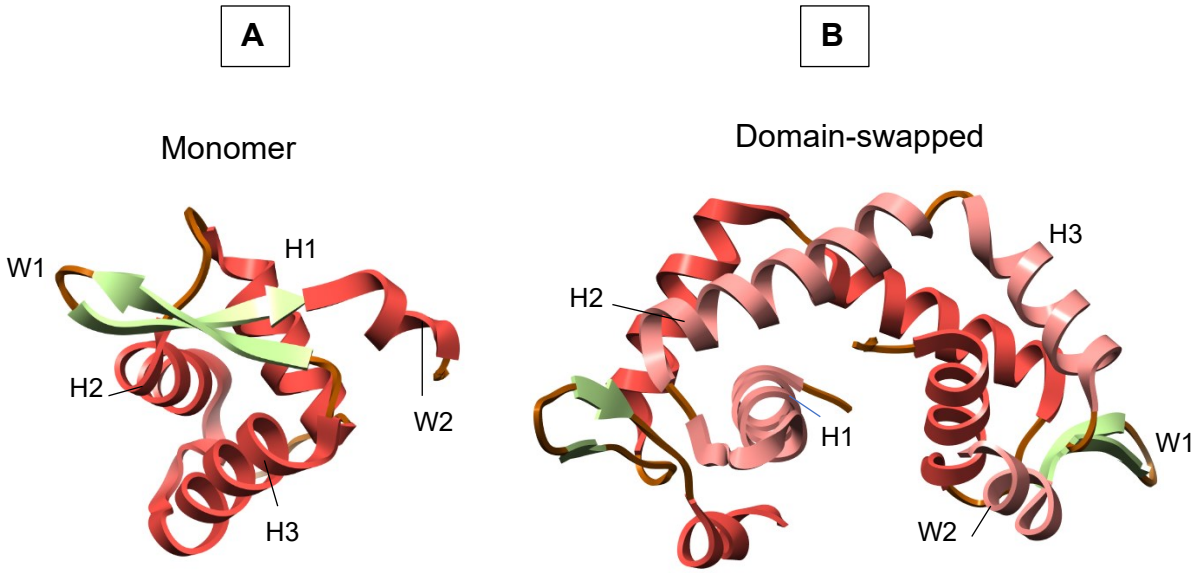


Figure 10. Illustration of the FOXP2 forkhead domain monomer and domain-swapped dimer.

(A) the FHD monomer comprises three α-helices (H1, H2, H3) and a single β-sheet comprising two antiparallel β-strands that are linked by a wing motif (W1). **(B)** the FHD domain-swapped dimer is formed by the extension of H3 and strands 2 and 3, one monomer is shown in pink and the other in red.

Further studies in this work, particularly focussing on Trp537 which is found at the protein: DNA interface, imply that the wt FHD does not insert as deeply into the major groove of DNA, as compared to the mutant. It might be possible, that the wild-type, due to its more flexible conformation, may not reside with the DNA for as long as the monomeric mutant does and as a result, does not insert as deeply into the major groove of the DNA, therefore resulting in a weakened binding-affinity. Perhaps the leucine zipper domain could offer more stability and facilitate deeper insertions into the major groove- this is yet to be explored.

1.5.2.2 Leucine zipper oligomerisation

As previously mentioned, FOXP2, as well as the other members of the FOXP subfamily has been shown to display homo- and heterodimerisation (Li *et al*., 2004). The LeuZip was accredited to these dimerisation events that were observed, as dimerisation was abolished when the LeuZip was interfered with (Li *et al*., 2004).

Whilst leucine zipper oligomerisation is speculated to exist, there is no clear mechanism detailing its contribution to oligomerisation events within FOXP2. The proximity of the LeuZip to the FHD raises questions about its spontaneous involvement in the dimerisation of FOXP2 when an FHD domain-swapped dimer form. Perhaps the leucine zipper plays a passive role in aiding dimerisation in FOXP2, but this has yet to be explored. The role of the FOXP3 leucine zipper has been more well-defined. Mutations in the LeuZip domain of FOXP3 have been shown to disrupt homo-associations and impede Treg function both *in vivo* and *in vitro* (Chae *et al*., 2006; Lopes *et al*., 2006; Song *et al*., 2012). Furthermore, it was discovered that an intact leucine zipper was required for effective FOXP3 FHD DNA binding. This antiparallel, coiled-coil domain is reported to have a flexible conformation that tightens when forming a dimer, necessary to bind DNA with a high affinity (Song *et al*., 2012). Where FOXP2 is concerned, studies by Häußermann and team in 2019, discovered that the FOXP2 LeuZip is essential for FOXP2 oligomerisation. The researchers compared the binding constant (K_D) of the FOXP2 variant possessing an intact LeuZip domain and FHD with a FOXP2 variant baring a broken leucine, heptad repeat. The results showed drastic changes in K_D, with the protein with the intact LeuZip yielding a K_D (protein-protein) of ~ 2 nM, whilst the variant with the broken heptad repeat yielded a K_D (protein-protein) of ~ 300 nM. Disrupting the leucine, heptad repeat sequence in the LeuZip of FOXP2 was therefore shown to destabilise the homo-association by ΔΔG ~3 kcal/mol.

1.6 DNA and FOXP2

1.6.1 DNA-binding associated with FOXP2

DNA-binding in FOXP2 is linked directly the forkhead domain, which directs the specificity of the interaction to the target sequence. The DNA-binding consensus sequence, recognised by FOXP2, is 5'-CAAATT-3'. Sequence-specific FOXP2 DNA-binding is peculiar in that van der Waal's forces stabilise the contacts to a great extent and hydrogen bonding at the major groove is minimal. This could confer binding of different non-consensus DNA sequences, as long as the minimum number of hydrogen bonds are present at the core (to stabilise the FOXP2-DNA complex) and the potential DNA-binding sequence shape is complementary to the FOXP2 interface (Stroud *et al*., 2006).

Studies by Webb and team in 2017 investigated the affinities with which the FOXP2 FHD binds DNA and revealed that high-affinity sequences remained bound to the FHD for longer periods as compared to low-affinity sequences. Furthermore, increased FHD-DNA interactions were accounted for, along with observable structural changes, when the FOXP2 FHD was bound to high-affinity sequences. FOXP2-DNA-binding had been categorised according to sequence affinity. On one extreme, FOXP2 was suggested to rapidly scan the genome, where low-affinity sequences occurred but bind more securely at high-affinity sites to induce conformational changes, facilitating functional binding.

Studies by Stroud *et al.,* in 2006 resolved the FOXP2, forkhead domain crystal structure which suggests that FOXP2 binds DNA via the FHD as one of two complexes, either as a monomer or a dimer (refer to Fig. 11). However, this is yet to be definitively confirmed as the domain-swapped dimer may occur as an artefact of crystallisation. The study also proposed that monomeric FOXP2 binds DNA more tightly than the FOXP2 dimer (Stroud *et al.,* 2006).

The FHD crystal structure reveals the FHD DNA-binding interface and shows how helix 3 interacts with DNA, with key amino acids facilitating hydrophobic interactions and hydrogen networking with DNA (Stroud *et al*., 2006). Certain stabilising amino acids interact with helix 1 and 2 to help wedge the recognition helix within the major groove of the DNA (Stroud *et al*., 2006). The interface between FOXP2 and DNA is further stabilised by Thr508, Arg583, Arg584, and Tyr509 (located at the end of the N-terminal of helix 1), which display electrostatic interactions with DNA, as well as hydrogen bonding and van der Waal's contacts (Stroud *et al*., 2006).

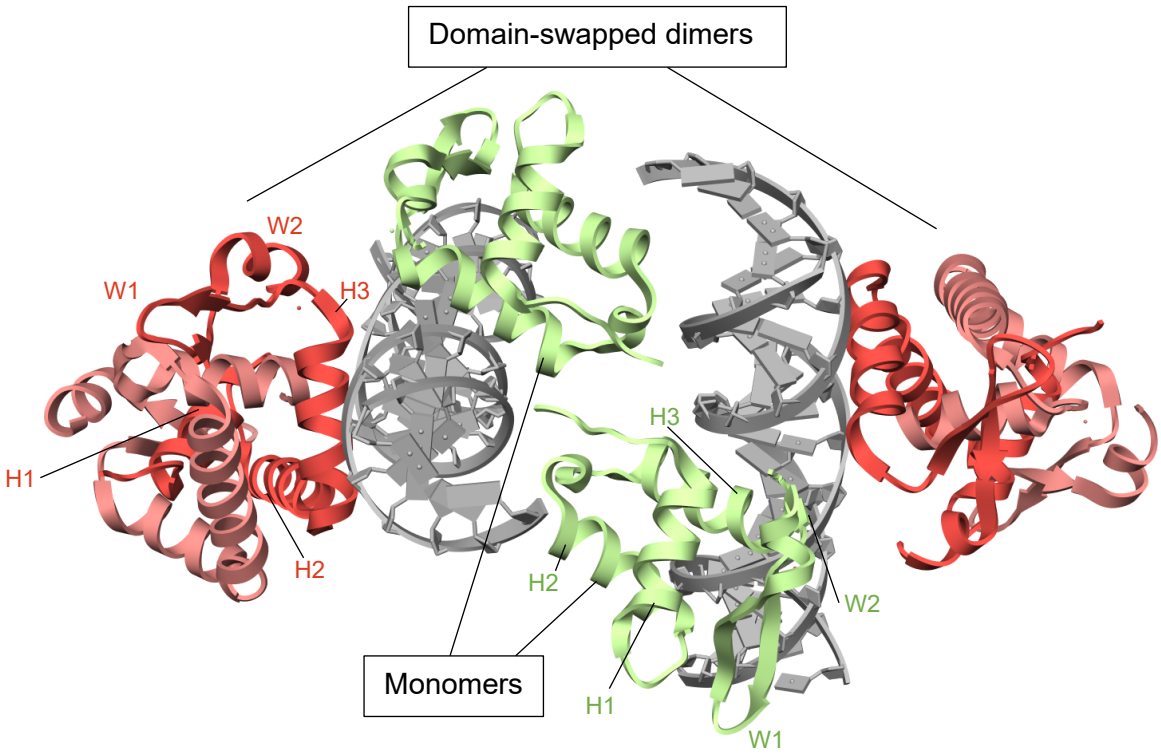


Figure 11. Structure of the FOXP2 FHD bound to DNA.

The crystal structure displays 6 copies of FHD in complex with DNA, 4 of which form domain-swapped dimers (orange) and the remaining two copies exist as monomeric forms of the FHD (green) (PDB code: 2A07). Adapted from: Stroud *et al*., 2006.

The DNA-binding properties of the FOXP2 FHD were characterised in a study that used site-directed mutagenesis to prevent the reversibility of the monomeric (A539P FOXP2 FHD variant) and dimeric (F541C FOXP2 FHD variant) forms of the FOXP2 FHD. This was done so that each state could be evaluated in terms of its DNA-binding properties (Morris and Fanucchi, 2016). The F541C variant was found to bind DNA as both a dimer and a monomer, whereas, the wild-type FOXP2 FHD was found to bind DNA, exclusively, as a monomer (Morris and Fanucchi, 2016). Increased affinity and entropy of binding were observed for the wild-type FHD as compared to the A539P variant (monomer) and the wild-type FHD was found to have greater conformational flexibility in the hinge-loop region (Morris and Fanucchi, 2016). Interestingly, an

increase in conformational freedom was accounted for in the FHD when the leucine zipper was introduced, which was suggested to promote dimerisation when the FHD would bind DNA (Morris and Fanucchi, 2016).

Comparing the structural folds of certain FOX TFs (See Fig. 12) with FOXP2 shows some variation and some similarities in wing 1, the C-terminal regions, and the H2-H3 turn. The FOXP2 structure forms short H2–H3 α-helical turns, whereas these regions form coil structures in the other FOX TFs (Tsai *et al*., 2007). The amino acid composition and length of the wing 1 motif vary among FOX TFs and is comparatively shorter in the FOXP2 protein. This results in a shortened turn that does not associate with DNA at all. The FOXP2 forkhead domain forms an α-helix at the C-terminal end, whereas FOXO3a and FOXA3 form coiled structures (Tsai *et al*., 2007). Even though the structure of the C-terminal regions FOXP2 varies considerably among the other FOX TFs, it is still positioned in a particular orientation that facilitates DNA-binding (Tsai *et al*., 2007).

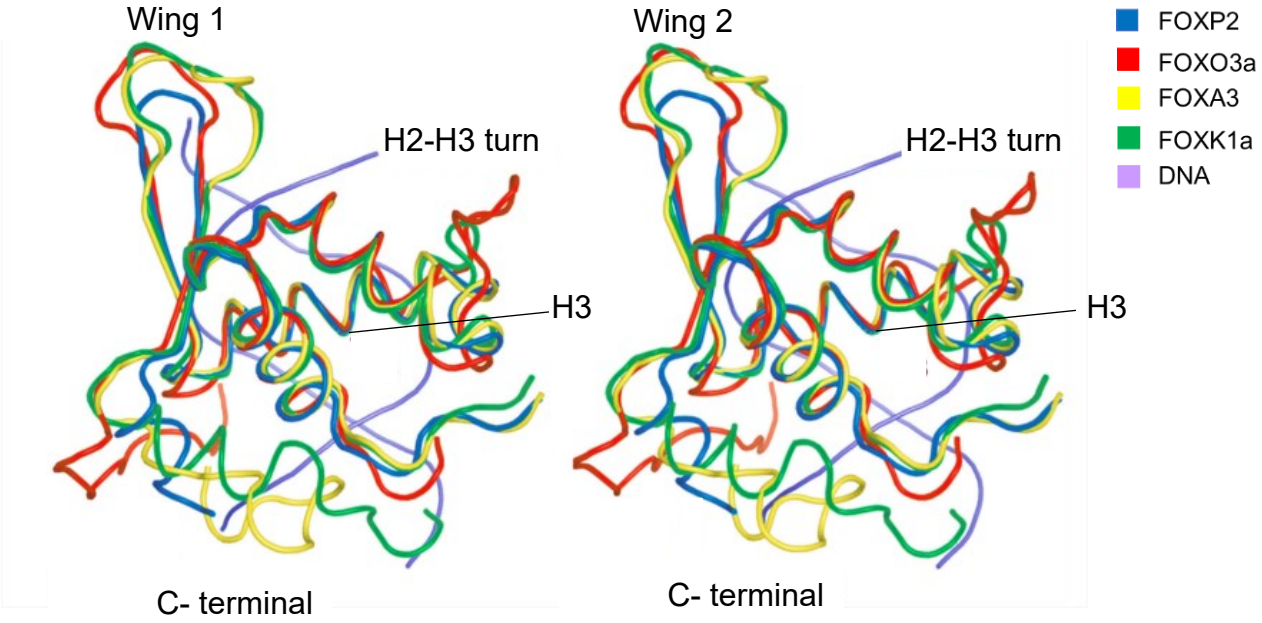


Figure 12. Superimposition of FOX transcription factors to highlight structural differences and similarities upon DNA-binding.

FOXA3, FOXK1a, FOXP2, and FOXO3a are shown in yellow, green, blue, and red respectively and DNA is shown, in purple. [Adapted from: Tsai *et al*., 2007]

Since the recognition helix (H3) in FOXP2 (and other members of the FOXP subfamily) is short in comparison to other FOX TF's and DNA-binding affinity of the FHD is relatively weak (Li *et al*., 2004), it is likely that higher-order complexation of FOXP2 or protein-protein interactions is required for specific DNA-binding (Stroud *et al*., 2006). Further investigations by Webb *et al*. in 2017 suggests that the structure of FOXP2 is subject to change when complexed with high-affinity DNA sequences. Whether the FOXP2 oligomeric state could also be influenced by such DNA sequences is not known.

1.7 Rationale

Interestingly, FOXP proteins have evolved to multimerise while their other family members did not. One might question what the significance would be of this oligomerisation event. It is likely that since proteins (transcription factors included), tend to oligomerise to increase their stability, the formation of FOXP2 oligomers will stabilise the protein structure to allow it to perform its function effectively.

It is important to reiterate that the FOXP subfamily are the only FOX TFs that have been shown to dimerise via the FHD through key evolutionary events. Furthermore, the FOXP proteins contain a LeuZip interface that likely further facilitates some form of oligomerisation. This divergence from other FOX TFs, suggests that dimerisation may be critical in directing the specificity of the role and function of the FOXP subfamily.

Considering oligomerisation of other families of TFs is a common means of regulating transcription, it is plausible that the FOXP subfamily may be employing a similar mechanism of regulation through oligomerisation. Indeed, if FOXP2 regulated transcription of its target genes through oligomerisation, then its oligomeric state would be directly linked to its interaction with DNA. While work has been done on the oligomerisation of apo FOXP2, which showed increased levels of complexation for various FOXP2 constructs (Häußermann *et al.,* 2019; Thulo *et al.,* 2021), there is no current literature available on the oligomeric state in the presence of DNA. Knowledge of this would directly impact our understanding of the mechanism of action of FOXP2.

In light of this, an investigation into the stability of FOXP2 in the absence and presence of DNA, in conjunction with information on the oligomeric quaternary structure, would give critical insight into the importance of the DNA and hence the DNA-binding

interaction on the structure, folding, and stability of the protein. This would allow the mechanism of DNA-binding to be more clearly defined and information on whether or not the protein needs to unfold before binding DNA can be established.

Apart from the potential oligomerisation capacity of the leucine zipper, its structural relevance and contribution to the FOXP2 protein is not clearly defined. Indeed, no solved structure exists for the full length FOXP2 protein and only the forkhead domain has been characterised in this way. Thus, investigating the dynamics of the leucine zipper domain needs to be achieved and compared to that of the forkhead domain to understand how both entities contribute to DNA-binding.

Thus, a comprehensive investigation, considering all factors that contribute to FOXP2 structure and its mechanistic function is what is required to elevate and broaden our understanding of FOXP2 structure and DNA-binding.

Chapter 2: Aim and Objectives

2.1 Aim

The effect of DNA on the oligomerisation of FOXP2 has not been fully considered or defined, especially at the FOXP2 oligomerisation interfaces, that is, the leucine zipper domain and the forkhead domain. It has been established that the forkhead domain functions as the FOXP2 DNA-binding domain and has been shown to bind DNA as both a monomer and a domain-swapped dimer (*in vitro*) (Stroud *et al.,* 2006). Indeed, the formation of the domain-swapped dimer is contestable since it could be the result of a crystallography artefact. The leucine zipper domain is associated with dimerisation potential, as it forms homo- and heterodimers with its FOXP counterparts (Sin *et al.,* 2015). Thus, what still needs to be assessed in the leucine zipper domain and the forkhead domain is the structural implications of having DNA involved. Determining the structural changes that occur at both these domains is necessary to understand the global (FOXP2) and local (domain contribution) significance of these structural changes and the effect this has on the DNA-binding mechanism.

Therefore, the aim of this study was to gain a comprehensive understanding of the FOXP2 DNA-binding mechanism by comparing the DNA-binding capabilities, stability, and dynamics of both the leucine zipper domain and the forkhead domain in the presence and absence of DNA.

2.2 Objectives

- Overexpress protein using a bacterial recombinant expression system and purify the two FOXP2 constructs (LZ-END and FHD) for structural study analyses.
- Investigate and compare the LZ-END and FHD DNA-binding capabilities by assessing binding affinities through fluorescence anisotropy.
- Assess and compare the LZ-END and FHD fold stability via urea-induced equilibrium unfolding and spectroscopic techniques.
- Use hydrogen deuterium exchange mass spectrometry to assess protein dynamics in the presence and absence of DNA

Chapter 3: Materials and Methods

3.1 Experimental constructs

3.1.1 Protein constructs

This study investigates two FOXP2 constructs: the leucine zipper-end FOXP2 (LZ-END) and FOXP2 forkhead domain (FHD). The LZ-END construct (40,69 kDa) comprises the leucine zipper domain (LeuZip) and forkhead domain (FHD) and extends from residue 372 to the end of the FOXP2 sequence (residue 715). The forkhead domain (FHD) construct comprises the isolated winged-helix motif that extends from residue 504 to residue 594 (10,24 kDa). The experimental sequences of the LZ-END and FHD constructs are shown in Fig. 13 below. The LZ-END construct excludes the poly-glutamine region and the zinc finger (ZF) domain as these regions tend towards protein aggregation and result in a decreased yield of purified FOXP2. Furthermore, studies to date have suggested only the LeuZip domain and the FHD being associated with dimerisation and DNA-binding, respectively, and since oligomerisation is the focus of this study, only the regions encompassing the dimerisation interface are used.

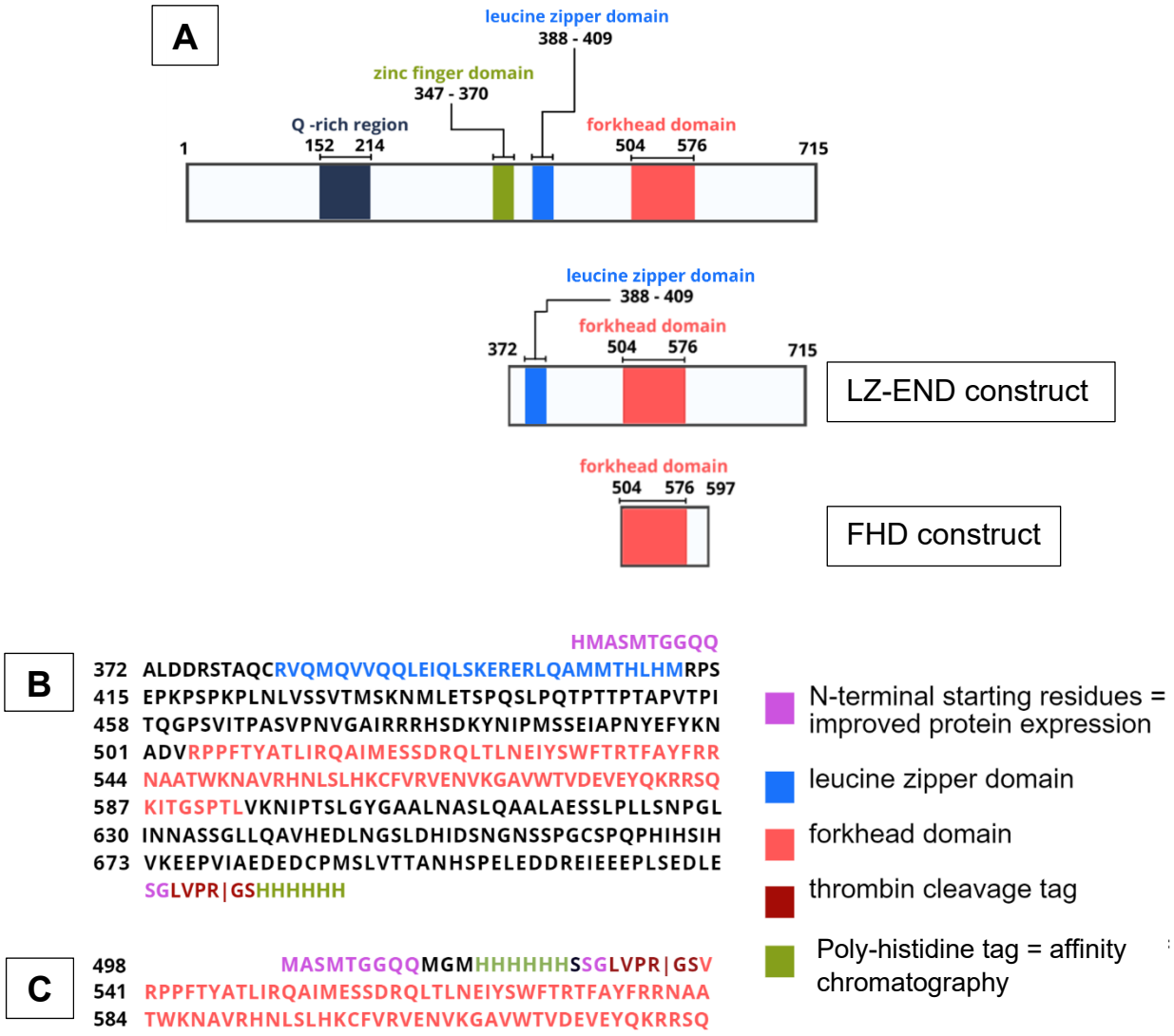


Figure 13. Amino acid sequences of the FOXP2 LZ-END construct and FHD construct, inclusive of poly-histidine tags and thrombin cleavage sites.

(A) The LZ-END construct comprises the leucine zipper domain (blue) and the forkhead domain (red) till the end of the sequence. The FHD construct comprises the isolated forkhead domain (red). **(B)** The FOXP2 LZ-END sequence spans positions, 372 to 715 and encompasses N-terminal starting residues that make for improved protein expression (purple), the leucine zipper domain (blue), the forkhead domain (red), a thrombin cleavage tag (maroon) and a poly-histidine tag at the C-terminus (green), used for affinity chromatography purification. **(C)** The FHD construct spans 504 to 594 (red), with the other highlighted portions functioning in the same manner as the LZ-END construct, however, the thrombin cleavage tag (maroon) and a poly-histidine tag (green) are located at the N-terminus.

3.1.2 Oligonucleotides

All the experimental procedures assessing LZ-END and FHD structure involve DNA. The FOXP2 FHD binds the following cognate DNA sequence with high affinity: 5'-TGTTTAC-3' (Nelson *et al.,* 2013) and this is referred to as Nelson DNA in this study. The oligonucleotide was acquired from IDT®, Whitehead Scientific (South Africa) as a double-stranded sequence, bearing unreactive flanking regions on either end: 5'-TTAGGTGTTTACTTTCATAG-3'. The molecular weight of the double-stranded oligonucleotide is 12 230.1 Da.

3.2 Transformation

Both the LZ-END and FHD construct sequences were codon optimised for expression in *Escherichia coli* and the sequences were inserted into a pET-11a plasmid (Novagen, Germany, GenScript Biotech Corporation, USA), bearing ampicillin resistance, and were transformed into competent T7 Express pLysS *E. coli* cells (New England Biolab, USA). This bacterial expression system produces soluble proteins at a reasonable quantity (Thulo *et al.,* 2019).

For both constructs, 300 ng of the relevant pET-11a plasmid was incubated with 100 µL of competent T7 *E. coli* cells on ice for 30 minutes. To ensure plasmid entry into the cells, the cells were heat shocked at 42°C for 45 seconds and immediately placed on ice for 5 minutes. To provide growth nutrients, 950 µL S.O.C. media [2% (w/v) tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, and 20 mM glucose] was added to the cells, following incubation at 37°C for 1 hour at 230 rpm. To select for successful transformants, the cells were spread and grown on LB-agar plates [1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 1.5% (w/v) agar] that were supplemented with 100 µg/ mL of ampicillin. The cells were incubated at 37°C for 16 hours and single colonies were harvested and resuspended in LB broth to make aliquots of a cell culture glycerol stock.

3.3 Protein expression and cell harvesting

To produce large quantities of the recombinant protein, the transformed *E. coli* cells were overexpressed by inducing the intracellular *lac* operon machinery (Baneyx, 1999). To ensure a good LZ-END protein yield, 6 L of cell culture was expressed in YT medium [1.6% (w/v) tryptone, 1% (w/v) yeast extract, and 0.5% (w/v) NaCl] from a 250 mL stock inoculated with the LZ-END glycerol stock (1:1000) and 100 µg/ mL of

ampicillin. This concentrated culture was grown for 16 hours at 37°C with constant shaking at 230 rpm. The 6 L expression media was inoculated with the 16-hour cell culture in a 1:20 ratio and also inoculated with 100 µg/ mL ampicillin. The culture was grown at 37°C with constant shaking at 230 rpm until the desired optical density (OD_{600}) of 0.6 was observed. Subsequently, the culture was induced to expression with 0.5 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) at 30°C for 4 hours. For the FHD expression, 2 L of culture produced a viable protein yield. The cell culture was inoculated with a 1:50 FHD glycerol stock and 100 µg/ mL of ampicillin. The induction conditions of the FHD construct differed to that of the LZ-END construct in that the cells were left to express for a longer time, 22 hours at 20°C.

Post induction, LZ-END and FHD cultures were centrifuged at 5000 ×g for 15 minutes at 4°C. The LZ-END pellets were resuspended in 35 mL equilibration buffer supplemented with 1 mg/ mL of chicken egg yolk lysozyme [20 mM Tris-Cl, pH 7.5, 500 mM NaCl, 50 mM imidazole]. The imidazole salt prohibits the non-specific of contaminating proteins. The FHD pellets were resuspended in 35 mL equilibration buffer supplemented with 0.1 mg/ mL of chicken egg yolk lysozyme (required to disrupt the bacterial cell wall and release the protein) and bearing a marginally lower concentration of imidazole (30 mM) in the equilibration buffer. A concentration of 1 mM of phenylmethylsulphonyl fluoride (PMSF) was added to the harvested LZ-END and FHD cell culture. PMSF was added to prevent the proteolytic degradation of the target protein. For storage, the LZ-END lysate was placed in a -20°C freezer immediately whilst the FHD lysate was left to incubate at 30°C for 30 minutes before freezing at around -20°C. This helped to improve the stability of the longer LZ-END protein and ensured that greater yields of recovered protein for the FHD.

In preparation for protein purification, the LZ-END and FHD lysates were thawed at 20°C and sonicated (Sonicator Ultrasonic Processor from Misonix Incorporates, USA) at 60 amplitudes with incubation on ice for 30 seconds in between 5x 20-30 second sonication steps. To prevent excessive bacterial DNA contamination, 1 µg/mL DNase was added to each lysate. The lysed LZ-END and FHD cells were left to incubate at room temperature for 15 minutes and 30 minutes, respectively. The difference in incubation times limits protein loss for the LZ-END construct, while 30 minutes is required for the DNase to fully integrate and mix with the more viscous FHD lysate.

The cells were centrifuged at 24 000 ×g at 4°C for 20 minutes (LZ-END) and 30 minutes (FHD). The supernatant, bearing the soluble target protein, was retained.

3.4 Purification

The purification of both the LZ-END and FHD constructs was achieved using immobilised metal-affinity chromatography (IMAC). The constructs were purified using 5 mL HisTrap™ HP IMAC columns, prepacked and charged with Ni Sepharose™ High Performance agarose beads (GE Healthcare Bio-Sciences, Sweden).

Histidine displays a distinctive intermolecular binding capacity (see Fig. 14) that resembles the bond formation of a tridentate ligand (Valenti, De Pauli and Giacomelli, 2006). It can facilitate octahedral bond coordination with its carboxyl and imidazole group, the amino nitrogen and the immobilised Ni^{2+} ions (Valenti, De Pauli and Giacomelli, 2006). This is interaction was leveraged with a poly-histidine affinity tag insertion at the C-terminal end of the LZ-END and N-terminal of the FHD. Purification via affinity tags makes for effective selectivity of target proteins, greatly aiding in their immobilisation amidst the milieu of non-target protein. This heightened selectivity is usually sufficient to permit a one-step purification. A neutral to basic pH ensures that the histidine imidazole groups remain unprotonated and can participate in electron donor interactions with the chelated Ni(II) ions (Falke and Corbin, 2013). Simply increasing the concentration of imidazole throughout the wash steps lends excess imidazole ions that compete for binding to the metal-charged resin and ultimately results in the elution of the His-tagged protein (Yip, Nakagawa and Porath, 1989).

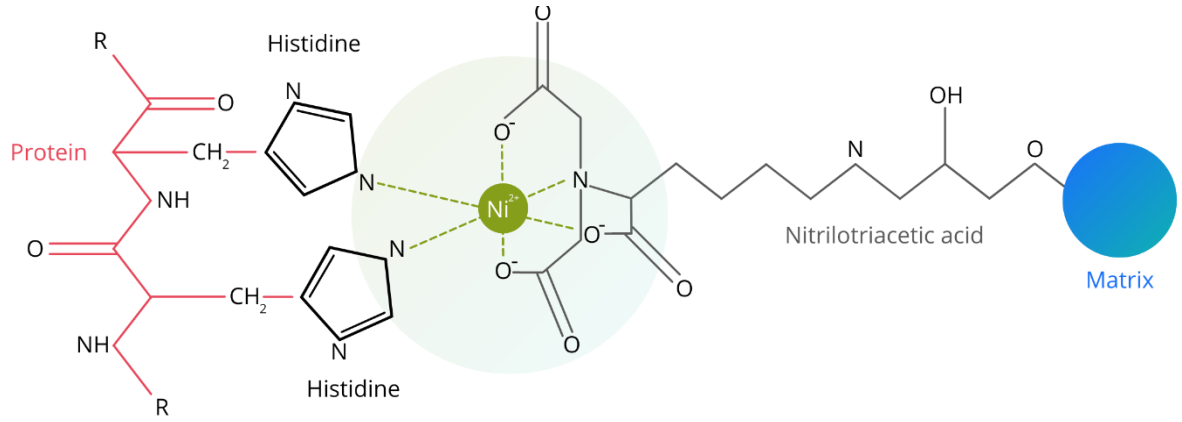


Figure 14. Structural representation of the non-covalent bond formation between histidine and Ni^{2+}.

The histidine tag can be placed near the N- or C-terminus and must be accessible to the nickel-nitrilotriacetic acid (Ni-NTA) matrix. The pyridine nitrogen in the histidine imidazole group can coordinate bond formation with transition metals such as Ni^{2+} and by increasing the free imidazole concentration of the system, the imidazole competes for binding at the immobilised resin and the His-tagged protein is able to elute from the system (Valenti *et al.,* 2006).

The wash steps for both constructs differed marginally in terms of the concentration of the reagents that were used and also in the duration of the wash steps. This information has been tabulated below:

Table 1. IMAC purification wash steps for the LZ-END and FHD constructs:

Construct	Buffer composition	Wash step (mL)
LZ-END	Salt wash: 20 mM Tris-Cl, pH 7.5, 1.5 M NaCl and 50 mM imidazole	50
	Imidazole wash: 20 mM Tris-Cl, pH 7.5, 500 mM NaCl and 100 mM imidazole	
	Elution wash: 20 mM Tris-Cl, pH 7.5, 500 mM NaCl and 300 mM imidazole	
FHD	Salt wash: 20 mM Tris-Cl, pH 7.5, 1.5 M NaCl and 30 mM imidazole	40
	Elution wash: 20 mM Tris-Cl, pH 7.5, 500 mM NaCl and 500 mM imidazole	

Between each wash step, equilibration buffer [20 mM Tris-Cl, pH 7.5, 500 mM NaCl and 50 mM imidazole (LZ-END) or 30 mM imidazole (FHD)] was run through the GE AKTA Prime Liquid Chromatography System. The eluted target proteins were collected in 6 mL fractions and stored at 4°C. The IMAC columns were stored in 20% ethanol at 4°C.

3.5 Protein molecular weight and homogeneity

The relative molecular weight and homogeneity of the LZ-END and FHD constructs were assessed using a fundamental analytical technique, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Laemmli, 1970). Solubilising the proteins at high temperatures in the presence of a reducing agent such as β-mercaptoethanol and an anionic surfactant like SDS, promotes protein denaturation into individual polypeptide chains by disrupting disulfide bond formation and proper protein folding. The SO_4^{2-} group of the SDS molecule confers an overall negative charge on the polypeptide chain. The denatured polypeptide chains migrate through a discontinuous polyacrylamide gel matrix of delimited porosity by the application of a voltage gradient and are separated according to their molecular masses. The denaturation process gives all proteins the same hydrodynamic volume-to-charge ratio and ensures that protein migration is solely dependent on the molecular mass of a protein, with smaller proteins migrating faster through the pores of the gel matrix than larger proteins. Determining the relative molecular weight of a protein can be quantified by comparison with standard protein molecular weights on a calibration curve.

The LZ-END and FHD constructs were resolved on discontinuous glycine and discontinuous tricine polyacrylamide gels, respectively. The two different tris-gel systems were employed on the basis of improving their resolved resolution by accommodating for differences in the constructs' molecular masses (Schägger, 2006). Protein constructs were heated in reducing sample buffer [10% SDS (w/v), 10% β-mercaptoethanol (v/v), 20% glycerol (w/v), 0.05% (w/v) bromophenol blue, 150 mM Tris-Cl, pH 7.0] at a 1:1 ratio for 5 minutes at 95°C. The samples were briefly vortexed to ensure the sample solutions were well combined.

The glycine gel system comprised a 4% stacking gel [500 mM Tris-Cl, pH 6.8] and 12.5% separating gel [250 mM Tris-Cl, pH 8.8]. The tricine gel system comprised a

4% stacking gel [6.25% (v/v) 3× gel buffer: 3 M Tris-Cl, pH 8.45, 1M HCl, 0.3% SDS] and a 16% separating buffer [19% (v/v) 3× gel buffer: 3 M Tris-Cl, pH 8.45, 1 M HCl, 0.3% SDS]. The glycine and tricine gels were polymerised by the addition of 10% ammonium persulfate (APS) and 0.2 – 0.5 mM tetramethylethylenediamine (TEMED). Electrophoresis for both systems was employed using the Bio-Rad miniPROTEAN Electrophoresis System (Bio-Rad, USA) apparatus and occurred at 40 V for migration through the stacking gel, which was increased to 140 V thereafter for a total run-time of 1,5 hours. The BLUeye Prestained Protein Ladder (Sigma-Aldrich, USA) was used to extrapolate the relative molecular weight of migrated proteins. The gels were stained in Coomassie brilliant blue R-250 overnight and subsequently destained [dH_2O, methanol, and acetic acid: 50/40/10 (v/v/v)] to reveal distinct bands of migrated protein.

3.6 Protein purity and concentration

Most of the gross protein contaminants are removed after purification and the target protein purity is accounted for in terms of the abundance of trace contaminants, usually nucleic acids, or detergents. The purity threshold or percentage purity that is regarded as 'acceptable' is dependent on the experimental application (Rhodes and Laue, 2009). By exploiting (through various techniques) the chemical or structural properties of the contaminants or target protein itself, it is possible to gauge the purity of a protein.

To assess the level of nucleic acid contamination and protein concentration, UV-Vis spectroscopy was employed. Since nucleic acids absorb light at 260 nm and proteins at 280 nm, their relative abundance in a sample can be assessed. Typically, an A_{280}/A_{260} ratio of ~1.8 is indicative of pure DNA, whilst ratios lower than ~1.6 indicate the prevalence of protein or phenols (Lucena-Augilar *et al.,* 2016). At a wavelength of 280 nm, tryptophan, and tyrosine (although to a lesser extent) absorb light, therefore, by applying the Beer-Lambert Law, the concentration of a protein can be directly calculated from its measured absorbance:

$$A_{280} = \varepsilon[M^{-1}cm^{-1}] \times c[M] \times l[cm] \qquad \textit{[equation 1]}$$

Where A_{280} is the reported absorbance (unitless), ε is the molar extinction coefficient of the protein and is measured in $M^{-1}cm^{-1}$. The concentration of the protein (c) is measured in terms of molarity (M) and the optical path length of the incident light (*l*) is measured in centimetres (cm). The molar extinction coefficients of the LZ-END and

FHD constructs were calculated from the amino acid sequences of each protein using the protein identification and analysis tools on the ExPASy server (Gasteiger *et al.,* 2003) and are as follows: 26 870 $M^{-1}cm^{-1}$ and 22 460 26 870 $M^{-1}cm^{-1}$, respectively.

3.7 Structural characterisation

3.7.1 Secondary structure characterisation

Far-UV circular dichroism (CD) spectropolarimetry is a technique that characterises the secondary structure of proteins by measuring differences in the optical absorptivity of left- and right-handed circularly polarised light (Rodger and Nordén, 1997). When plane polarised light traverses optically active compounds, such as proteins, elliptically polarised light is emitted, and the magnitude of the dichroic event is given by (Rogers *et al.,* 2019):

$$\phi = \frac{\pi}{\lambda(n_L - n_R)} \qquad \textit{[equation 2]}$$

Where ϕ measures ellipticity (in radians) of the emergent elliptically polarised light and at wavelength λ, n_L and n_R are the absorptivity indices of left- and right-handed circularly polarised light.

The measurement of the spatial arrangement of amino acids yields information about the secondary conformation of a protein. Secondary structural elements exhibit distinct far-UV CD profiles over a range of wavelengths (160 nm – 240 nm), producing characteristic absorption and emission profiles for a particular protein (Kelly *et al.,* 2005). In CD spectropolarimetry, the Beer-Lambert Law is applicable and the mean residue ellipticity (θ_{MRE}) can be measured:

$$\theta_{MRE} = \frac{100 \times \theta_{mdeg}}{c \times n \times l} \qquad \textit{[equation 3]}$$

Where θ_{MRE} is the mean residue ellipticity measured in $deg \cdot cm^2 \cdot dmol^{-1}$, c is the protein concentration (μM), n is number of residues, and l is the path length of the incident plane polarised-light, measured in centimetres (cm).

The secondary structures of the LZ-END and FHD constructs were assessed in the absence and presence of DNA to investigate any structural changes upon the addition of DNA. Proteins were dialysed against CD buffer [10 mM HEPES, pH 7.5, 50 mM Na_2SO^4] and centrifuged at 10 000 ×g for 5 minutes, to remove aggregates.

Measurements were taken in triplicate on a J-1500 Circular Dichroism Spectrophotometer (Analytical Solutions, South Africa), temperature regulated at 20°C. The protein concentration was 5 µM and the DNA concentration was 2.5 µM, for both independent LZ-END and FHD experiments. Experiments were conducted under reducing and non-reducing conditions by the addition or omission of 2 mM dithiothreitol (DTT). Experimental parameters were set using Spectra Manager™ software (Jasco, USA). The set wavelength started at 250 nm and ended at 180 nm, with a pitch of 0.2 for the LZ-END and 0.5 for the FHD. Continuous measurements were taken over 5 accumulations at a speed of 200 nm/ min with a response rate set at 1 second, standard sensitivity. The bandwidth for the LZ-END was set at 5 nm and at 1 nm for the FHD. Buffer and DNA contributions were subtracted from the triplicate average of the relevant sample falling under a high-tension voltage of 700 V. Data was reported in terms of mean residue ellipticity (θ_{MRE}).

Secondary structure content contribution for each construct (in the apo and DNA-bound forms) were predicted using the DichroWeb server (Whitmore and Wallace, 2004, 2008). To obtain quantitative data on protein secondary structures in terms of specific structural elements, the CONTINLL analysis algorithm was applied using Reference set 7, which was optimized for the 190 nm to 240 nm range of wavelengths. Two-sample t-tests were employed to determine the level of significance between observed differences between the apo and DNA-bound constructs

3.7.2 Tertiary structure characterisation

Fluorescence spectroscopy is a technique that is used to characterise the tertiary structure of a protein by harnessing the intrinsic optical properties of tryptophan (Lakowicz, 2013; Ghisaidoobe and Chung, 2014). Whilst chromophores like tyrosine and phenylalanine can fluoresce, the higher molar extinction coefficient of tryptophan produces an amplified signal when measured (Hannemann *et al.,* 2000). Tryptophan absorbs UV light at 295 nm, but its emission is highly dependent on its polarisation environment (ranging from 300 nm to 350 nm) thus, tryptophan residues in proteins serve as a good indicator of local structural behaviour (Lakowicz, 2013). Emission from the indole ring of tryptophan is influenced by the polarity of the buffer solution, protein concentration, pH, and temperature (Lakowicz, 2013). These experimental conditions can quench fluorescence intensity and produce bathochromic shifts (emission at lower

energy/ longer wavelengths) or hypsochromic shifts (emission at higher energy levels/ shorter wavelengths). Hyperchromic and hypochromic shifts are a result of increases or decreases in fluorescence intensity.

The tertiary structure of the LZ-END and FHD constructs were assessed in the absence and presence of DNA, with the intention of investigating changes to the local FOXP2 environment. Protein samples were dialysed in a suitable spectroscopy buffer [10 mM HEPES, pH 7.5, 50 mM Na_2SO_4] and centrifuged at 10 000 ×g for 5 minutes, to remove aggregates. Emission spectra was measured on an FP-6300 fluorescence spectrophotometer (Jasco, South Africa), temperature-regulated at 20°C. Protein concentration remained constant at 5 µM for both independent LZ-END and FHD experiments and at 2.5 µM for DNA. Experimental parameters were set using Spectra Manager™ software (Jasco, USA). Fluorescence parameters were set at emission mode with the photon multiplier tube (PMT) set at 400 V and the bandwidth at 5 nm. Buffer and DNA contributions were subtracted from the triplicate average of the samples. The data was reported as relative fluorescence intensity.

3.8 Equilibrium unfolding experiments

3.8.1 Urea Induced Equilibrium Unfolding

Equilibrium unfolding assays describe the folding mechanism and pathway that an unfolded protein undergoes to assume its native fold (Dill, 1990; Shirley, 1995; Pace *et al.,* 2000). The nature of the unfolding event is characterised by the protein folding funnel that illustrates the progression from an unfolded state towards the folded state, with several partially folded intermediates along the folding funnel, should they exist (Dill, 1990, Pace *et al.,* 2000).

Denaturation curves are useful for detecting variations in conformational stability between proteins (Pace, 1986a). Deductions can be made about the protein's folding mechanism, revealing the type of unfolding mechanism it may follow. Moreover, when a denaturation curve shows multiple steps, it usually indicates the presence of multiple domains within the protein, which allows for estimation of the domains' relative stabilities. This type of comparison can be made for the FHD and LZ-END construct, bearing in mind that the LZ-END construct is comprised of both the FHD (with its flanking unstructured regions) and the leucine zipper domain, and will provide insight

into the domain-specific stabilities that can be connected to the stability of the whole FOXP2 protein.

The LZ-END and FHD constructs (5 µM each) were denatured for 1 hour, in increasing concentrations of freshly prepared urea, ranging from 0 – 9 M, and in 0,25 M increments. Refractometry was employed to ensure that the desired urea concentrations were accurate. Urea was prepared in sodium sulfate buffer (10 mM HEPES at pH 7.5 with 50 mM sodium sulfate) for unfolding monitored by far-UV CD spectropolarimetry and intrinsic tryptophan fluorescence emission. The urea stock and protein were filtered before use and absorbance at 340 nm was performed to ensure that the protein solution was free of aggregates. After an hour of incubation at room temperature, the full-spectra measurements were recorded using the FP-6300 fluorescence spectrophotometer (Jasco, South Africa) and the J-1500 circular dichroism spectrophotometer (Analytical Solutions, South Africa). Temperatures of 20°C were maintained by a Peltier Thermostatted Cell Holder (Merck, USA).

Data fitting

In urea equilibrium unfolding, the Gibbs free energy change (∆G°) of protein unfolding is calculated by measuring the equilibrium constant (K) of the unfolding reaction at different concentrations of urea using spectroscopic techniques (Pace and Scholtz, 1997). The equilibrium constant (K) is related to the free energy change (∆G°) by the following equation:

$$\Delta G° = -RTln(K)$$

Where *R* is the gas constant, *T* is the temperature in Kelvin, and *ln* is the natural logarithm. By measuring the equilibrium constant at different urea concentrations and fitting the data to a linear model, the ∆G° of unfolding can be obtained (Pace and Scholtz, 1997). The urea concentration at which ∆G° is zero is called the midpoint of denaturation (Cm) and it corresponds to the concentration of urea at which the native and denatured states of the protein are in equal proportions (Myers, 1995; Pace and Scholtz, 1997). In addition to this, the *m*-value can be derived, which is a measure of the cooperativity of the unfolding transition and reflects the extent to which the denaturant affects the stability of the protein structure (Myers, 1995). Proteins with high *m*-values are more cooperative and have more compact, stable native states,

whereas proteins with low *m*-values are less cooperative and have more flexible, unstable native states (Myers, 1995).

Two-state unfolding mechanism

A two-state unfolding mechanism is comprised of 3 regions: (a) the pre-transition region, where the protein is predominantly in the native state, (b) the transition region, which reveals the slope of unfolding/ the cooperativity of unfolding and (c), the post-transition region, where the protein is predominantly denatured.

For a monomeric protein, the two-state mechanism reflects that the protein can exist in only one of the two states, that is, native (N) or denatured (D). Consequently, the total of the fraction of protein present in the solution that is native (F_N) and the fraction that is denatured (F_D) amounts to 1.

$$F_N + F_D = 1$$

$$\therefore F_N = 1 - F_D$$

In the following equation, *y* relates to the unfolding spectroscopic measurement retrieved from the intrinsic fluorescence and circular dichroism probes. In relation to the equation above, y, along the transition region is expressed as:

$$y = Y_N F_N + Y_D F_D$$

where Y_N and Y_D are the *y* values of the native and denatured protein, respectively.

Combing the above equation produces:

$$\therefore y = Y_N(1 - F_D) + Y_D F_D$$

$$\therefore y = Y_N - Y_N F_D + Y_D F_D$$

$$\therefore y - Y_N = F_D(Y_D - Y_N)$$

$$\therefore F_D = \frac{Y_N - y}{Y_N - Y_D}$$

Data fitting: three-state unfolding mechanism

The three-state model chosen for the fit of the LZ-END unfolding curve assumes that the native protein is dimeric and produces a dimeric intermediate (2N → I2 → 2U). In the following sets of equations, *y* again relates to the unfolding spectroscopic measurement retrieved from circular dichroism experiments pertinent to the LZ-END construct unfolding. In this model, *y*, along the transition region is expressed as:

$$Y_N = Y_{N0} + M_N(x)$$

$$Y_D = Y_{D0} + M_D(x)$$

$$Y_i = Y_{i0} + M_i(x)$$

Where Y_{N0} is the y intercept of the native protein baseline Y_{D0}, that of the denatured protein baseline and Y_{i0}, the baseline of the intermediate. The dependence of the signal due to native, denatured and intermediate protein on urea concentration is given by M_N, M_D, and M_i.

The three-state model assumes the existence of intermediate species in equilibrium, which results in the apparent extent of unfolding, F_{app}, differing from the two-state unfolding transition (N→D). This difference is determined by the concentration of intermediates present:

$$F_{app} = Z_i(F_i) + F_U$$

$$Z_i = \frac{y_i - Y_N}{Y_D - Y_N}$$

Where Z_i denotes the fractional change in *y* that occurs during the protein's transition from the native to the intermediate state.

3.9 DNA-binding experiments

3.9.1 Electrophoretic mobility-shift assays

In order to assess or account for the DNA-binding functionality of the LZ-END and FHD constructs, electrophoretic mobility shift assays (EMSAs) were performed. It is a non-denaturing technique; thus, DNA-binding of the intact native protein can be assessed, as well as the number of protein-DNA complexes. This is particularly important as it serves as an indicator of higher-order complexity in the presence of DNA.

It is a simple and rapid method used to detect protein-DNA interactions (Fried, 1989). The protein and DNA are combined, and the solution mixture is subjected to electrophoresis under native conditions through a polyacrylamide or agarose gel matrix (Holden and Tacon, 2011). Post electrophoresis, the distribution of protein-DNA species is evaluated via DNA staining (Hellman & Fried, 2007). Principally, the rate of migration of the DNA-protein complex is slower than the free DNA fragment. Should an interaction exist, the DNA will bind to the protein and form a complex that is in equilibrium with the unbound molecules (Hellman & Fried, 2007). When the rate of association (*k*a) is greater than the rate of dissociation (*k*d), a distinct band will be observed on the gel. A protein capable of oligomerisation may bind DNA as a higher-order complex and thus multiple bands would be observed on the gel (Holden and Tacon, 2011).

EMSAs were employed as functionality check on the experimental constructs as only correctly folded proteins are able to bind DNA. Samples were prepared in 1 x EMSA binding buffer [10 mM HEPES, 100 mM KCl, 1 mM $MgCl_2$, 10% glycerol, 0.1 mg/ mL BSA at pH 7.5] and DNA was left to bind to the LZ-END and FHD constructs for an hour at 4°C, in two independent experiments and at DNA: protein ratios of 3:1 to 1:10. Continuous 10% polyacrylamide gels were run at 160 V at 4°C in TBE buffer [890 mM Tris base, 890 mM boric acid, 20 mM EDTA disodium salt at pH 8.3] using the Bio-Rad miniPROTEAN Electrophoresis System (Bio-Rad, USA). The electrophoresis run time for the FHD was 1.5 hours and 2-2.5 hours for the LZ-END. The DNA-protein complexes were stained with SYBR™ Gold and visualised using the Molecular Imager® Gel Doc™ XR System (Bio-Rad Laboratories, USA).

3.9.2 Fluorescence Anisotropy

The *modus operandi* of transcription factors is to execute the regulation of genes, which requires them to bind promoter regions on target DNA. Estimating the binding affinity between a transcription factor and its target DNA can reveal much about the strength of the interaction. In this study, evaluating the binding affinities of both the LZ-END-DNA construct and the FHD-DNA construct would assess the effect of the leucine zipper on binding affinity and whether it weakens or strengthens the interaction. Fluorescence anisotropy was used to determine the binding specificities of the LZ-END-DNA and FHD-DNA complexes.

Fluorescence anisotropy measures the rotational diffusion of a fluorescent molecule and is predicated on the fact that the polarization of fluorescence emission from a molecule depends on the rotational diffusion of that molecule (Heyduk *et al.,* 1996; Favicchio *et al.,* 2009). Anisotropy is the ratio of the fluorescence intensity emitted in parallel and perpendicularly to the plane of the incident polarised light (Heyduk *et al.,* 1996; Favicchio *et al.,* 2009). As the rotational diffusion increases, the depolarisation becomes greater, which results in a decrease in anisotropy. Smaller fluorescent molecules display a lower anisotropy than larger molecules and this is particularly useful in studying protein-protein and protein-DNA interactions (LiCata & Wowor, 2008; Favicchio *et al.,* 2009)

Thus, the LZ-END-DNA and FHD-DNA interactions were assessed using fluorescence anisotropy. Cognate Nelson DNA (20 bp) labelled with a fluorescent 5-carboxy-X-rhodamine tag served as the fluorescent probe of the protein-DNA interaction. Theoretically, the unbound labelled DNA would display a low anisotropy since it is only 20 base pairs long, whereas the larger complex would display higher levels of anisotropy.

Two independent experiments were constructed and binding affinities (K_D) for each complex (LZ-END-DNA and FHD-DNA) were deduced. Anisotropy experiments were conducted on a Perkin Elmer LS-50B Luminescence Spectrometer (Perkin Elmer, USA) at 20°C. Nelson DNA, at 200 nM, was labelled with 5-carboxy-X-rhodamine at the 5’ end. The labelled DNA was mixed with increasing concentrations of either the LZ-END or the FHD construct, in DNA: protein ratios ranging from 2:1 to 1:30 in sample binding buffer [10 mM HEPES, pH 7.5, and 100 mM NaCl]. The excitation wavelength was set at 580 nm and emission at 605 nm at integration time of 3 seconds and with

a 5 nm slit width. The G-factor of the unbound DNA was measured (0,94349) to ensure the accurate calibration of the spectrometer's magnetic field strength, with reference to the DNA. Each experiment was performed in triplicate (biological replicates) with 10 technical replicates per measurement. DNA binding isotherms were plotted for each construct and fit to single-site specific binding model, with the following equation:

$$y = \frac{Bmax(x)}{K_D(x)} \qquad \textit{[equation 5]}$$

With *Bmax* indicating the maximum specific binding and K_D is the dissociation constant, in the same units as *x* (μM).

3.9.3 Hydrogen deuterium exchange mass spectrometry

Hydrogen deuterium exchange mass spectroscopy (HDX MS) is an indispensable tool used in the field of biophysics and biochemistry that can be used to make inferences about the conformational dynamics of a protein (Englander, 2006). In doing so, HDX MS data offers insight into protein-protein and protein-DNA interactions and conformational dynamics (Engen and Smith, 2000).

In this study, the conformational dynamics of the LZ-END construct were assessed in the absence and presence DNA. The conformational dynamics of the isolated FHD has been previously characterised in the absence and presence of DNA, however, the leucine zipper region and flanking unstructured regions surrounding the FHD have not. Assessing LZ-END dynamics in this study, provides insight into the domain-specific conformational changes associated with DNA-binding and assesses the effect of the leucine zipper domain and unstructured regions on the DNA-binding capacity of the FHD.

Principally, HDX is the replacement or exchange of labile hydrogens in the amide backbone of a protein, with its heavier isotope, deuterium (Chowdhury *et al.,* 1990). The native protein can become 'exchange-competent' when incubated in a deuterium-based solvent (Engen and Wales, 2015). The rate of HDX is primarily determined by solvent accessibility/ local environment, side chain effects and the strength of the hydrogen bonding in solution (Engen and Wales, 2015). HDX coupled to mass spectroscopy yields information about the global (top-down) and local (bottom-up) exchange-protection by revealing changes in the arrangement or orientation of certain

structural motifs that impact on the overall conformation of the protein. An overview of the HDX experimental workflow is shown below in Fig. 15.

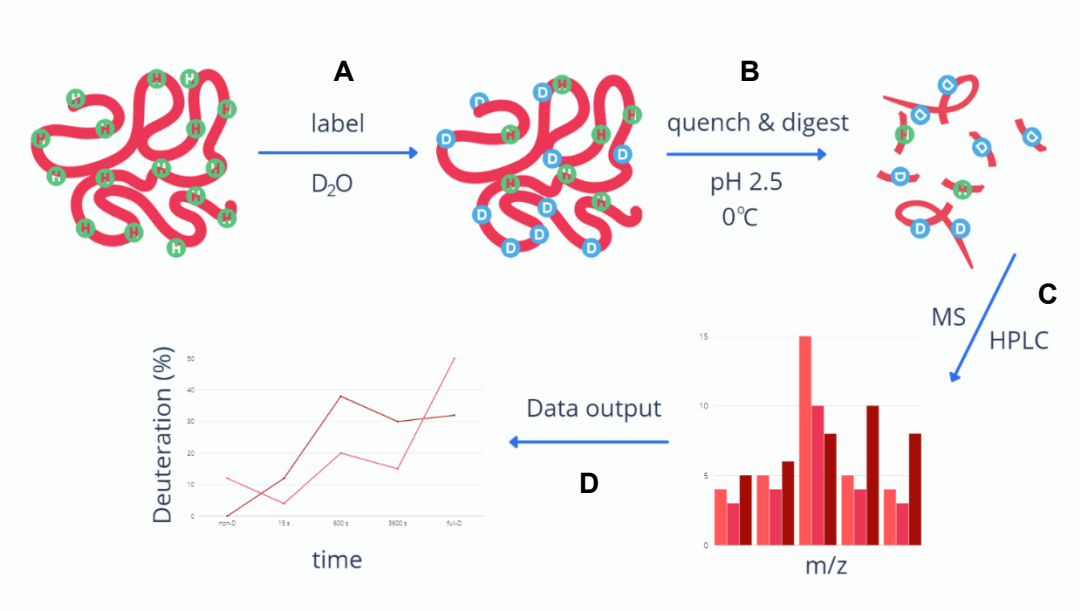


Figure 15. Schematic illustrating the experimental workflow of a typical bottom-up HDX-MS setup.

(A) As a first step, the experimental protein is labelled by dilution (10- to 50-fold) in a 90-99% deuterated buffer at pH ~7.0. **(B)** The protein in D_2O is then subjected to proteolytic digestion at quench conditions (pH 2.5, at 0°C) , which is usually performed by pepsin since it is stable at acidic pH (2.5). **(C) T**he fragmented peptides are then separated by HPLC and undergoes an identification process by the mass spectrometer. **(D)** The MS data is then analysed to reveal deuterium uptake plots for each fragment. [Figure adapted from: Rand *et al.,* 2014].

Proteolytic digestion is relevant to the bottom-up regime, in order to recover localised information, which requires improved resolution – at the peptide or residue level (Engen and Wales, 2015). The pH is maintained at 2.5 and at 0°C to limit back exchange as a 10-fold increase in exchange rates is otherwise observed for every pH unit above pH 4.0, thus, digestion occurs at quench conditions (pH 2.5, 0°C). In the top-down approach, the digestion step is omitted, as the proteins are ionised intact, to reveal global conformational dynamics. The peptides generated in the bottom-up digestion are then subjected to reverse-phase liquid chromatography (LC) separation, which has the highest peak resolving power compared to other LC methods. The chromatography system is coupled to a mass spectrometer that sorts through the generated peptidic fragments and generates a mass-to-charge spectrum for each

peptide. The deuterium uptake per peptide is calculated in reference to the non-deuterated control and deuterium uptake plots can be generated.

HDX-MS was performed at CSIR in Pretoria, South Africa. The labelling, quenching, and pepsin cleavage procedures were performed on the PAL HDX system (Leap Technologies, USA) which is used in conjunction to an Agilent 1100 high-power liquid chromatography (HPLC) system (Agilent Technologies, USA). Mass spectrometry was performed using the AB Sciex 6600 TripleTOF mass spectrometer (Sciex, USA). Labelling of the LZ-END protein construct involved incubation (approximately 10-3600 s) of the 30 µg of LZ-END construct, with and without 30 µg of Nelson DNA, in a deuterated binding buffer [20 mM Tris, 500 mM NaCl, 50 mM imidazole, pH 7.6] at an 8-fold dilution. Non-deuterated controls were constituted in non-deuterated binding buffer.

The sample was quenched to prevent back exchange, by two-fold dilution in an acidic quenching buffer, at 0°C in the LEAP column compartment (Leap Technologies, USA). The LZ-END construct was then subjected to pepsin digestion using an online Porozyme immobilised pepsin chromatography column, at 4°C for approximately 30 seconds and then desalted on an Acclaim PepMap trap column (0.3 × 5 mm) (Thermo Fisher Scientific, South Africa). Subsequently, the protein sample was loaded onto a Kinetex C18 reverse phase chromatography column (Thermo Fisher Scientific, South Africa) eluted by an acetonitrile (ACN) gradient at 200 µL/min, for 10 minutes. The peptide fragments were then separated onto the mass spectrometer in positive mode and according to the following parameters:

Table 2. MS1 and MS2 analysis settings

Parameter	Settings
Cycle time	3.5 seconds
Number of cycles	423
Ion source gas: GS1 & GS2	30 psi
Curtain gas	30 psi
temperature	150°C
Ion spray voltage	5500 V

Peptides were identified using PEAKS® Studio 11 software and deuteron incorporation assessed using the HDXaminer software (Sierra Analytics, USA). Mass spectroscopy analysis over 15, 600, and 3600 seconds revealed the rate of deuterium uptake of the DNA-bound and unbound forms of the LZ-END construct. Since the LZ-END protein has no solved structure, its conformational dynamics were mapped to the FOXP2 AlphaFold structure (Jumper *et al.,* 2021).

Deuterium back exchange was corrected according to the following equation:

$$D = \frac{M_t - M_{0\%D}}{M_{100\%D} - M_{0\%D}} \text{ x } N \qquad \textit{[equation 7]}$$

Where, M_t, $M_{0\%D}$, and $M_{100\%D}$ represent the averaged masses of the centroid isotope of the peptide fragments at exchange time, *t*, for the non-deuterated and fully deuterated control, respectively. The number of exchangeable peptide hydrogens is represented by N and excludes non-exchanging resides such as proline and residues at the N-terminus due to their rapid exchange rate.

Chapter 4: Results

Presented in this section, is the collection of findings that were observed in probing the structural characteristics and differences between the LZ-END and FHD constructs, in the absence and presence of DNA. These results are organised methodically to reflect the progression of research, with each data set serving as a foundation or premise for the next. Mapped below, in Fig. 16, is an overview of the experimental progression of this study in terms of the research objectives and techniques employed to achieve them. Broadly, they can be categorised as preparative experiments, structural characterisation, and functional characterisation. The preparative experiments constitute protein over-expression, purification, purity and size confirmation and assessing functionality. Structural characterisation was assessed at the secondary and tertiary level. Functional characterisation involved interrogating the DNA-binding affinity of both constructs, which informs the relative strength of the DNA-protein interaction. Protein backbone dynamics for the LZ-END informs the conformational structural contributions made by the leucine-zipper domain and how that impacts the dynamics of the FHD and its flanking regions. Lastly, stability studies provided insight into the folding mechanisms at play, in each construct.

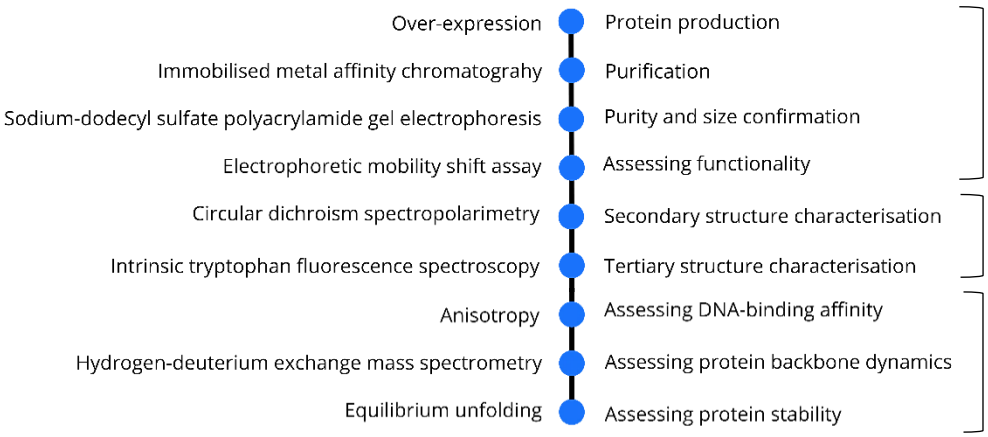


Figure 16. Illustration of the experimental progression and datasets presented in the study.

The preparative experiments **(i)** enabled the overexpression subsequent purification (via immobilised metal affinity chromatography) of the LZ-END and FHD constructs. Pure protein was obtained for both the LZ-END and FHD constructs, and this was assessed using sodium dodecyl sulfate polyacrylamide gel electrophoresis. As a prerequisite to structural and functional characterisation, electrophoretic mobility shift assays were performed to confirm protein activity. Structural characterisation **(ii)** was performed at the secondary and tertiary level. Functional characterisation **(iii)** assessed the DNA-binding affinity of Nelson DNA towards both constructs. Hydrogen-deuterium exchange mass spectrometry revealed the backbone dynamics of the LZ-END construct, which was compared to a prior study characterising the dynamics of the FHD, in the absence and presence of DNA. Lastly, urea-induced equilibrium unfolding informed protein stability by the extraction of thermodynamic parameters.

4.1 Expression and purification of the LZ-END and FHD constructs

Successful interrogation of the LZ-END and FHD structure and function required high working concentrations of pure, soluble protein. Protein expression is perhaps the first port of call in establishing the amount of soluble protein that can be recovered under optimised conditions. The expression protocol sought to produce the maximum amount of recoverable protein using *E.coli* cells under a T7 promoter. Protein

expression for the LZ-END construct proceeded over 4 hours and 22 hours for the FHD construct.

Given that an OD_{600} of 1.0 correlates to a culturability of 1.3 x 10^7 cells/ mL, at the point of induction (OD = 0.5) for both constructs, the cellular recoverability amounted to ~6.5 x 10^6 cells/ mL and continued to increase in the 4- or 22-hour expression period. Ultimately, an average of 400 mg – 600 mg of pure target protein is recovered. Notably, a 6 L expression culture is at minimum, necessary to recover ~500 mg of the LZ-END construct, whilst only a 1.5 L expression culture of the FHD produces the equivalent amount. This could be due to the tendency towards insolubility, which could be attributed to the unstructured regions of the LZ-END construct.

The constructs were purified via immobilised-metal affinity chromatography, via a Ni^{2+} charged column for the FHD construct (Fig. 17B) and a Co^{2+} charged column for the LZ-END construct (Fig. 17A). Different divalent cations were used to charge the columns as this improved the binding of the construct to the column and the recovered yield. FHD and LZ-END expression and purification has been previously optimised by Morris and Fanucchi, 2016 and Thulo *et al.,* 2019, respectively.

For the LZ-END construct, unbound gross proteins were immediately eluted in the flow-through fraction (see Fig 17A), at the application of a salt wash at ~15 mL, the non-target proteins were eluted, as well as at the point of application of the imidazole wash at ~40 mL. The elution buffer contains a 10-fold increase of imidazole, which is necessary to compete for binding at the resin. This enables the LZ-END construct to be eluted at ~63 mL, with a sharp peak observed at ~65 mL, followed by peak broadening. This broadening could be due to overloading the column with the sample. This may cause the resin to become saturated with protein, which leads to peak broadening. Nonetheless, all the eluted LZ-END protein was collected, and the level of purity was satisfactory for downstream experiments.

Similarly, for the FHD, the unbound or non-target protein was eluted immediately in the flow through fraction. The application of salt wash (~20 mL) was necessary to rid the column of non-specifically bound gross protein and at the point of elution of the FHD construct, a large sharp peak was observed, and pure protein was recovered. As seen from the corresponding gels in Fig. 17B, the purification produced sufficiently pure protein and did not require additional purification steps.

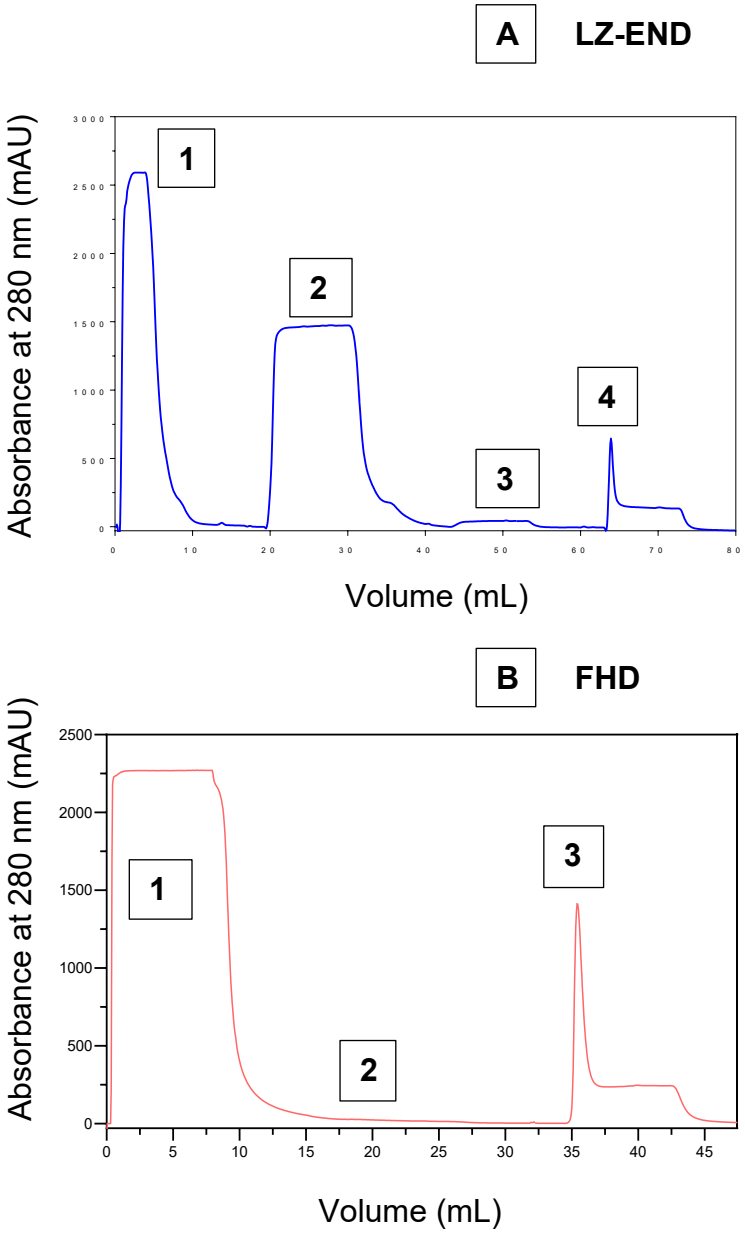


Figure 17. Elution profiles of the LZ-END and FHD construct, purified via IMAC

(A) For the LZ-END, protein absorbance was measured at 280 nm over 80 mL and at a flow rate of 5 mL/min. A series of wash steps was employed to remove non-target protein: **(1)** the flow through of unbound/ non-target protein, at low salt and low imidazole concentration, **(2)** application of a salt buffer, **(3)** 2-fold increase in imidazole concentration and **(4)** 10-fold increase in imidazole, resulting in the elution of the His-tagged LZ-END construct. **(B)** Protein absorbance was measured at 280 nm over 50 mL and at a flow rate of 5 mL/min. **(1)** shows the flow through of unbound/ non-target protein, at low salt and low imidazole concentration, **(2)** the application of a highly concentrated NaCl buffer at 20 mL, **(3)** the application of an imidazole wash, resulting in the elution of the His-tagged FHD construct. The pH was maintained at 7.5 throughout both purifications and no additional purification steps were required, since sufficiently pure protein was collected for both constructs.

4.2 Protein homogeneity and molecular weight confirmation

Indeed, downstream experiments required pure protein, but equally as important is the verification of the molecular weight of the constructs. Therefore, to assess protein homogeneity and verify the molecular weight of the LZ-END and FHD (post purification), SDS-PAGE was performed.

Shown below in Fig. 18 are the SDS-PAGE gels illustrating the purification of both constructs. The calibration curves are included so as to size the proteins on the gel and to establish whether the correct proteins were purified. For the LZ-END construct (Fig. 18A) it appears that trace amounts of protein may have eluted in the salt wash. The salt wash is necessary to remove the remaining gross protein, but perhaps a lower salt concentration would be more suited to this application and could negate the premature elution of trace amounts of the his-tagged LZ-END protein. Nonetheless, good yields of sufficiently pure protein were collected in the elution fractions (E1 – E6 of Fig 18A). Moreover, the calibration curve (Fig. 18C) confirmed that the bands produced in E1 – E6 correspond to a molecular weight of 46 kDa, which is the accurate.

Similarly, for the FHD construct, whilst the gross or unbound protein eluted in the flow through, a trace amount of protein was eluted in the salt wash (Fig 18B). Since no other bands or signs of additionally gross protein is observed in the salt wash lane, perhaps a salt-wash is not necessary to include since it only dislodged a small proportion of the target protein. In elution lanes E1 – E5, sufficiently pure FHD protein was collected. The calibration curve (Fig. 18D) also confirms the accurate relative molecular weight of the construct, which is 12 kDa.

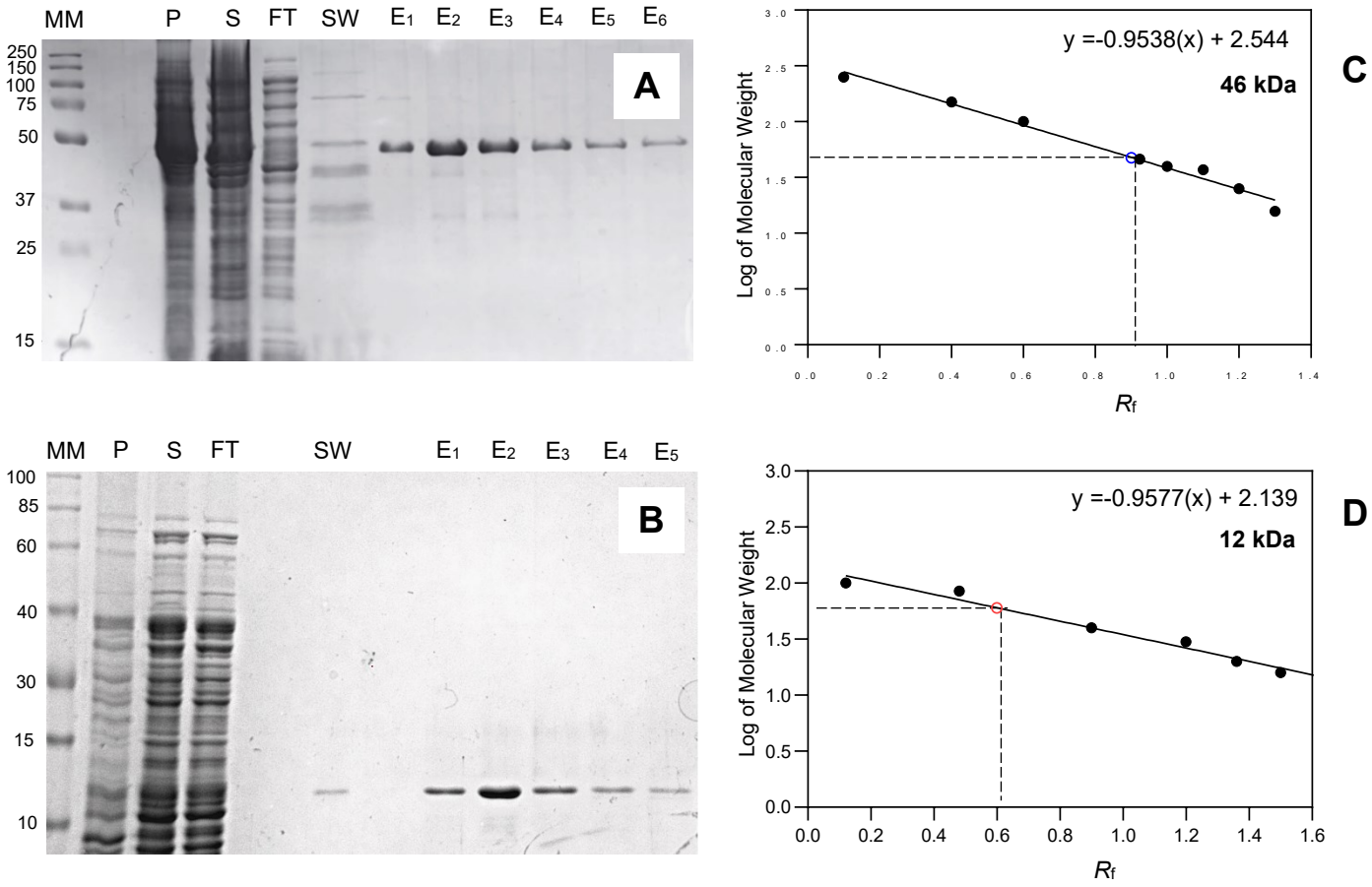


Figure 18. SDS-PAGE and calibration curve analyses of the purified LZ-END and FHD constructs.

(A) The LZ-END purification was resolved on a 12,5% glycine gel. The resuspended pellet (P) and lysate supernatant (S) are shown alongside the flow through (FT) of unbound protein. The salt wash (SW) shows the removal of column bound non-target protein and the elution of pure LZ-END is shown from E_1 to E_6.**(B)** The FHD purification was resolved on a 16% tricine gel for improved resolution. Similarly, pellet (P), supernatant (S) and flow through (FT) are shown along with the salt wash (SW) and the elution of pure FHD protein from E_1 to E_5. Molecular weight markers (MM – (A) Precision Plus Protein Dual Colour Standards and (B) Penn Protein Ladder), shown in both gels were used to construct the calibration curves for each protein. The logarithm of the molecular weight is plotted against the relative mobility for each marker band that migrates. The extrapolated molecular weight of the LZ-END **(C)** is 46 kDa and 12 kDa for the FHD **(D)** which are both as expected.

4.3 Assessing secondary structure

The amino acid composition of a protein informs its secondary structure and fold. Secondary structure motifs include α-helices, β-strands and random coils, which assume a fold that is held together by a hydrogen network of bonds.

It is important to confirm that the expressed and purified protein is folded in its native state as this lends to protein functionality (Greenfield, N.J., 2006). In this case, circular dichroism spectropolarimetry was implemented to assess the secondary structure of the LZ-END and FHD constructs in the absence and presence of DNA. The changes in the degree of ellipticity displayed over a wavelength ranging from 190 nm to 240 nm was monitored for both apo and DNA-bound forms of each protein. This allowed for a comparison of any structural changes induced upon DNA binding in both proteins. The supposition is that any quantifiable differences between the two constructs in the apo and bound form will inform the domain-specific structural contribution, within the FOXP2 protein. The circular dichroism results are shown in Fig. 19 below.

In both experiments, spectra were measured in the absence and presence of Nelson DNA to assess the changes in secondary structure upon the addition of DNA/ DNA-binding. Both the LZ-END and FHD display ellipticities that resembles a predominantly α-helical fold, in both the apo and DNA-bound forms. An α-helical fold exhibits a positive peak at 190 nm and a double negative peak at 208 nm and 222 nm, which is evident from the spectra (Fig. 19A and B). Protein aggregates that could potentially display increased absorbance at 200 nm were removed from the experimental solutions immediately before measurements were taken. This was performed by centrifuging the solutions at 5000 ×g for 5 min and then by ensuring that the absorbance reading at 320 nm was low (indicative of aggregate-free protein). Indeed, the absorbance observed at ~200 nm is valid and owes to a more folded protein, for both constructs.

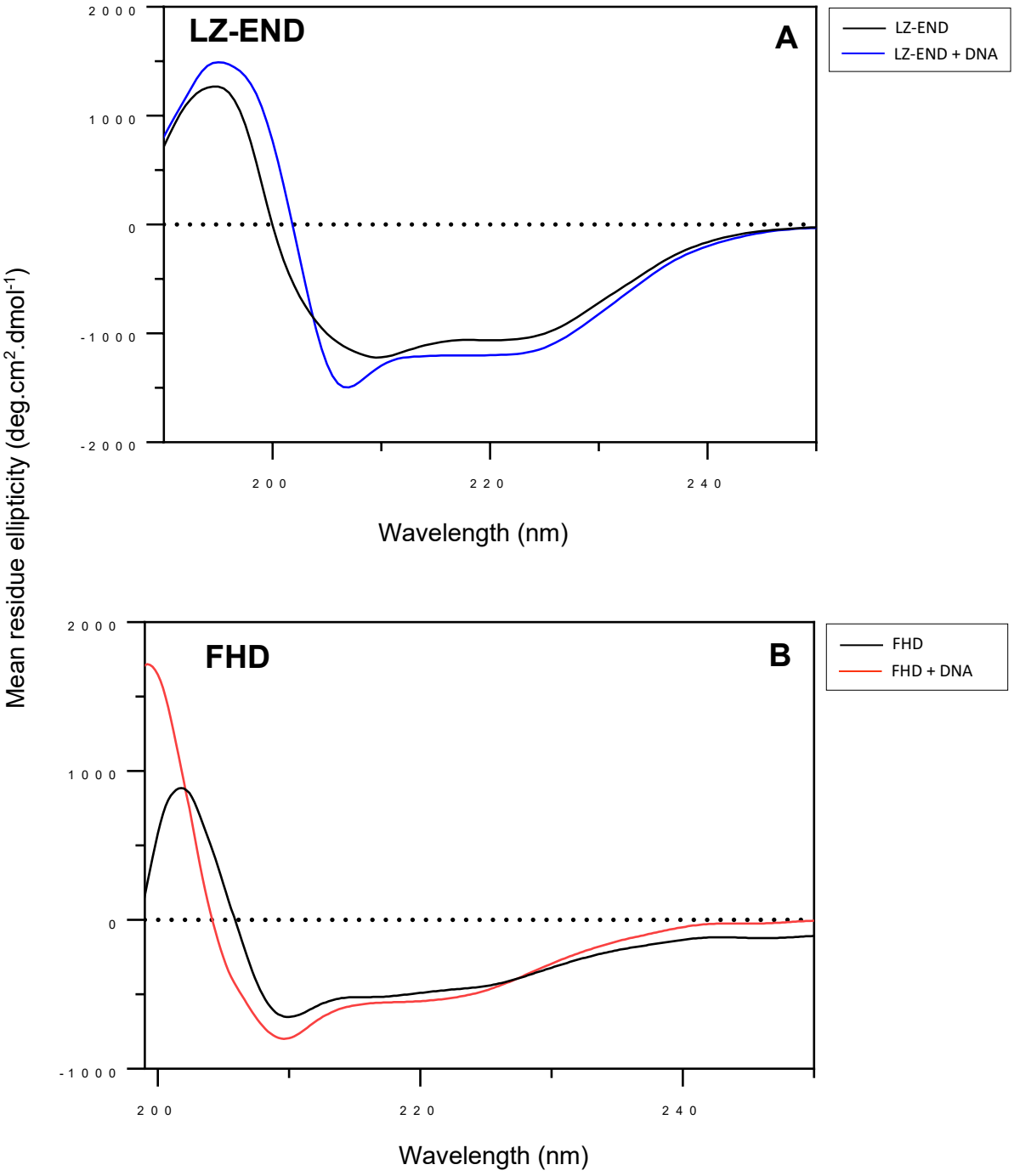


Figure 19. Circular dichroism spectropolarimetry profiles of the LZ-END and FHD constructs in the absence and presence of DNA.

Over the wavelength range (190 nm to 240 nm), the mean residue ellipticity of the apo and DNA-bound of both constructs were measured. **(A)** The apo LZ-END is shown in black, the bound form in dark blue and **(B)** the apo and bound forms of the FHD are shown in black and red, respectively. Protein concentration was kept at 5 μM for both constructs, with DNA at 2,5 μM. The spectra indicate a secondary structure that is predominantly α-helical under both conditions for both proteins, although some differences are apparent.

In the LZ-END:DNA spectra, the negative peak at 208 nm appears to be more distinct than observed in the apo form and the positive peak 190 nm displays a greater magnitude than that of the apo form. These findings may indicate that the LZ-END construct becomes more structured upon the addition of DNA. Similarly, the negative peak at 208 nm for the FHD:DNA spectra is slightly more defined than that of the apo FHD, with the peak at 190 nm significantly greater than the apo form as well. Again, this may suggest that the presence of DNA results in a more structured FHD.

4.4 Investigating the tertiary structure of the LZ-END and FHD constructs

The tertiary structure of a protein is the 3-D conformation of a polypeptide chain in space. The adopted conformation is a result of inter- and intra-molecular bonding patterns. Exterior hydrophilic interactions and interior hydrophobic interactions stabilise the conformation, rendering the tertiary structure biologically functional when in its native state, which is typically the state of lowest energy.

Tryptophan serves as a useful probe to assess the tertiary structure environment of a protein since it is highly sensitive to local changes in the protein environment. Exploiting the intrinsic fluorescence emission properties of tryptophan allows for the indirect measurement of structural changes that a protein may experience in different environments. Therefore, intrinsic tryptophan fluorescence spectroscopy was employed in this study to observe the effects of DNA on the tertiary structure of both the LZ-END and FHD constructs, so as to distinguish the tertiary structural contribution of the leucine zipper domain and how that compares to the forkhead domain. Since intrinsic tryptophan fluorescence is a good measure of changes in the protein's local environment, the addition of DNA would reveal the tertiary response of each of these domains in relation to the DNA (See Fig. 20). Furthermore, denaturing both constructs would also reveal domain-specific behaviour in terms of unfolding (See Fig. 20 and section 4.8). Notably, the tryptophan probes are solely located within the forkhead domain, thus distinctions between the tertiary structural changes of each domain can only be quantified if produced spectra are clearly disparate.

For both constructs, the apo and DNA-bound bound forms display fluorescence profiles that result in a double peak. The buffer components are constant in both cases (see section 3.7.2) and at a functional pH of 7.5. In this environment, irrespective of

the addition of DNA, the tryptophan residues which are found solely in the forkhead domain region (which is present in both constructs) appears to fluoresce in a dualistic environment. The double peak was observed at 330 nm and 340 nm. At these wavelengths the emission of the buried Trp548, found at the hydrophobic core, as well as the solvent accessible Trp533 and Trp573 residues is observed.

The apo FHD fluorescence displays a greater intensity than that of the apo LZ-END construct, with the DNA-bound forms displaying intensities slightly lower than the natives forms of both proteins. The decrease in fluorescence intensity is linked to the fact that the two proteins are of course unidentical, thus the observed decrease in fluorescence intensity for the LZ-END could be attributed to the presence of the leucine zipper and unordered regions. This may imply that the leucine zipper and flanking regions be oriented in a way that masks some of the fluorescence signals. Since the unordered regions randomly orient themselves in solution, it is possible that they could quench the fluorescence emission in the LZ-END construct. Not forgetting that the leucine zipper itself may also act in this capacity and quench the fluorescence by interacting with DNA. The quenching in both DNA-bound constructs can also be attributed to the fact that Trp573 in the FHD directly interacts with DNA when bound, this is in accordance with the FHD crystal structure (Stroud *et al.,* 2006).

For the denatured constructs, a hypsochromic shift was detected and produced a single peak at both wavelengths, thus signalling the complete denaturation of the FHD and LZ-END construct, as the buried tryptophan (Trp548) now contributes to the single emission intensity observed in the peak, indicating the complete exposure of the hydrophobic core. There is not much difference between the denatured spectra of both constructs, therefore this may not reveal the unfolding events of the leucine zipper domain.

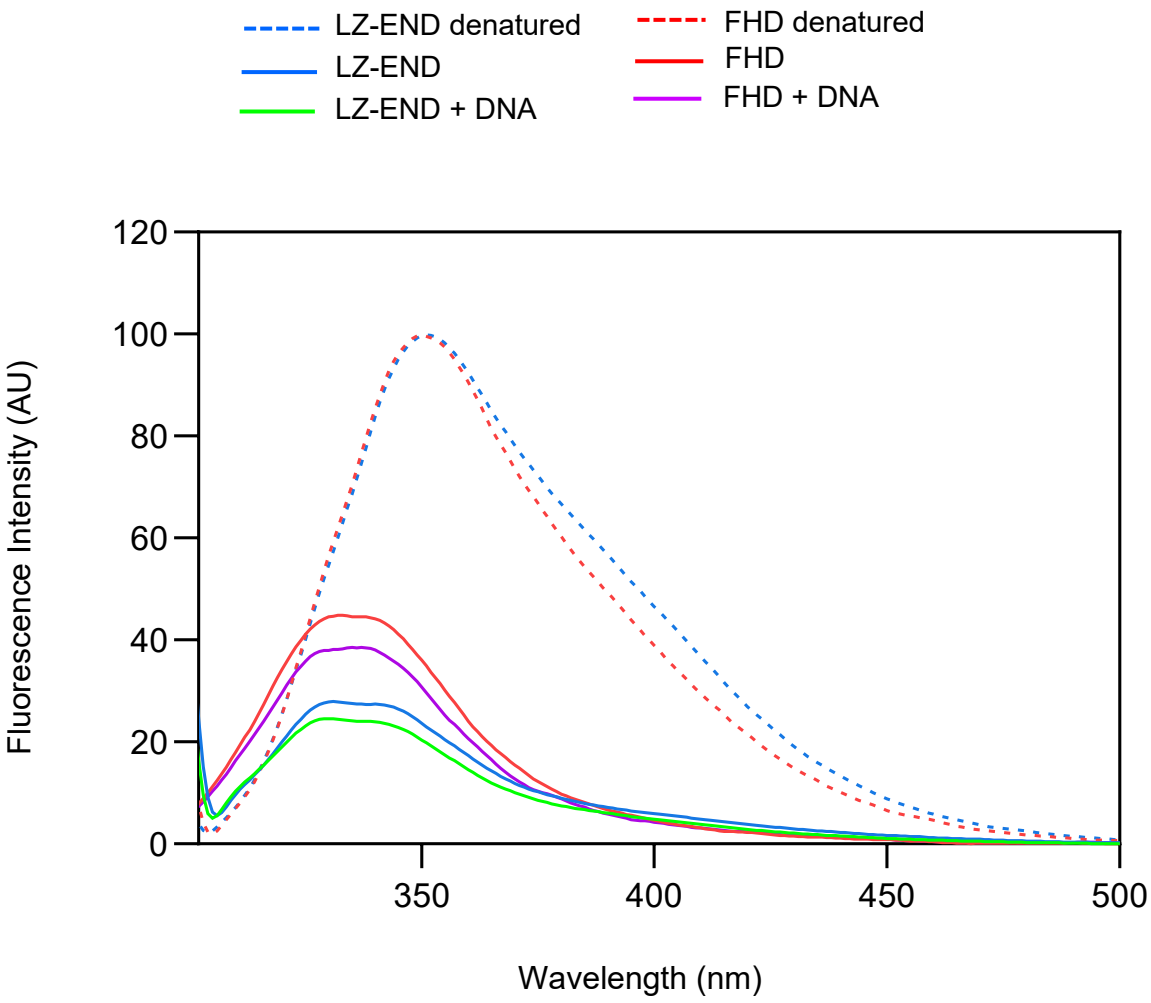


Figure 20. Intrinsic tryptophan fluorescence emission spectra of the LZ-END and FHD constructs in the absence and presence of DNA and 8 M Urea.

Both constructs were excited at 295 nm and their native, DNA-bound and denatured structures were compared. The fluorescence intensity of both constructs was monitored over the wavelength range of 300 nm to 500 nm. The native FHD (red) has a slightly greater intensity than its DNA-bound form (purple). Similarly, the native LZ-END (blue) also displays a slightly higher intensity than its DNA-bound counterpart (green). The addition of DNA may quench tryptophan emission to some extent as might the leucine zipper domain in the LZ-END construct. The denatured FHD (red dashed line) shows the same level of peak intensity as for the denatured LZ-END construct (dashed blue line).

4.5 Assessing LZ-END and FHD functionality and oligomeric state via electrophoretic mobility shift assays

Downstream experiments seeking to characterise FOXP2 structure and function require correctly folded, functional protein. Since FOXP2 is indeed a transcription factor, a key step in ensuring its functionality is to assess its ability to bind DNA. This was investigated via electrophoretic mobility shift assays. Increasing concentrations of protein were incubated with a standard amount of cognate Nelson DNA and left to bind for an hour so as to reach equilibrium.

As seen in Fig. 21A, the FHD construct bound to the cognate DNA, as shown by the faint top row of bands, from a DNA:FHD ratio of 1:5, up until the highest ratio of 1:40. From a ratio of 1: 0.5 to 1:3, no binding was observed, as the band produced in the free-DNA lane had appeared at these ratios as well. Contrary to the FHD, the LZ-END bound the cognate DNA from a DNA:Protein ratio of 3:1 up until a ratio of 1:5. More interestingly, the formation of higher-order complexes was observed for the LZ-END construct. This is seen by the multiple bands produced, which are clearly evident at ratios of 1:1 to 1:5. Whilst the oligomeric states observed is undefined, that is, it cannot be said whether the species are dimeric or of a higher-order complexity, this result confirms the oligomeric potential of the LZ-END construct. Indeed, the FHD has been shown to form a domain-swapped dimer *in vitro*, however, no oligomeric species were observed on the FHD EMSA. Therefore, in the LZ-END construct, the addition of the leucine zipper (which has been shown to dimerise with other FOXP proteins) may be the contributing factor to the observed oligomerisation.

Furthermore, seeing that the LZ-END bound DNA at a DNA:Protein ratio of 1:3 and binding of the FHD to DNA was only produced at a ratio of 1:5, this infers that more FHD molecules are required to bind DNA as compared to the LZ-END. This may also suggest that the FHD binds with weaker affinity than does the LZ-END. Additionally, since the LZ-END comprises the DNA-binding, forkhead domain but also the leucine zipper domain and unstructured regions that flank the FHD, the evidence of binding at lower concentrations of the LZ-END (as compared to the FHD), may suggest that the leucine zipper domain plays a cooperative role in binding.

Nelson DNA

5’-AGGTGTTTACTTTCATAG-3’

3’-TCCACAAATGAAAGTATC-5’

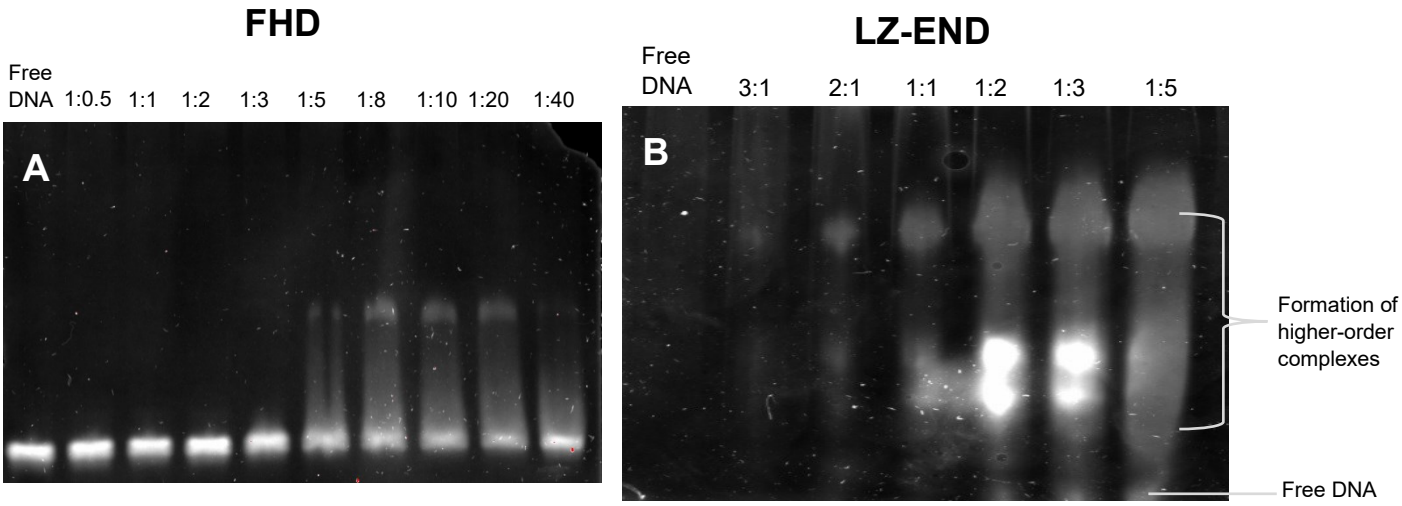


Figure 21. Native electrophoretic mobility shift assays for the LZ-END and FHD constructs.

Increasing concentrations of the FHD and the LZ-END were incubated with a constant concentration of 1 μM Nelson DNA. The samples were resolved on a non-denaturing 10% (w/v) polyacrylamide gel. **(A)** For the FHD, binding is observed from a ratio of 1:5 – 1:40, whilst unbound DNA is shown in the preceding lanes. **(B)** For the LZ-END, binding is observed at a ratio of 3:1 and continues to 1:5. From a ratio of 2:1 to 1:5, additional DNA-LZ-END complexes are observed.

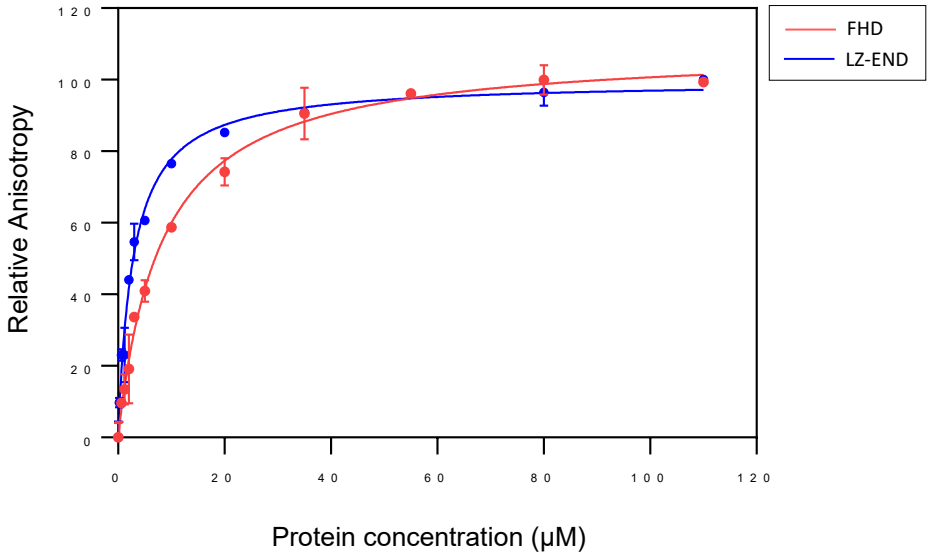
4.6 Investigating FHD and LZ-END stability via fluorescence anisotropy

Following the confirmation of DNA-binding in both constructs (EMSAs, Fig. 21), the strength of the DNA-binding interaction was needed to be assessed. Thus, the following sets of experiments compares the DNA-binding affinity of the LZ-END and FHD constructs. Both constructs possess the DNA-binding forkhead domain, but only the LZ-END construct bears the leucine zipper domain and additional disordered regions. Thus, differences in DNA-binding affinity could solely be attributed to the lack or presence of the leucine zipper domain, as all experimental conditions were constant for both independent experiments.

Fluorescence anisotropy provides a means of measuring the protein-DNA binding interactions or affinity, by exploiting the excitation properties of fluorescently-labelled Nelson DNA (5-carboxy-X-rhodamine fluorescent label) that bound to the protein in solution. Principally, the tumbling motion of the heavier labelled DNA-protein complex is slower than the unbound entities. Fluorescence of the complex in motion was detected and DNA-binding curves were produced (see Fig. 22) over increasing concentrations of protein. The anisotropy experiments characterised the strength of the protein-DNA association for both constructs. The DNA-binding isotherms were fit to single-site binding model (refer to section 3.9.2).

From the isotherms, dissociation constants for both proteins were derived. The K_D of the FHD for the cognate DNA was found to be 8.2 μM ± 1.54 μM while the LZ-END construct has a K_D of 2.8 μM ± 0.35 μM for the same sequence. This difference indicates that the FHD binds DNA with weaker affinity than the LZ-END construct and according to the two-sample t-tests, the difference was confirmed to be statistically significant (see Fig. 22 below). This means that the LZ-END requires a lower concentration of protein to achieve the same amount of DNA binding as the FHD. Overall, this may suggest that the leucine zipper (and potentially disordered regions) may promote DNA-binding in conjunction with the forkhead domain, which may play a bigger role in directing the specificity of the interaction (Fanucchi and Morris, 2016). The maximum specific binding (B_{max}) was calculated to be 108.7 ± 3.8 and 99.67 ± 2.6 for the FHD and LZ-END respectively. It is a measure of the maximum amount of DNA that binds to the protein. The lower B_{max} observed for the LZ-END may infer that the leucine zipper domain and the disordered flanking regions orient in a way that decreases the probability of the FHD making contact with DNA. Since the B_{max} in the

FHD is slightly higher, this may suggest that the absence of the leucine zipper domain/ disordered regions allows for DNA to bind the FHD, without hinderance.



Consensus sequence: 5'-/56-ROXN/TTAGG**TGTTTA**CTTTCATAG-3'

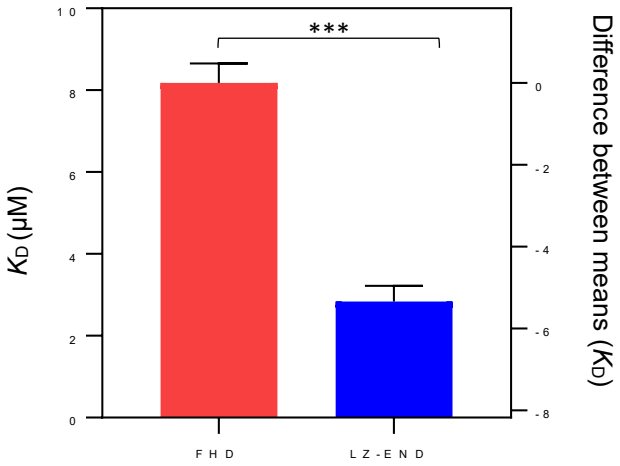


Figure 22. DNA-binding isotherm illustrating the independent binding interaction of the LZ-END and FHD constructs with DNA.

The Nelson DNA concentration was kept constant at 200 nm whilst protein concentration ranged from 0 μM – 120 μM and each experiment was performed in triplicate. The isotherms were fit to single-site, specific binding model: R^2_{FHD} = 0.9884 and $R^2_{LZ\text{-}END}$ = 0.9860. The maximum specific binding (B_{max}) was calculated to be 108.7 and 99.67 for the FHD and LZ-END respectively. The K_D for the LZ-END and FHD are 2,823 μM, 95% CI [2.449,3.186] and 8,173 μM, 95% CI [7.180,9.320], respectively, thus indicting that the LZ-END binds with higher affinity as compared to the FHD. Two-sample t-test revealed that there was statistically significant difference between the K_Ds (*p < 0.05, ** p < 0.01, and *** p < 0.001)

4.7 Investigating the conformational dynamics of FOXP2 by hydrogen-deuterium exchange mass spectrometry

Hydrogen deuterium exchange mass spectroscopy (HDX MS) is an indispensable tool used in the field of biophysics and biochemistry that can be used to make inferences about the conformational dynamics of a protein (Englander, 2006). In doing so, HDX MS data offers insight into protein-protein and protein-ligand interactions, and the conformational dynamics of a system (Engen and Smith, 2000). Principally, measuring the rate of deuterium uptake over time reveals the regions of the protein that are dynamic, in comparison to the control.

The conformational dynamics of the LZ-END construct was monitored via HDX MS in the absence and presence of its cognate double-stranded DNA (Nelson DNA) and was performed at the CSIR, Pretoria. HDX MS has been previously performed for the isolated FOXP2 FHD construct and the monomeric mutant A539P variant (Morris *et al*., 2018). The study showed that the isolated FHD adopts a flexible backbone when associating with DNA, forming a dynamic “open” structure, as compared to the mutant which had reduced flexibility due to a proline insertion (Morris *et al*., 2018). Considering this finding, an investigation into the change in conformational dynamics that could be directly attributed to the leucine zipper domain and FHD flanking regions in the near-full length protein, was carried out. The leucine zipper has been implicated in DNA-binding, therefore assessing its interaction with DNA, amidst the DNA-binding forkhead domain, provides both a global and local depiction of how these structural motifs dynamically cooperate to ensure DNA binding and if the coiled unstructured regions participate in the interaction at all.

The rate of deuterium uptake of the apo and DNA-bound LZ-END construct was measured over 15, 600, and 3600 seconds and compared to the non-deuterated control of both constructs and their fully deuterated forms. The time-resolved conformational dynamics of the LZ-END construct shows the level of isotopic exchange at each aforementioned time point, which helps to illustrate the motions of the protein. The addition of DNA informs which sections of the protein change in conformation, in order to directly bind to the DNA or adapt to support binding. The terms ‘shielded’ or ‘protected’ describe the portions of the protein that seem to be buried or sequestered from solvent and thus slow- or non-exchanging. Whilst the term ‘dynamic’ refers to the portions of the protein that are more solvent accessible and

fast-exchanging. Flexible regions of the protein display a combination of both forms of exchange.

The conformational dynamics of the apo and DNA-bound LZ-END construct was mapped to the AlphaFold FOXP2 model (AF-O15409-F1) that was retrieved from the AlphaFold Protein Structure Database (Jumper *et al.,* 2021). The model showcases the full-length protein, however, for the purposes of this study, conformational dynamics was mapped to the LZ-END portion of the full protein. Shown below is the predicted local distance difference test (pLDDT) for model confidence, for the FOXP2 AlphaFold model.

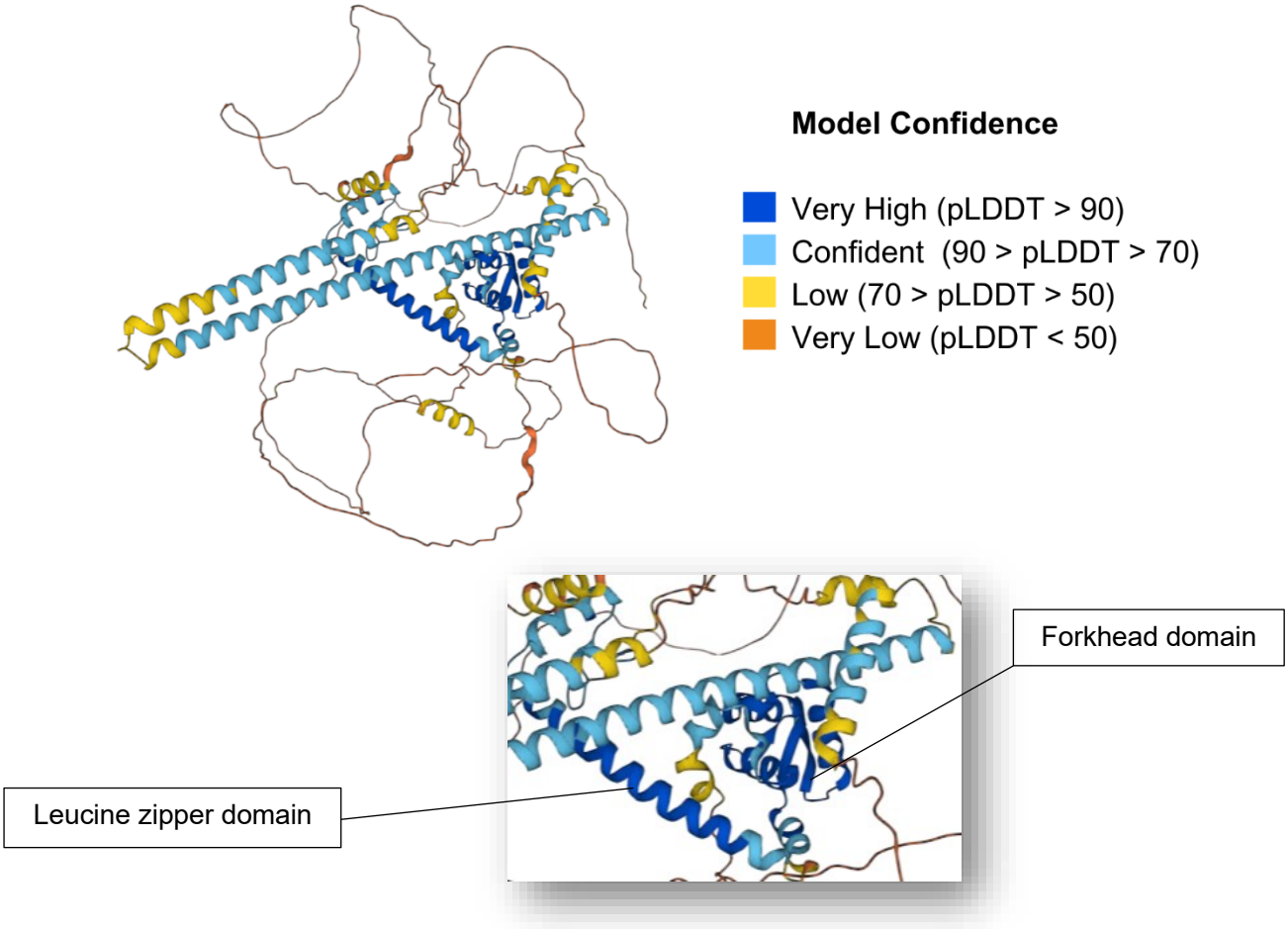


Figure 23. FOXP2 AlphaFold model with predicted local distance difference test score.

The predicted local distance difference test (pLDDT) score (0-100) evaluates the modelling confidence of each residue. Values greater than 90 are indicative of high confidence regions (blue), whilst values below 50 indicate low confidence regions (orange). The predicted secondary structure of the leucine zipper domain and the forkhead domain have pLDDT scores greater than 90 (very high confidence). Whilst the unstructured regions have a very low confidence score (less than 50).

The HDX data presented in this section has been divided into three key parts. First, the dynamics of the isolated leucine zipper domain is reported (Fig. 23), then the isolated forkhead domain (Fig. 24) and lastly the conformational dynamics of the entire LZ-END construct, inclusive of the unstructured regions (Fig. 25).

4.7.1 FOXP2 leucine zipper conformational dynamics

Comparing the apo and DNA-bound forms of the leucine zipper region helped to understand if the addition of DNA would cause it to observe a conformational change (see Fig. 23 below). In other classes of transcription factors, the leucine zipper region solely facilitates DNA-binding or works with other domains to mediate the interaction (Tarazona *et al.*, 2019). In Fig 23, the deuterium uptake of just the leucine zipper domain is illustrated to focus on its conformational dynamics. Unlike the quantification of conformational dynamics of the isolated FHD construct (Morris & Fanucchi, 2018), the conformational dynamics of the LZ-END needs to be considered in terms of the forkhead domain as well. Comparisons between these two constructs will reveal the effect of the leucine zipper on the dynamics of the FHD (see section 5, Discussion). Thus, in this section, the dynamics of the leucine zipper is first accounted for or more deeply investigated.

The *in vitro* motions of the apo and DNA-complexed leucine zipper domain display a mixture of high dynamism and also high protection from exchange at certain residues. Interestingly, the DRSTAQCRV sequence was 90% dynamic in both the apo and DNA-bound forms, for the full duration of exchange. However, residues such as, Leu-390, Glu-391, Met-402, and Met-410 showed increased protection from exchange, having only a 10% level of dynamism in both forms and across all time points. Stark differences in the dynamics between both forms was mostly evident at 15 seconds, at the very start of the exchange. At that stage, the DNA-bound form showed increased levels of protection from exchange. By the full duration of exchange, 3600 seconds, there was not much difference between the apo and complex dynamics, although the aforementioned protected residues are 10% - 30% more protected when in the presence of DNA.

Overall, it appeared as though the advent of DNA-binding did not induce significant changes on the structural dynamics of the leucine zipper domain, or at least the change could not be quantified. This observation was qualified by the difference maps

shown in Fig. 23. Here the leucine zipper domain is coloured in grey, illustrating that the change in dynamics between each from was negligible and there was possibly no difference between the dynamics. This could be demonstrating that the leucine zipper domain either does not make contact with the DNA or perhaps, when the contact was made, there was no conformational change to signify the binding. However, Gln-402 appears red (90%) in the difference map at 15s, indicating that this residue is highly dynamic in the initial exchange event but at 600s and 3600s it appears to be 30% protected. Perhaps this residue (Gln-402) and the other protected residues shown at 3600s (Leu-390, Glu-391, Met-402, and Met-410) play a role in increasing DNA-binding affinity or interact with DNA in some form.

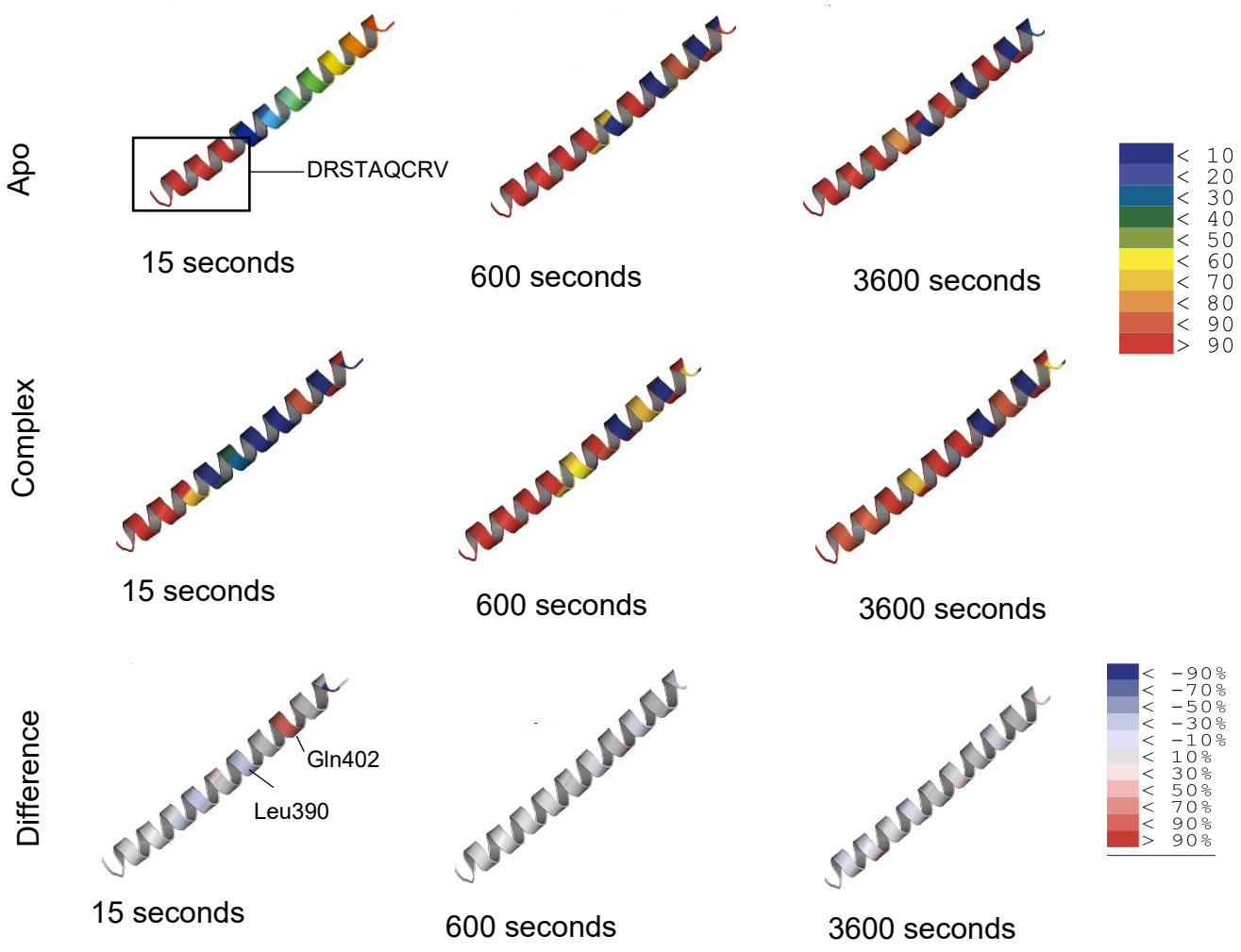


Figure 24. Comparison of the time-lapsed conformational dynamics of the leucine zipper domain in the absence and presence of DNA.

Apo and complex (DNA-bound) heat maps were mapped to the AlphaFold FOXP2 structure to visually represent the time-based dynamics of the leucine zipper domain, While the leucine zipper remains solvent exposed at all time intervals in the apo form, the DNA-bound form is initially protected from exchange and becomes more solvent exposed at 3600s, with conformational dynamics comparable to that of apo form at 3600s.

4.7.2 FOXP2 forkhead domain conformational dynamics

In this section, the conformational dynamics of the forkhead domain is dissected. It is important to reiterate that the dynamics of the forkhead domain is considered in the context of the larger LeuZip variant and not as an isolated entity. The isolated forkhead domain undoubtedly binds DNA and the hypothesis regarding the conformational dynamics of the forkhead domain within the LZ-END construct was that an increased level of protection from exchange would be observed at the FHD, relative to the leucine

zipper domain and unstructured regions and indeed, the results confirmed the hypothesis. The FHD dynamics are shown in Fig. 24.

Helix-3 of the FHD

As seen in Fig. 24, the apo form contrasts greatly to the DNA-bound complex, in that helix-3 (DNA-recognition helix) was found to be highly dynamic at 15 and 600 seconds (90%) and less so after 3600 seconds (60%) in the apo form but showed increased levels of protection in the complexed form. Interestingly at 15 seconds, helix-3 displayed a 50% protection in the complex and then a 10% reduction in protection was observed at 600 seconds. However, at 3600 seconds, shielding was increased to 70%. The dynamism in the apo form and intense shielding observed in the complex can be attributed to DNA binding at helix-3 as binding between the DNA and helix-3 prevents solvent accessibility to a large degree, ultimately reducing the level of exchange that would otherwise be observed.

Helix-2 and helix-1

Similar to H-3, helix-2 (H-2) also showed fluctuations in conformational dynamics. In the apo form, H-2 displayed 50% protection against exchange as early as 15 sec, but this decreased at 600 sec and 3600 sec with H-2 becoming highly dynamic (90%). Interestingly, in the DNA-bound complex form, as H-3 became increasingly protected from exchange, H-2 did not. At 15 sec, H-2 was found to 90% buried, notably, H-3 did not show levels of protection as great at this, however, at 600 sec and 3600 sec, the buried H-2 become more solvent exposed (60 % and 70% respectively).

The Ala516-Ile517-Met518 portion of H-2 did not show any conformational changes in the complex form and remained 90% buried for the entire exchange period. Similarly, in helix-1 (H-1), Leu527-Asn528 and Tyr531 also remained 90% protected from exchange in the complex form but were shown to be highly dynamic in the apo form. Perhaps these highly shielded regions of H2 and H1 are sequestered from the solvent in the presence of DNA, as shown in the FHD-DNA complex dynamics. Besides the Ala-Ile-Met portion of H-2 that is tightly shielded against exchange, the rest of the helix became increasingly dynamic as H-3 became increasingly protected from exchange. This perhaps suggests that H-2 is moderately extended away from the H-3-DNA interaction, perhaps to allow for the contact to favourably form.

Wing-1

Unsurprisingly, the connecting loop of the β-sheet wing region seemed to display high dynamism in both apo and complex forms, varying between 60-90% dynamism. The β-strands, however, were found to be 90% shielded against exchange, even in the apo form. This could be due to the extensive hydrogen bonding network that is responsible for holding β-sheet structure together.

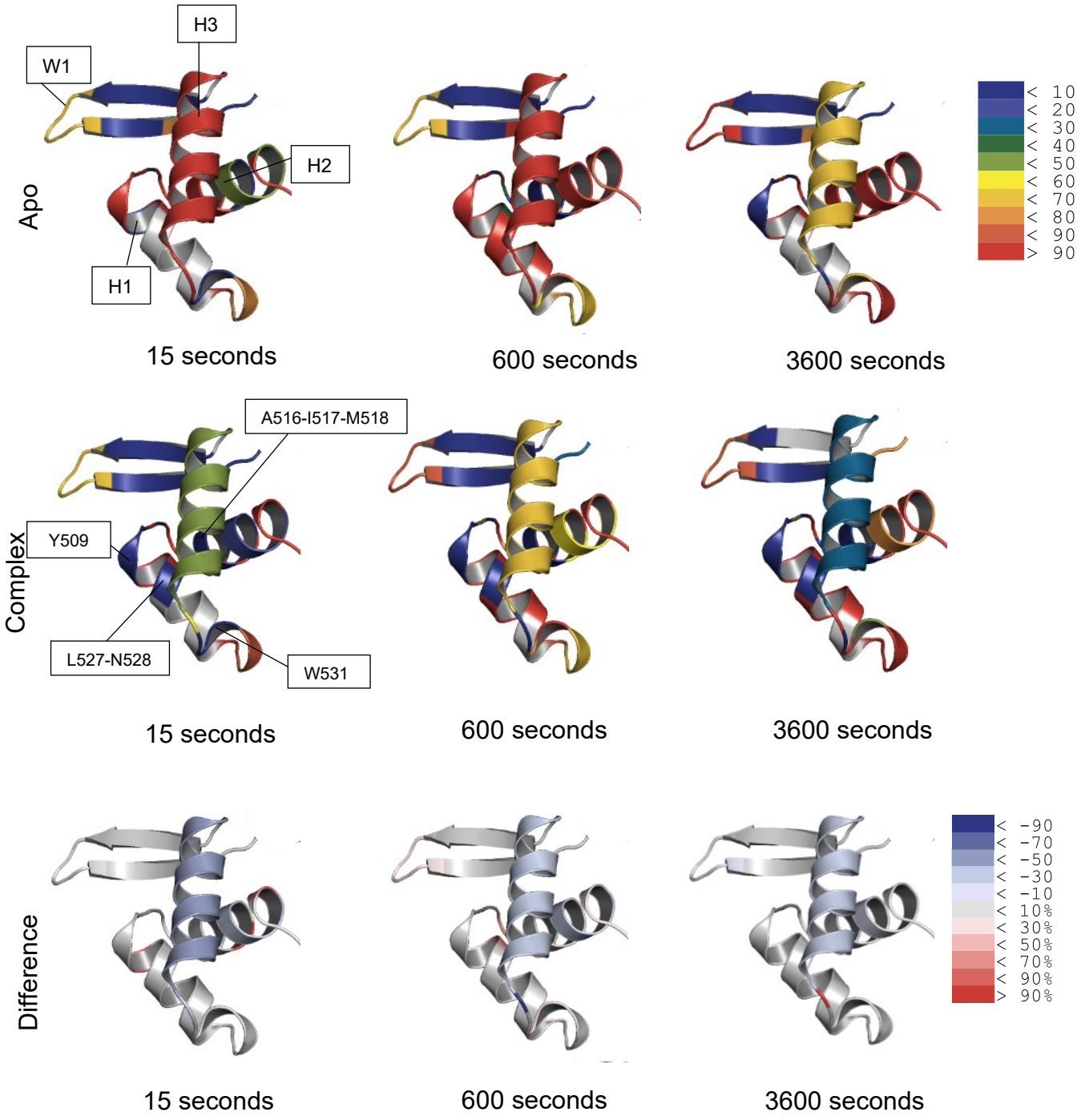


Figure 25. Comparison of the time-lapsed conformational dynamics of the FOXP2 forkhead domain of the LZ-END construct, in the absence and presence of DNA.

Apo and complex (DNA-bound) heat maps are mapped to the AlphaFold FOXP2 structure to visually represent the time-based dynamics of the forkhead domain, within the LZ-END construct. The apo, complex, and difference maps structures of the FHD are shown, each at 15s, 600, and 3600s. The level of exchange in the apo and complex structures is indicated by the warmth of the colours, with cooler colours (blue) indicating protection from exchange and warmer colours (red) indicative of increased levels of exchange. The colour scale in the difference maps also indicates the level of exchange in the same way as the apo and complex, with grey however, indicating no conformational change. In the Apo rendering, the forkhead domain becomes more solvent protected from 15s - 3600s, however in the complex, at 15s it appears to be initially protected (buried and compact) from exchange and gradually becomes more solvent exposed or dynamic at 3600s, with conformational dynamics comparable to that of apo form at 3600s.

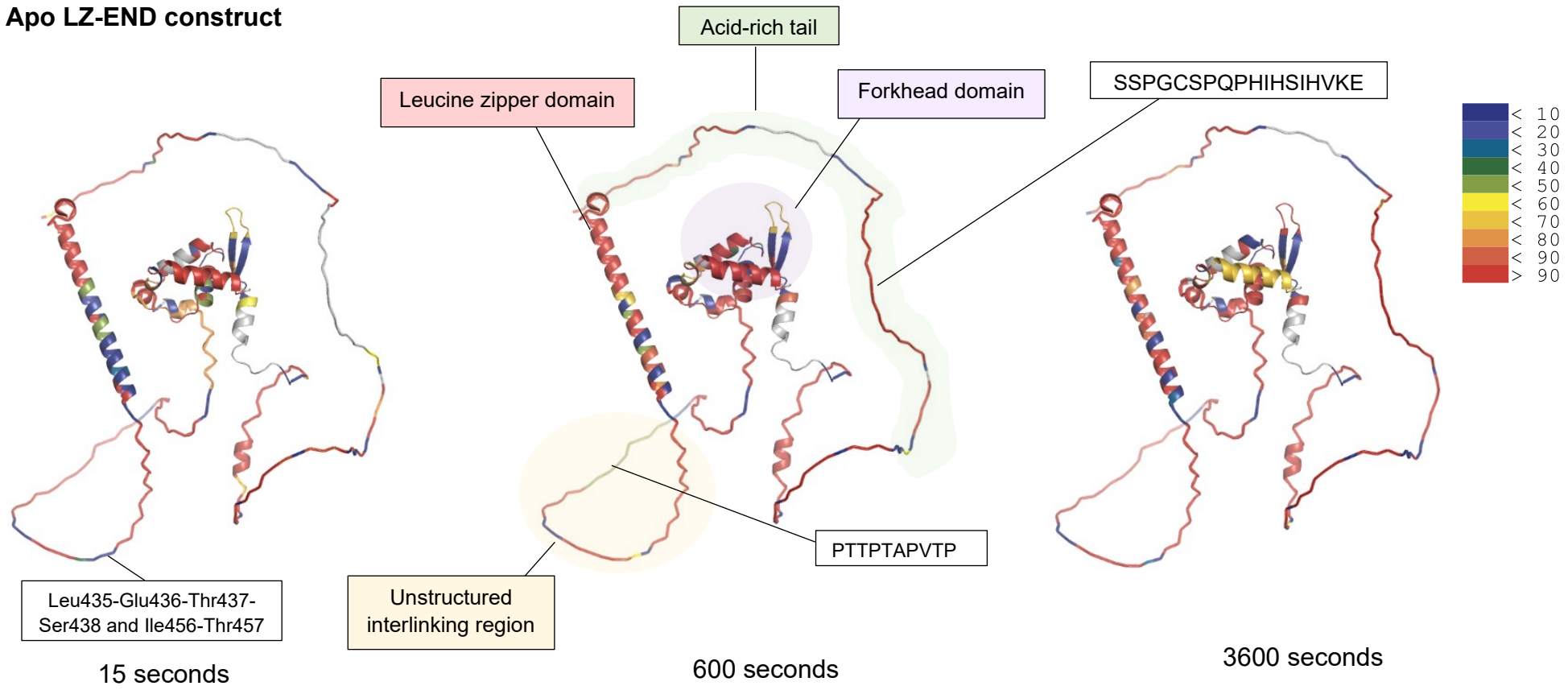


Figure 26. Comparison of the apo time-lapsed conformational dynamics of the full FOXP2 LZ-END.

The dynamics of the full apo LZ-END construct is shown at 15, 600, and 3600 seconds. The unstructured interlinking region is highlighted in yellow, the isolated FHD in purple and the acid-rich tail in green. Both the unstructured interlinking region and the acid-rich tail appears to be highly dynamic throughout the full period of exchange. Heat map key indicates the percentage of exchange protection associated with each mapped colour.

Complex LZ-END construct

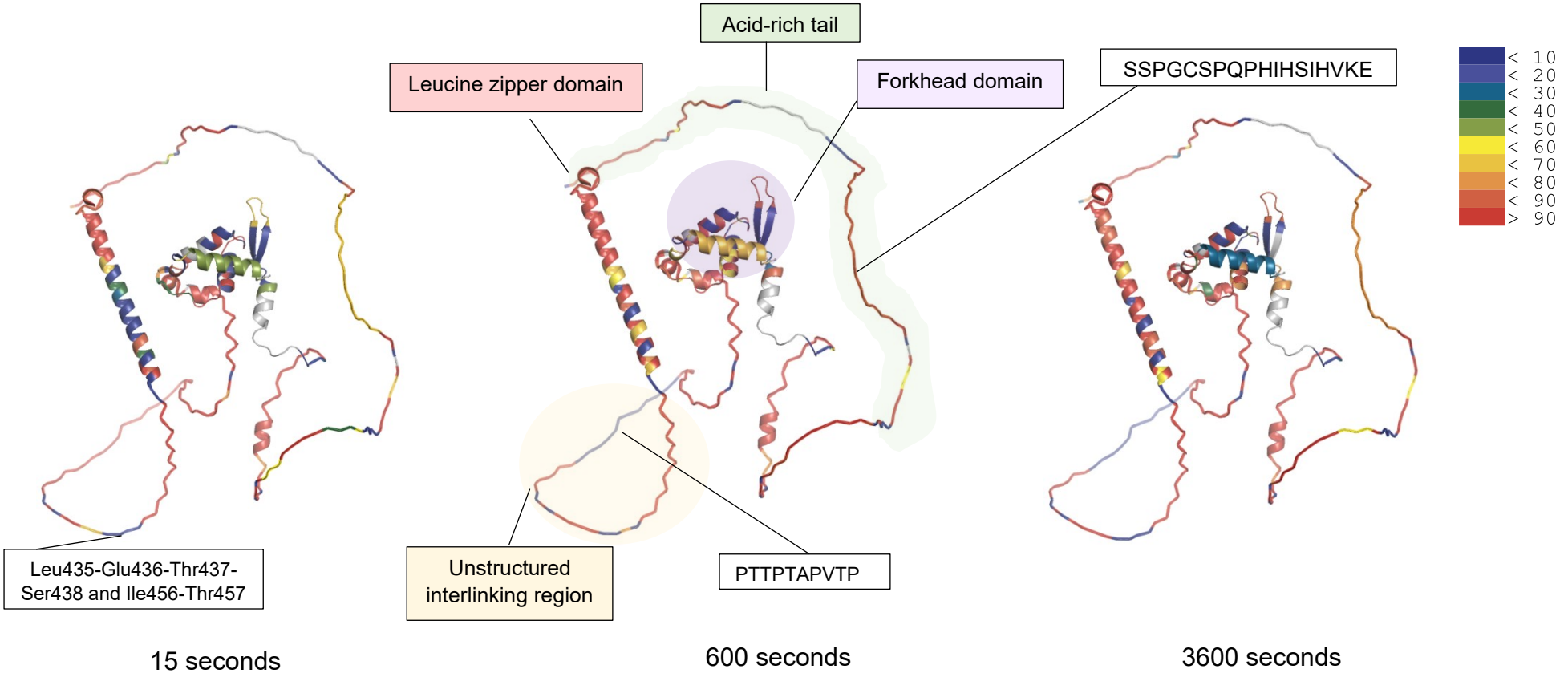


Figure 27. Comparison of the DNA-complexed, time-lapsed conformational dynamics of the full FOXP2 LZ-END construct.

The dynamics of the full complexed LZ-END construct is shown after incubation with D_2O for 15, 600, and 3600 seconds. The unstructured interlinking region is highlighted in yellow, the isolated FHD in purple and the acid-rich tail in green. Certain portions of the unstructured interlinking region show increased levels of protection from exchange and whilst the acid-rich tail appears highly dynamic, it too observed a reduction in deuterium uptake (~20-30%), but not by a significant amount. The unstructured interlinking region may be implicated in DNA-binding. The heat map key indicates the percentage of exchange protection (hence the degree of flexibility) associated with each mapped onto the AlphaFold structure.

The focus of this section of the results pertains to conformational dynamics of the unstructured loop regions of the LZ-END construct which contribute greatly to the overall secondary structure of the protein. In particular, the dynamics of the interlinking loop region (ILR) that connects the leucine zipper domain to the forkhead domain and the c-terminal acid-rich tail as seen in Fig. 25 and 26 are described in detail.

The interlinking loop region

In the apo structure, the ILR appeared to be highly dynamic (90%) at 15 sec and 3600 sec. Interestingly, at 600 sec, the PTTPTAPVTP sequence (coloured green) seemed to be protected from exchange as deuterium uptake was reduced by 30%. Certain portions of the ILR assumed distinct levels of protection and dynamism that remained throughout the full duration of exchange, this is confirmed by the deuterium uptake plots shown in Fig 25. Notably, Leu435-Glu436-Thr437-Ser438 and Ile456-Thr457 did not exchange and observed a protection factor of 90%.

In the complex structure, the IRL appeared to mimic the dynamics of the apo structure at 15 sec, in that it appeared to be highly dynamic. However, at 600 and 3600 sec, the PTTPTAPVTP sequence did not exchange and was 90% protected from exchange.

Because the IRL is indeed an unstructured region and therefore more likely to be solvent accessible, the shielding observed in the complex structure of the PTTPTAPVTP sequence infers that this portion of the IRL may interact with the DNA itself or other regions of the protein.

The acid-rich tail

The acid-rich tail (ART) constitutes a large portion of the LZ-END construct and spans from Glu642 up until the end of the sequence, ending at Glu715. In the apo structure, the ART was found to be highly dynamic, for the most part. The dynamics of the SSPGCSPQPHIHSIHVKE region were found to be inconclusive at 15 sec, however, at 600 and 3600 sec, it was shown to be highly dynamic (90%).

In the complex structure, a reduction in deuterium uptake was observed (30% reduction) at 15 sec and then increased to 90% at 600 sec, indicating the occurrence of high levels of exchange. This apparent dynamism was subsequently decreased (marginally), to 70%. The fluctuating levels of deuterium uptake for the ART may simply point to electrostatic repulsions that could occur between the negatively

charged sidechains of aspartic acid and glutamic acid and the negatively charged DNA backbone, thus rendering it mostly solvent accessible and likely unable to make contact with DNA.

4.8 Investigating equilibrium unfolding of the LZ-END construct and FHD

Equilibrium unfolding data provides important information about the stability and folding mechanism of a protein. In this study, measuring the conformational stability of the LZ-END and FHD constructs provides insight on the level of stability of their native conformations, which can then be compared to their unfolded conformations. This can be determined from the transition midpoint (Cm) which is the denaturant concentration at which the protein is half-unfolded and speaks to the ability of the proteins to maintain their native structures under denaturing conditions. Additionally, the folding mechanism of both constructs can be determined. By analysing the profile of the unfolding curve, one can determine whether the constructs undergo a two-state or multi-state unfolding transition. The latter of which would infer the presence of intermediates. Considering that LZ-END comprises two functional domains (the leucine zipper domain and the forkhead domain), any deducible differences in stability and the folding mechanism would link directly to the absence or presence of the leucine zipper and thus enlighten the fold contribution of each domain and its impact on the FOXP2 protein.

Experimentally, probing the unfolding pathway the constructs required denaturation. In this case, 9 M urea was used to denature the FHD and LZ-END constructs and unfolding was monitored, sequentially, by circular dichroism spectropolarimetry at 222 nm (for both constructs) and intrinsic tryptophan fluorescence spectroscopy at 330 nm and 340 nm for both constructs. Typically, the wavelengths selected to monitor unfolding spectroscopically are chosen on the basis that the unfolded and folded proteins display the greatest change in conformation. As shown previously, both constructs exhibited a predominantly α-helical fold, displaying a spectral minima at 208 nm and 222 nm. Since the trough at 222 nm is clear and pronounced, equilibrium unfolding was monitored by measuring circular dichroism at 222 nm. Following a similar line of thought, the wavelength chosen to monitor unfolding via fluorescence spectroscopy was 350 nm since the denatured peak was produced at this wavelength.

4.8.1 Establishing Reversibility

Protein stability can be measured by deriving thermodynamic parameters of protein unfolding at equilibrium. By assessing the denaturation midpoints (Cm), the cooperativity (*m*-value) and the change in Gibbs free energy (ΔG_{H_2O}) one can assess the stability of the leucine zipper domain and forkhead domain and understand how these domains support the overall stability of the FOXP2 structure. Hence, reversibility of unfolding was established so as to ensure that all measurement were made at equilibrium, this is shown in Fig. 27 below.

Both the LZ-END and FHD constructs were denatured in 8 M urea and left to equilibrate; changes in recoverability was assessed at the tertiary level (at 350 nm via fluorescence spectroscopy) and the secondary structure level (222 nm via circular dichroism). In order to ensure that both proteins are able to recover their native folded structure after denaturation, the denatured and native spectra were compared to the recovered spectrum. The proteins were first denatured in 8 M urea and then diluted 10-fold to recover their native fold. Since the dilution process dilutes the urea from 8 M to 0.8 M, 0.8 M urea was included in the native sample to ensure an accurate comparison.

At the tertiary structural level, the LZ-END observed ~82% recovery and the FHD a ~94% recovery. Secondary structural changes for the LZ-END saw a ~96% recovery and the FHD a ~90% recovery (Fig. 27). It is important to note that the refolded proteins contain trace amounts of urea - unlike the native spectra. These trace amounts of urea in the buffer result in increased signal noise (increased HT voltage) and thus results below 210 nm were unreliable. However, there was enough evidence (from the CD and intrinsic tryptophan fluorescence data) to demonstrate that the native structure can be reassumed after denaturation for both proteins and thus, further equilibrium unfolding studies could therefore be performed.

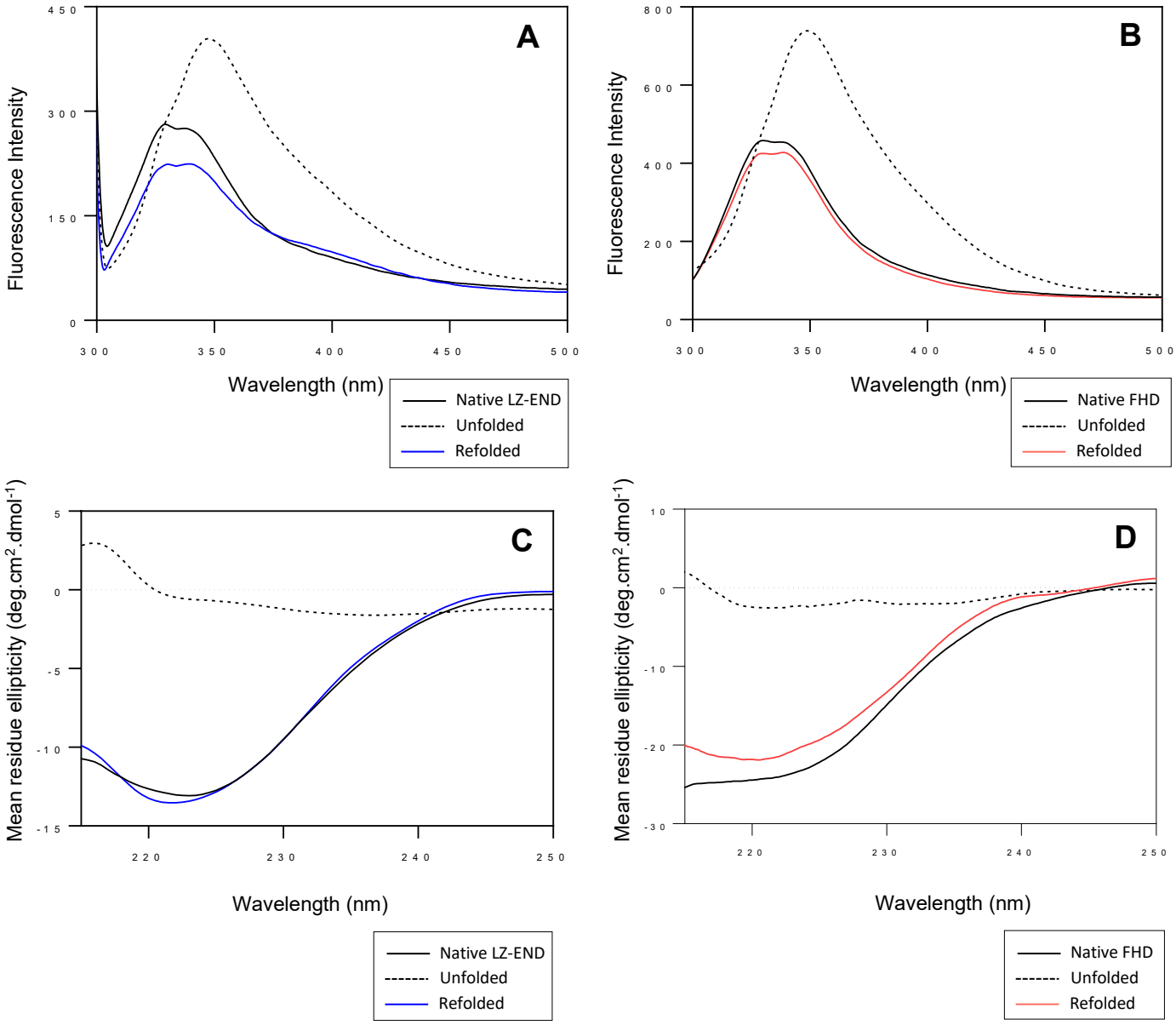


Figure 28. Native, unfolded and refolded secondary and tertiary profiles of the LZ-END and FHD construct.

Recovery of native secondary and tertiary structure of both LZ-END **(A and C)** and FHD **(B and D)** showed that both protein structures were recoverable after denaturation. Tertiary structural changes were monitored via intrinsic tryptophan fluorescence emission and showed a ~82% recovery for the LZ-END **(A)** and a ~94% recovery for FHD **(B)**. Secondary structural changes were monitored via circular dichroism spectropolarimetry at 222 nm. The LZ-END construct had a ~96% recovery **(C)** and the FHD had a ~90% recovery **(D).**

4.8.2 Urea-induced equilibrium unfolding assessed at the secondary and tertiary structure level

Tryptophan residues serve as the spectroscopic probe in assessing the unfolding trajectories of a protein, since they are a good reporter of changes in the protein's local environment. Hence, monitoring the intrinsic tryptophan fluorescence of a protein as a function of increasing urea concentration, one can observe changes in the protein's fluorescence emission spectra that correspond to changes in the protein's conformation and stability.

Since the tryptophan probes are located solely within the forkhead domain, any observable changes in the unfolding curves for both constructs can be attributed to the leucine zipper domain and the FHD disordered flanking regions. Thus, these experiments were performed to investigate whether the unfolding of the LZ-END construct, particularly the leucine zipper domain (since it is a functional FOXP2 domain) would impact on the unfolding trajectory of the forkhead domain. Furthermore, since spectroscopic measurements performed via circular dichroism is based on the differential absorption of circularly polarised light by the protein backbone, the measurements are not biased towards unfolding of the FHD alone. Therefore, comparing the stabilities and unfolding mechanisms at both levels helps to tease apart the domain-specific impact on FOXP2 stability and its folding mechanism.

Unfolding of the LZ-END and FHD tertiary structure

Monitoring unfolding via intrinsic tryptophan fluorescence spectroscopy for both constructs produced unfolding curves that were fit to a two-state model (See Fig. 28A). This model assumes that only two conformational states exist, the native conformation and the denatured conformation (N→D). As seen in Fig 28A below, both unfolding curves (LZ-END and FHD) are essentially superimposable and observe a single, sharp transition and therefore both clearly follow a 2-state folding mechanism. For the LZ-END, the denaturation midpoint (Cm) was 6.2 M, which slightly higher than that observed for the FHD (5.9 M). This slight deviation may suggest that the tertiary structure stabilizing forces are disrupted at a marginally lower urea concentration for the FHD as compared to the LZ-END. This may be probable considering that the FHD is not the only domain that needs to be unfolded in the LZ-END construct. The ΔG_{H_2O} for the FHD construct was found to be 8.63 $kJ.mol^{-1}$ and 19.89 $kJ.mol^{-1}$ for the LZ-END construct. This result suggests that the FHD is less stable than the LZ-END

construct. This may be plausible considering once more the addition of the leucine zipper domain. Prior to this investigation, EMSAs revealed the oligomeric potential of the LZ-END (see Fig. 28B). Perhaps the substantial difference between these ΔG_{H_2O} could be attributed to the energy taxing disruption of the bonds formed at the dimerisation interface of the LZ-END construct. Of course, this would mean that the native form of the LZ-END construct is in fact a dimer or higher-order structure.

The *m* value, informs the cooperativity of the unfolding transition. For the FHD, *m* was 4.3 $kJ.mol^{-1}.M^{-1}$ and 3.0 $kJ.mol^{-1}.M^{-1}$ for the LZ-END. This suggests that FHD construct is slightly more compact in the native conformation than the LZ-END.

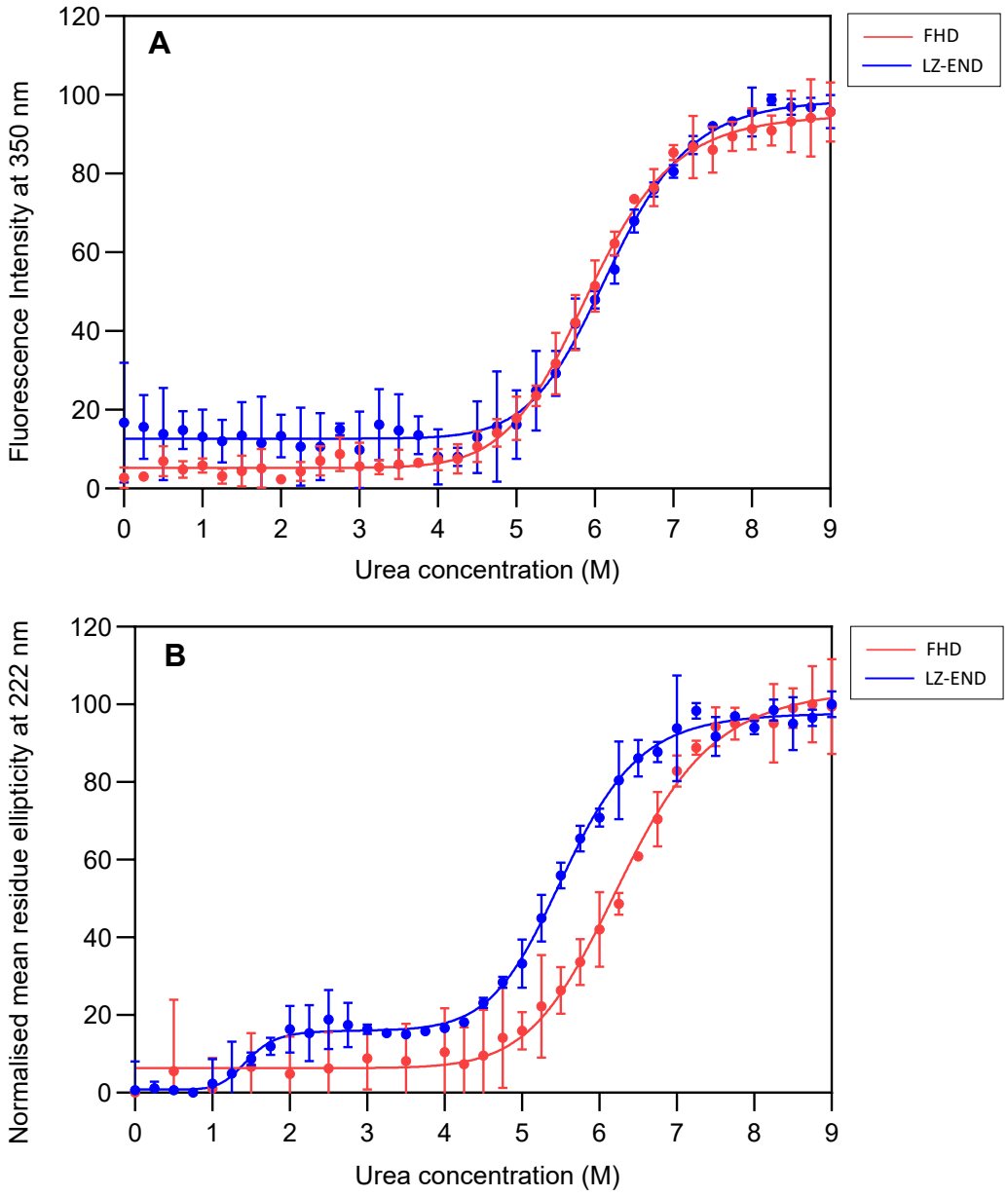


Figure 29. The equilibrium unfolding curves of the LZ-END and FHD at the tertiary and secondary structural level.

Urea-induced equilibrium unfolding was monitored via intrinsic tryptophan fluorescence spectroscopy and circular dichroism. The constructs were denatured increasing concentrations of urea (0 M – 9 M). The FHD (red) and LZ-END (blue) unfolding curves produced at the tertiary level is shown in **(A)** and was monitored at a λ_{max} of 350 nm**.** Both curves were fit to a two-state model, with R^2 = 0.985 for the FHD and R^2 = 0.971 for the LZ-END. Unfolding monitored at 222 nm at the secondary structural level for both constructs is shown in **(B)**. The FHD (red) followed a 2-state pathway with R^2 = 0.958 and the LZ-END (blue) displayed a three-state unfolding pathway, with R^2 = 0.989.

Unfolding of the LZ-END and FHD secondary structure

Monitoring unfolding via circular dichroism for both constructs resulted in unfolding curves that demonstrated a two-state model for the FHD construct but three-state model for LZ-END construct (See Fig. 28B). This follows that unfolding curve of the LZ-END construct revealed the presence of an intermediate (N→I→D), more specifically, a near-native intermediate. Indeed, the tertiary structural unfolding curves did not provide evidence for the LZ-END intermediate. This could be attributed to the fact that the unfolding probes at the tertiary level, which are tryptophan residues, are all found within the forkhead domain. This means that the formation of the intermediate probably stems from the leucine zipper domain since it was undetected at the tertiary level and the unfolding of the FHD construct at both levels is consistent with a 2-state model.

The Cm value of the FHD was found to be 6.2 M. Since the unfolding pathway of the LZ-END construct produced an intermediate, two Cm values were reported for this pathway. For the N→I transition, a Cm value of 1.5 M was obtained and for the I→D transition, the Cm value was 5.6 M. Notably, the Cm of the FHD construct is significantly higher than both Cm values produced for the LZ-END construct, which suggests that the FHD is more stable than the LZ-END construct.

For the LZ-END construct, a Cm on 1.5 M infers that near-native intermediate is destabilised at rather low concentrations of urea. The I→D transition is lower than that observed for the LZ-END at the tertiary level, which could mean that the secondary structure of the LZ-END construct is destabilised before the tertiary structure.

The ΔG_{H_2O} for the FHD was found to be 2.37 kJ.mol^{-1}, inferring that the secondary structure unfolds before the tertiary structure (ΔG_{H_2O} = 8.63 kJ.mol^{-1}). This may be plausible if conformation of the α-helices and β-sheet of the FHD are folded in a way that renders the tryptophan residues buried. Indeed, of the four tryptophan residues found in the FHD, two are solvent accessible and two are found within the hydrophobic core of the protein. Thus, destabilising the secondary structure may precede unfolding of the tertiary structure, in order for the structure to open up and reveal the hydrophobic core.

For the LZ-END, the ΔG_{H_2O} for the N→I transition was 7.7 kJ.mol^{-1}, whilst the I→D transition produced a ΔG_{H_2O} of 11.59 kJ.mol^{-1}. This means that the N→I transition is

less stable than the I→D transition. This makes sense since the intermediate is partly unfolding at low concentrations of urea. Interestingly, ΔG_{H_2O} of the LZ-END intermediate is higher than that observed for the FHD (2.37 $kJ.mol^{-1}$), which infers that the formation of the LZ-END intermediate is more stable than the isolated forkhead domain at the secondary level. Perhaps the leucine zipper domain plays a more prominent role in stabilising the intermediate, or perhaps even shields the FHD from the initial destabilisation of the secondary structural bonds. The I→D transition of the LZ-END requires the most amount of energy to become unfolded. Whilst this ΔG_{H_2O} for this transition is high, the ΔG_{H_2O} produced at the tertiary level for the LZ-END is even higher. Again, this suggests that the LZ-END tertiary structure is more conformationally stable than its secondary structure.

The *m*-value of the FHD at the secondary structural level was calculated to be 3.8 $kJ.mol^{-1}.M^{-1}$, with the LZ-END producing m-values of 3.7 $kJ.mol^{-1}.M^{-1}$ and 4.4 $kJ.mol^{-1}.M^{-1}$ for the N→I and I→D transitions, respectively. Comparatively, the compactness of the FHD is similar to that of intermediate produced (N→I) in the LZ-END. Perhaps the unfolding events at the N→I stage is comparable to the unfolding of the FHD. However, the I→D transition for the LZ-END construct observed a much higher *m*-value than the I→N transition and the *m*-value FHD. Fluorescence data probes the unfolding of the forkhead domain since the tryptophan residues are found solely within the forkhead domain. However, in the unfolding experiments, one observes the unfolding of backbone of both constructs, that is, the monitored unfolding is not limited to the forkhead domain alone, as is the case for fluorescence unfolding. Considering this, perhaps the higher m-value observed for the LZ-END I→D transition, demonstrates that the presence of the leucine zipper increases the stability of the LZ-END protein.

A summary of the results obtained from the thermodynamic parameters shown in table 3 below.

Table 3. Comparison of the thermodynamic parameters obtained from equilibrium unfolding curves of the LZ-END and FHD construct.

	Unfolding (3° structure)			**Unfolding (2° structure)**		
	ΔG_{H_2O} (kJ.mol^{-1})	Cm (M)	*m*-value (kJ.mol^{-1}.M^{-1})	ΔG_{H_2O} (kJ.mol^{-1})	Cm (M)	*m*-value (kJ.mol^{-1}.M^{-1})
FHD	8.63 ± 0.67	5.9 ± 0.45	4.3 ± 0.32	2.37 ± 0.94	6.2 ± 0.56	3.8 ± 0.64
LZ-END (N→I)				7.7 ± 0.32	1.5 ± 0.47	3.7 ± 0.85
LZ-END (I→D)	19.89 ± 0.54	6.2 ± 0.63	3.0 ± 0.43	11.59 ± 0.84	5.6 ± 0.65	4.4 ± 0.76

5. Discussion

The research undertaken in this study assesses the role of the FOXP2 leucine zipper as it cooperates with the forkhead domain to execute transcriptional function. Interrogating FOXP2 function is a complicated endeavor that benefits from investigating the structural organisation and conformation of key independently functional domains, the leucine zipper domain and the forkhead domain. The full length FOXP2 protein comprises a poly-glutamine region, zinc finger domain and the leucine zipper and forkhead domain as mentioned before, however, the leucine zipper domain and forkhead domain have been structurally and functionally implicated in DNA-binding and dimerisation events.

The forkhead domain is the conserved DNA binding domain among all members of the FOXP subfamily and is capable of binding DNA as a monomer and a domain-swapped dimer, *in vitro* (Stroud *et al.,* 2006). The leucine zipper domain has been shown to form homo- and heterodimers with FOXP1/2/4 (Li *et al*., 2004; Chae *et al*., 2006; Thulo *et al*., 2021). Thus, the leucine zipper domain and forkhead domain have proved to be key players in probing FOXP2 structure and its related function. The forkhead domain is the only domain of the entire FOXP2 protein to be solved via X-ray crystallography. So far, the only representation of the full-length structure of the entire protein has been achieved through AlphaFold, as shown below in Fig. 29, and according to the AlphaFold structure and other secondary structure prediction tools, a large proportion of the FOXP2 protein is indeed disordered.

Hence, in this study, partitioning the structural and functional contributions of the leucine zipper domain, forkhead domain and the flanking unstructured regions in the absence and presence of DNA provided both a global and local perspective on the cooperativity of these domains and regions, in relation to each other and their capacity for the DNA-binding interaction. This was achieved by characterising the DNA-binding capabilities, stability, and dynamics of FOXP2 LZ-END construct and the FHD construct. As mentioned prior, the LZ-END construct comprises the leucine zipper domain until the end of the FOXP2 sequence, which is inclusive of the forkhead domain. The approach was employed to understand the effect of the leucine zipper on the FHD (with its flanking regions) and whether the leucine zipper had a more prominent role in directing certain activities.

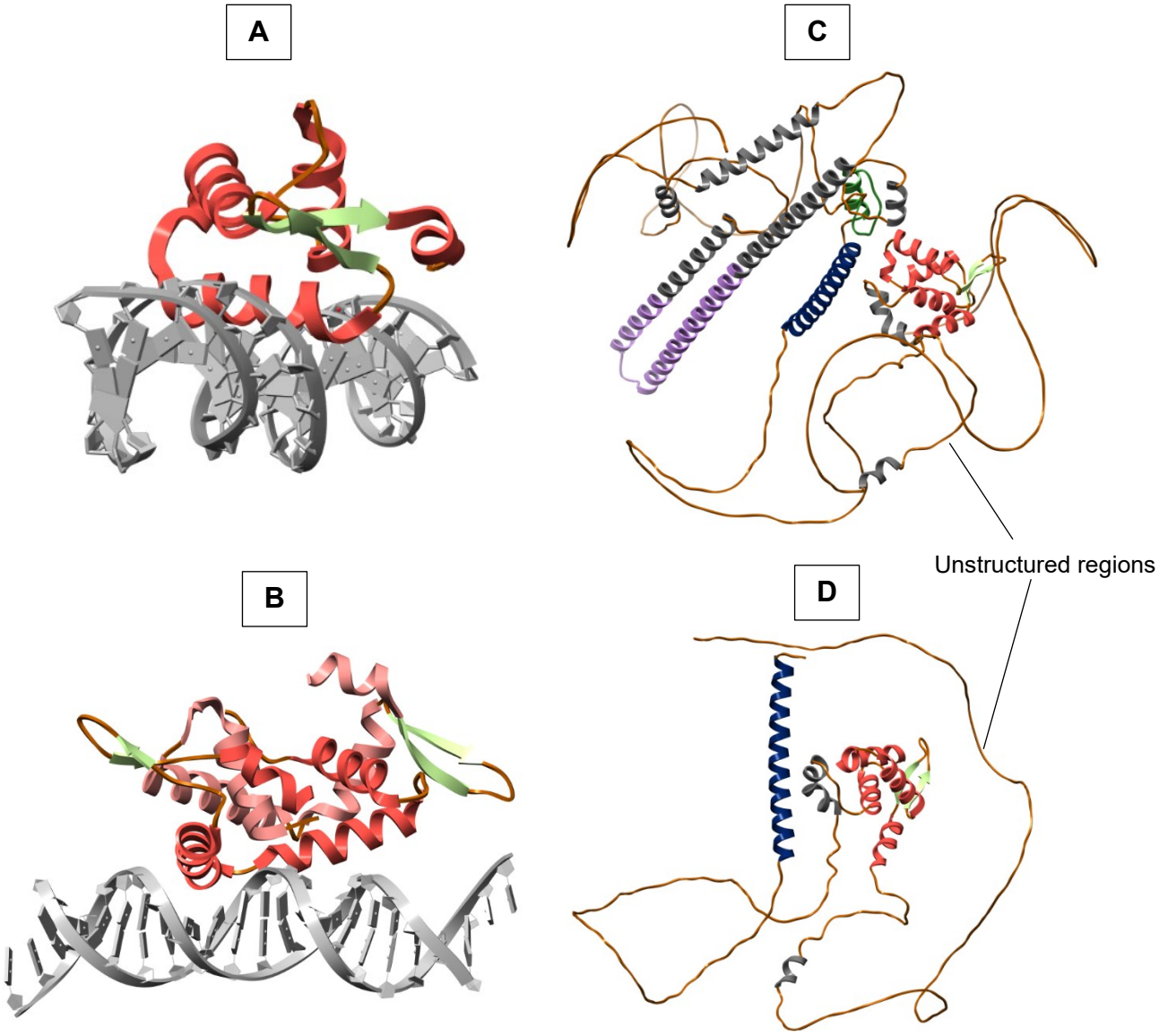


Figure 30. Comparative illustration of the isolated forkhead domain and full length FOXP2 structure.

(A) Shows the crystal structure of the monomeric forkhead domain bound to DNA and the domain-swapped dimer is shown in **(B)**, also bound to DNA (PDB: 2A07). The crystal structure of the solved forkhead domain can be directly overlaid with the forkhead domain shown in the full length FOXP2 AlphaFold structure in **(C)**, with all domains represented: forkhead domain (red), leucine zipper domain (dark blue), zinc finger domain (green) and the glutamine rich region (purple). The LZ-END construct used in this study, comprising the leucine zipper domain to the end of the FOXP2 sequence is shown in **(D)**. As shown in (C) and (D), a large proportion of the full length and LZ-END protein is unstructured and has critical roles to play in the FOXP2 protein, alongside the leucine zipper domain and forkhead domain.

Structural differences and oligomerisation potential of the LZ-END and FHD construct

Both constructs were shown to be predominantly α-helical in the absence and presence of DNA. In this study, the presence of DNA in both constructs seemed to increase the stability of the apo forms as both produced more distinct peaks upon binding to DNA. The secondary structural elements of the FHD have been characterised by X-ray crystallography (Stroud *et al.,* 2006). The FHD core comprises three stacked α-helices, with a fourth 3_{10} found between helix 2 and 3. The stacked α-helices are enclosed by an antiparallel β-sheet (Stroud *et al.,* 2006). Thus, the evidence of α-helical arrangements in the FHD crystal structure support the observed CD profile generated for the FHD. The LZ-END construct, however, has not been crystallized, albeit the FOXP2 AlphaFold structure, modelled the leucine zipper and FHD regions with high confidence. These domains were found to be predominantly helical in the AlphaFold FOXP2 model as well. Thus, the CD profiles produced for the LZ-END construct in this study corroborates the computational evidence.

Where the tertiary structure of the constructs is concerned, the tryptophan probes located within the forkhead domain indicate fluorescence emission originating from two distinct surroundings: the hydrophobic core and the outer surface. Whilst these environments are maintained in the absence and presence of DNA, the structural motifs pertinent to the LZ-END construct may facilitate intra-protein interactions or perhaps interactions with the DNA, as the fluorescence emission of the LZ-END: DNA complex was greatly quenched by the DNA. Perhaps the leucine zipper domain and unstructured regions may shield against intense emission.

This proposition is further supported by the evidence of higher-order structures that were evident from the EMSAs (See Fig 21). It is important to state that the FHD and LZ-END EMSA results are qualitative, not quantitative. They were performed to establish whether the protein constructs bind to the DNA, (which was demonstrated to be true) and therefore infer the proper functioning of the proteins. It is unfortunate that the free DNA migrated off the LZ-END EMSA, however, the gel is still able to address the issue of DNA-binding and functionality. Indeed, the EMSAs confirmed that both constructs assumed the correct fold as they were able to bind DNA.

Moreover, three different LZ-END:DNA complexes were observed on the LZ-END EMSA. This result suggests that homodimerisation is a key feature of the leucine

zipper domain, and even higher-order complexes may form when bound to DNA. The challenge remains in understanding the pathway employed to arrive at DNA-binding in the LZ-END construct, that is, understanding whether FOXP2 proteins follow the monomer or dimer pathway to achieve the observed binding and the interface of such associations. Investigating this phenomenon will provide insight into the LZ-END dimerisation interface and how the dimerisation of the leucine zipper domain is a prerequisite to DNA-binding or not and whether this happens in tandem with DNA-binding of the forkhead domain.

LZ-END oligomerization was explored by Thulo (2019), utilizing size exclusion chromatography. In this study, the oligomeric state of 5 constructs were assessed: FOXP1 LZ-END, FOXP2 LZ-END, FOXP1 LZ-FHD, FOXP2 LZ-FHD and the isolated FHD. Four constructs displayed oligomerisation and at varying degrees, however the FOXP2 LZ-END construct formed the largest quaternary complex, being octameric in size and also trimeric. The leucine zipper domain of FOXP1 and FOXP2 are identical (100% similarity), yet the FOXP1 LZ-END construct did not produce oligomers of the same size as the FOXP2 LZ-END construct, observing dimers and tetramers. Thulo (2019), attributed these differences to the unstructured regions flanking in the FHD in both constructs and proposed that these regions play a significant role in directing oligomerisation potential. The phenomenon of oligomerisation through the leucine zipper domain is not unique to LZ-END, as evidenced by similar findings in FOXP3 (Li *et al.,* 2004; Webb *et al.,* 2017) and FOXP2 (Häußermann *et al.,* 2019). However, the interactions observed in these cases did not involve DNA. Through electrophoretic mobility shift assays (EMSAs), it was confirmed that the oligomerization of the LZ-END does also occur in the presence of DNA.

In the same study by Thulo (2019), the isolated FHD was the only construct that did not oligomerise - SEC experiments showed that it remained a monomer, at physiological conditions. Notably, there is existing evidence indicating that the isolated FHD can adopt a domain-swapped dimeric configuration, albeit under conditions tailored for crystallization (Stroud *et al.,* 2006). The FHD domain-swapped dimer is capable of binding DNA, and in this study, EMSAs revealed that the isolated FHD can bind DNA as a monomer too.

DNA Binding capabilities

The LZ-END construct displayed greater affinity for DNA than the FHD construct. The forkhead domain has previously been to shown to bind DNA with weak affinity (Fanucchi and Morris, 2016), however, the fact that it is the primary means of binding one would assume it would display a greater affinity for DNA as compared to the LZ-END construct, but this is not so. This infers that the leucine zipper domain and perhaps the flanking disordered regions may support DNA-binding. According to the crystal structure of the isolated forkhead domain, helix-3 inserts into the major groove of the DNA. To date, this is the only known interface for binding where FOXP2 domains are concerned. However, the increased binding affinity shown for the LZ-END construct may cooperate with the forkhead domain and unstructured regions to support binding. Indeed, whilst the addition of the leucine zipper domain may seem to improve the affinity of FOXP2-DNA interaction, the forkhead domain plays an important role in directing the specificity of the DNA-binding interaction (Fanucchi and Morris, 2016, Webb *et al.,* 2017). DNA-binding specificity refers to the ability of protein to discriminate between different DNA sequences and bind selectively to its target sequence. Certainly, this finding helps to better distinguish the roles of the leucine zipper domain and the forkhead domain, which apportions the role of specificity in DNA-binding to the forkhead domain and enhanced affinity to the leucine zipper domain

Key structural features of the LZ-END construct interact with DNA

The conformational dynamics of the LZ-END construct was quantified using HDX MS, both in the presence and absence of cognate DNA (Nelson DNA). HDX MS has been used before to study the isolated FOXP2 FHD construct (Morris *et al.,* 2018). By assessing the conformational changes of the larger LeuZip variant, which includes the forkhead domain and its flanking regions, one can determine the effect of these motifs on the forkhead domain and more specifically, account for changes that occur to the construct when interacting with DNA. The leucine zipper has been implicated in DNA-binding, thus by analysing its interaction with DNA in the presence of the DNA-binding forkhead domain, one can gain a comprehensive understanding of how these structural elements work together, dynamically to ensure DNA binding.

In this study, the forkhead domain of the LZ-END construct (in the complex form) did not exhibit as great levels of protection from deuterium exchange, as reported for the isolated forkhead domain in the study by Morris (2018), in fact, shielding from exchange was observed at a maximum of 70% and this was reduced to 30% by the full duration of exchange (3600 s). Notably, while an increased level of shielding was observed in the forkhead domain of the LZ-END complex as compared to the apo form, it was not as buried as reported in the comparative study. In particular, helix-1 and helix-2 displayed a large degree of dynamism even in the complexed form, whereas in the isolated forkhead domain (Morris *et al.,* 2018), these helices were greatly shielded throughout the duration of exchange. In fact, helix-2 seemed to become more dynamic as helix-3 became less dynamic with the final structure existing in an open conformation as previously described by Morris (2018)

While helix 3 of FOXP2 is typically identified as the DNA-recognition helix, it's worth noting that helix 1 and helix 2 also engage with DNA and helix 3 (Stroud *et al.,* 2006). In the FOXP2 crystal structure, Tyr509 from helix 1, as well as Leu527 and Tyr531, establish interactions with helix 3 and the sugar-phosphate backbone (Stroud *et al.,* 2006). These interactions are instrumental in promoting the binding of H3 to DNA by inserting into the major groove.

The HDX results in this study affirm these interactions, particularly in helix 2. Notably, Leu527 and Tyr531 displayed protection from exchange throughout the entire exchange period in the complex form. This protection from exchange strongly suggests their involvement in interactions with helix 3 and the DNA phosphate backbone. However, when it comes to helix 1, no clear parallels between the DNA-binding interactions observed in the solved crystal structure (at the amino acid level) and the residue dynamics portrayed in this study could be drawn. However, the HDX results from this study reveal that Ala516, Ile517, and Met518 may engage with DNA or other FOXP2 components upon DNA binding, as they too exhibited protection from exchange throughout the entire exchange period.

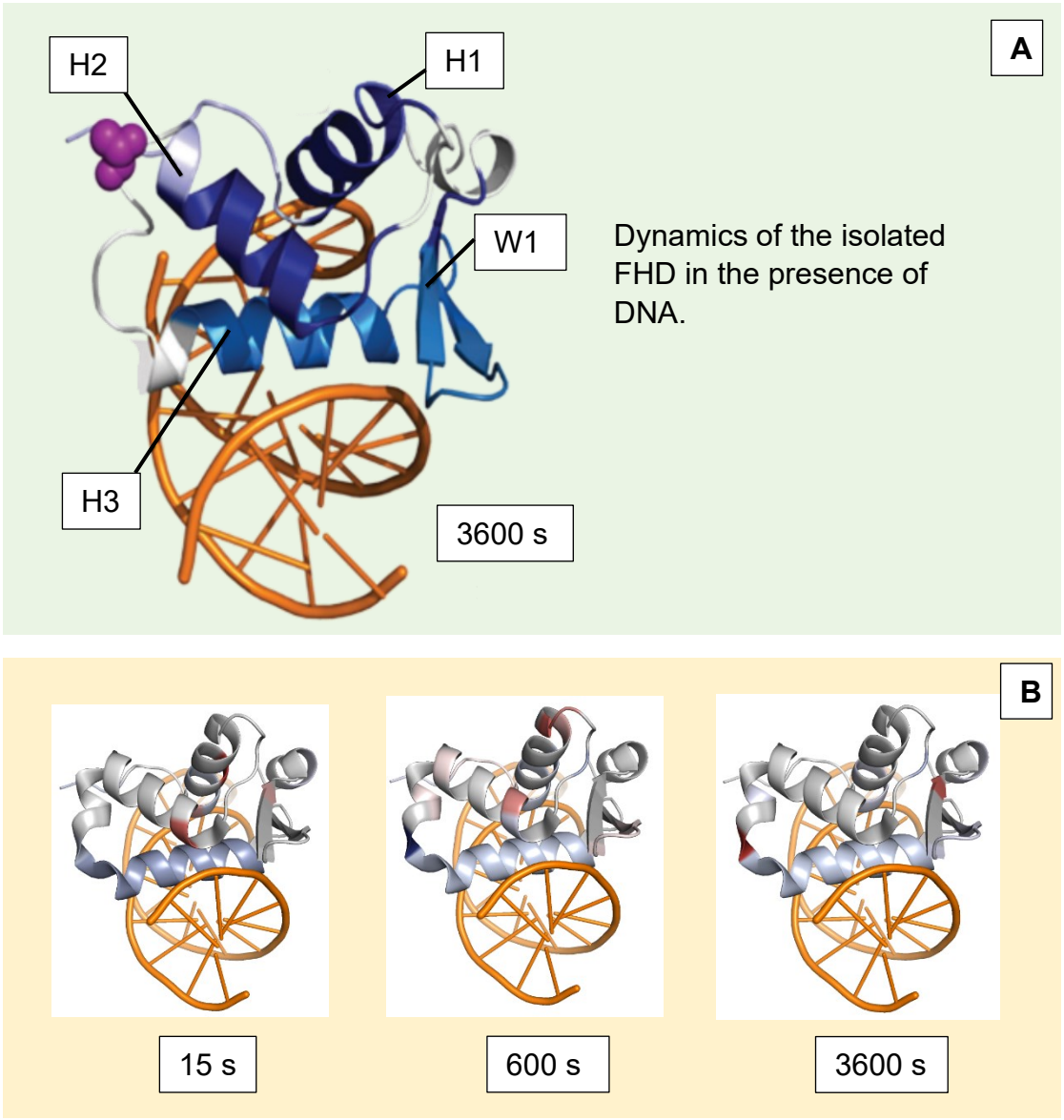


Figure 31. Comparing the conformational dynamics of the isolated forkhead domain (by Morris *et al.,* 2018) and the LZ-END construct in the presence of DNA.

(A) In a study by Morris *et al.,* 2018, hydrogen deuterium exchange mass spectrometry was performed for the isolated FOXP2 forkhead domain and the dynamics were compared to an A539P mutant (which disabled dimer formation). The conformational dynamics of the DNA-bound FHD was mapped to the solved structure (PDB: 2A07) and is shown in the green block at the full duration of exchange with all secondary structure features shown to be highly protected from exchange in the presence of DNA. **(B)** Comparatively, in this study only helix-3 displays a significant amount of protection, whilst helix-1 and helix-2 seem to show almost no change in dynamics in the presence of DNA, with only certain regions displaying an abstruse dynamism.

The difference maps revealed that the leucine zipper domain displayed fluctuations between levels of high dynamism and protection from exchange at the amino acid level. However, the overall local dynamics of the leucine zipper domain remained unchanged in the absence and presence of DNA. In FOXP3 the leucine zipper domain is shown to be necessary for DNA binding (Leng *et al.,* 2022). One would assume that the conformational dynamics of the FOXP3 leucine zipper domain allow for it to be in closer proximity to DNA or the forkhead domain in a manner that supports DNA binding. The leucine zipper in the FOXP2 protein show that this type of interaction may not necessarily exist in this manner and perhaps the presence of DNA does not induce conformational changes in the leucine zipper domain, but this does not negate the advent of potential DNA binding. This notion is put forward in light of the decreased dynamics observed for the FHD in the LZ-END construct as compared to the dynamics of the isolated FHD reported by the 2018 study. The reduced level of dynamics seen in the FHD domain may be attributed to the unstructured regions and leucine zipper domain that may be implicated in DNA-binding.

The conformational dynamics of the interlinking loop region (ILR) and acid-rich tail confirm the speculations about the role of these regions in DNA binding. Indeed, the ILR displayed increased levels of protection in the DNA complex form, in that it became 90% shielded from exchange. This infers that the ILR may interact with DNA and concurrently cooperate with the forkhead domain DNA-binding as postulated by Thulo *et al.* 2019. The dynamics of the acid-rich tail was reduced in the complex form but only marginally so. There is not enough evidence to support its participation in DNA binding as suggested by Thulo *et al.* 2019. The DNA backbone has an overall negative charge, thus interactions with the acid-rich tail is difficult to justify based on the fact that electrostatic repulsions would be commonly observed at such an interface. However, these regions do add to the complexity of the DNA-binding mechanism observed and may support theory that disordered regions, especially in transcription factors, become more ordered when bound to a ligand, or in this context, DNA.

The LZ-END displays a 3-state mechanism of unfolding and the FHD a 2-state mechanism of unfolding

Comparing the stabilities and unfolding mechanisms at the secondary and tertiary level helped to elucidate the impact of the leucine zipper domain and the forkhead domain on FOXP2 stability and allowed their folding mechanisms to be assessed. As the tryptophan probes are exclusively located within the forkhead domain, the hypothesis was that any differences in the unfolding curves of both constructs at the tertiary level could be ascribed to the structural integrity of the leucine zipper domain. Interestingly, both constructs revealed to follow a 2-state unfolding pathway (N→D) at the tertiary level. However, according to unfolding at the secondary level, the LZ-END was shown to follow a 3-state unfolding pathway (N→ I→ D). This pathway produced an LZ-END intermediate that was undetected at the tertiary level. Since the unfolding of the FHD construct at both levels conforms to a 2-state model, it implies that the intermediate is likely formed by the leucine zipper domain which is anticipated for multi-domain construct (Pace *et al.,* 2000).

Furthermore, the stability of the constructs was compared to probe whether the lack or presence of the leucine-zipper domain and FHD flanking regions would enhance FOXP2 stability. For the FHD, the denuration midpoints (6.2 M and 5.9 M) at the secondary and tertiary level are comparable to that observed for the LZ-END (6.2 M) at the tertiary level. This is in agreement with the notion that for both constructs (at the tertiary level), the unfolding events of the FHD is solely accounted for. That being said, the Cm for the I→D transition of the LZ-END is slightly lower (5.6 M) in comparison to the aforementioned Cm's, with the Cm of the near-native intermediate differing greatly from the rest. Therefore, the stability of the FHD (N→D) and LZ-END (I→D) reveals that the FHD is marginally more stable.

Comparing the ΔG_{H_2O} produced by each construct revealed that at the tertiary level the FHD (8.63 kJ.mol-1) is significantly less stable than the LZ-END (19.89 kJ.mol^{-1}) and this is consistent with that observed at the secondary level. This may imply that the leucine zipper domain adds an extra level of complexity to the unfolding events observed for the LZ-END as compared to the isolated forkhead domain. Prior investigations using EMSAs had demonstrated the oligomeric capacity of the LZ-END. Perhaps the significant difference in the ΔG_{H_2O} (between both constructs, at both levels) could be explained by the energetically demanding disruption of the bonds

formed at the dimerization interface of the LZ-END construct. Indeed, the bond formation at typical leucine zipper interface is quite complex. The leucine side chains interdigitate, forming a stable coiled-coil structure that holds the dimer together (Tropsha *et al.,* 1991; Alber, 1992; Tarazona *et al.*, 2019). The coiled-coil structure is stabilized by hydrophobic interactions between the leucine side chains and by inter-helical electrostatic interactions between oppositely charged residues (Hakoshima, 2001; Tarazona *et al.*, 2019). Thus, disrupting this complex network in the FOXP2 leucine zipper domain and FHD in the LZ-END complex may require more energy than that required for the FHD alone.

Conclusion

The research undertaken in this study sought to compare the DNA-binding capabilities, conformational dynamics, and stability of the FOXP2 forkhead domain and leucine zipper domain. Owing to the functional and structural differences that exist between these two key domains, their individual and cooperative impact on FOXP2 structure and function was assessed in the absence and presence of DNA. Indeed, both domains behave differently in the presence of DNA and these individual differences enlighten their specific roles within the FOXP2 protein. This study characterised, for the first time, the structural conformation of the leucine zipper domain and interlinking loop region in the presence of DNA and the data suggests that interlinking loop region may be implicated in DNA-binding. Additionally, the FOXP2 folding mechanisms and stability for each domain was characterized, revealing a 2-state mechanism for the FHD construct and a 3-state mechanism for the LZ-END construct. Fluorescence anisotropy studies showed that the FHD binds DNA with weaker affinity as compared to the LZ-END construct which may again be supported by the presence of the leucine zipper domain and possibly unstructured regions. Endeavoring to understand the structural attributes of these domains enlightens our understanding of proper FOXP2 functioning that occurs *in vivo.* The significance of investigations like this one is highlighted when elucidating mechanisms by with improper FOXP2 leads to disease, such as developmental verbal dyspraxia, autism spectrum disorder, and Alzheimer's Disease.

A: Heat maps indicating the level of isotopic exchange at 15s, 600s, and 3600s for the apo and DNA-bound forms of the LZ-END construct

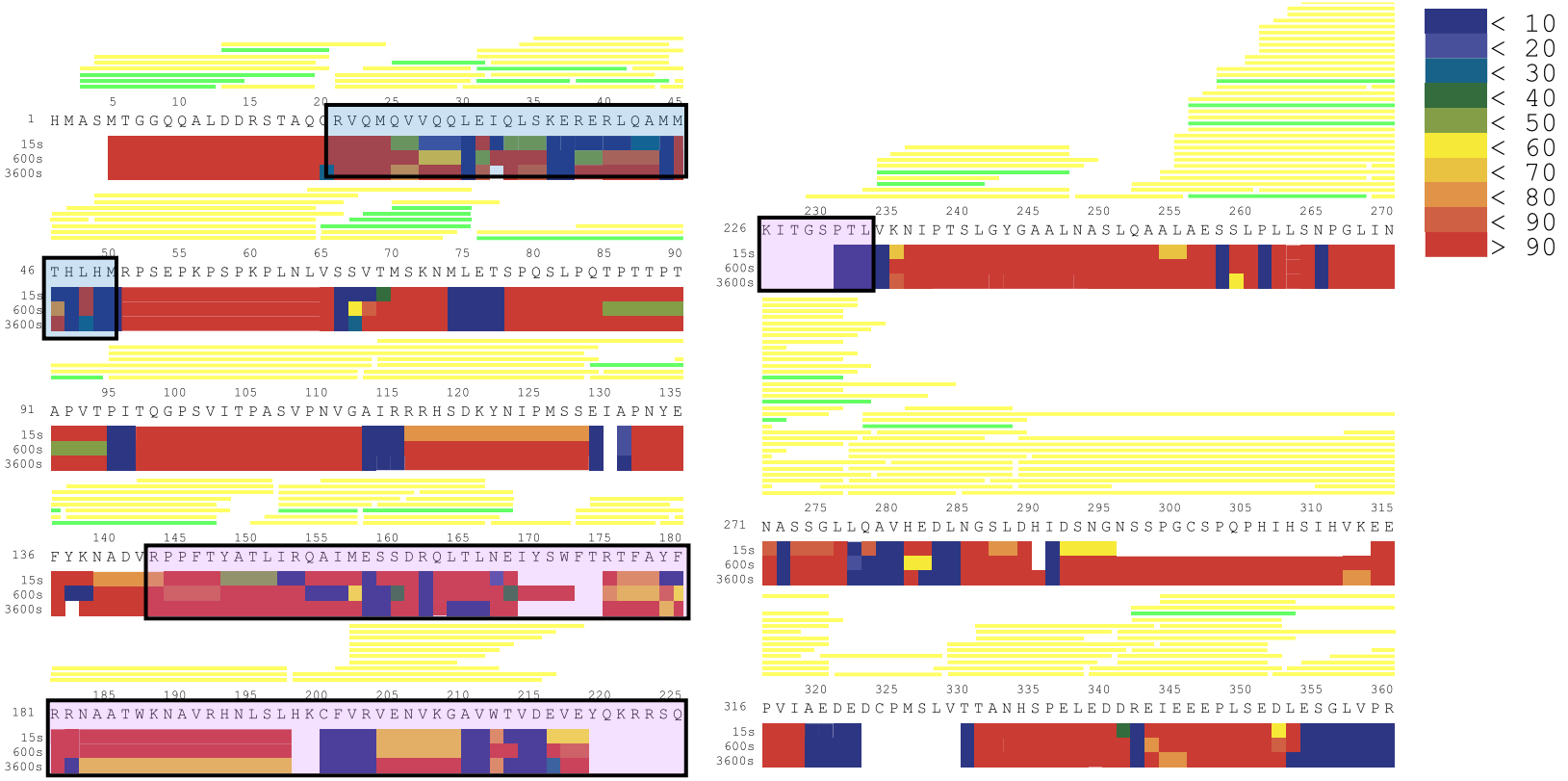


Figure 32. Heat map illustrating the backbone dynamics of the apo LZ-END protein.

The time-resolved level of isotopic exchange for the apo form of the LZ-END construct is shown. The colour spectrum indicates the level of isotopic exchange, i.e., regions that are highly dynamic or solvent accessible are indicated in red and regions that are more buried or inaccessible to solvent or to exchange are indicated in blue. The exchange is measured over 15, 600, and 3600 seconds for the entire LZ-END construct and for each residue, where there is enough sequence coverage. Sequence coverage is depicted by the yellow (medium coverage) and green (high coverage) bars. The leucine zipper domain (blue box) and forkhead domain (purple box) displays varying levels of dynamism, with certain residues resistant to exchange throughout the full time period.

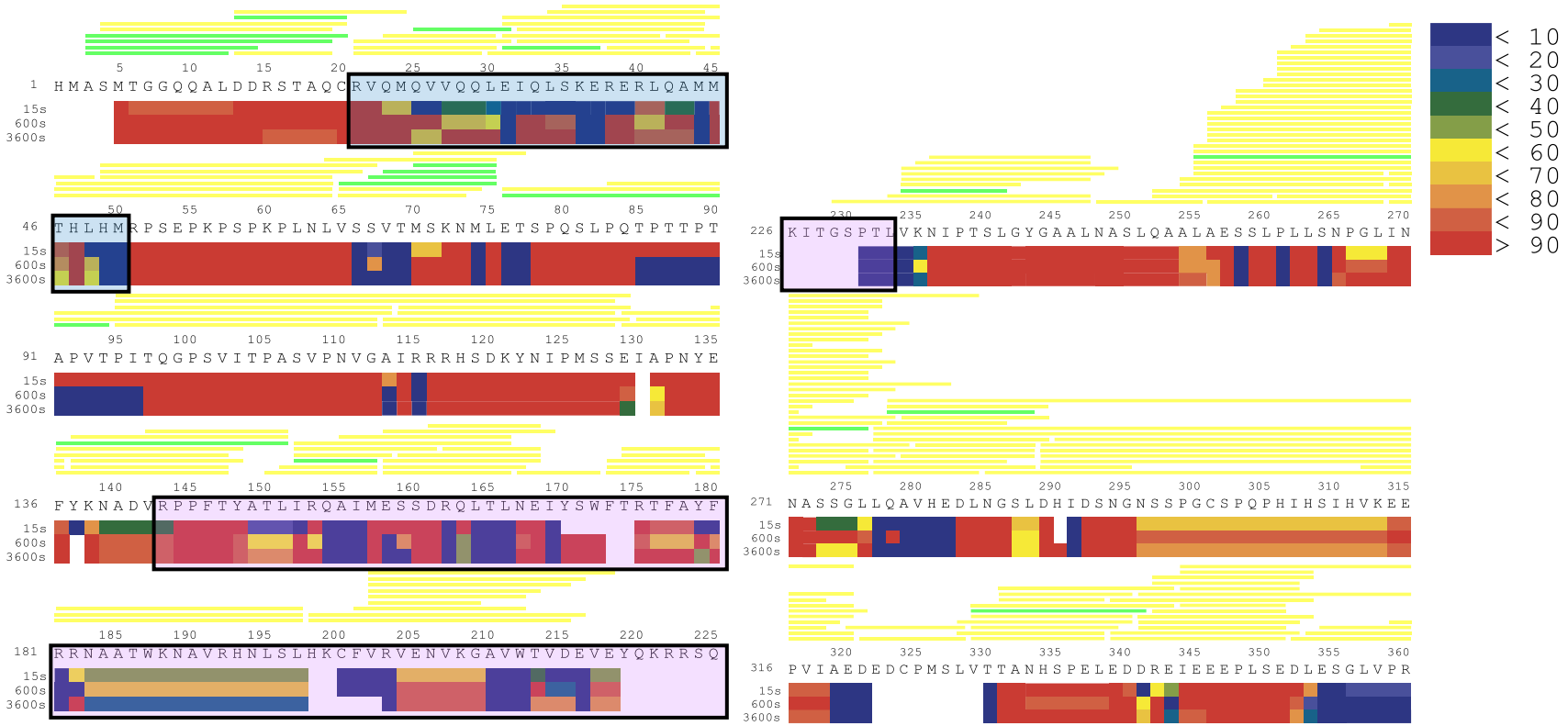


Figure 33. Heat map illustrating the backbone dynamics of the LZ-END protein in complex with DNA.

The time-resolved level of isotopic exchange for the complex form of the LZ-END construct is shown. Again, the colour spectrum indicates the level of isotopic exchange, i.e., regions that are highly dynamic or solvent accessible are indicated in red and regions that are more buried or inaccessible to solvent or to exchanging are indicated in blue. The exchange is measured over 15, 600, and 3600 seconds, for the entire LZ-END construct. Sequence coverage is depicted by the yellow (medium coverage) and green (high coverage) bars. The leucine zipper domain (blue box) seems to indicate similar levels of exchange as the apo form, showing varying levels of flexibility. Strikingly, helix-3 of the forkhead domain (purple box) shows a decreased level of exchange which can be directly attributed to the binding of DNA.

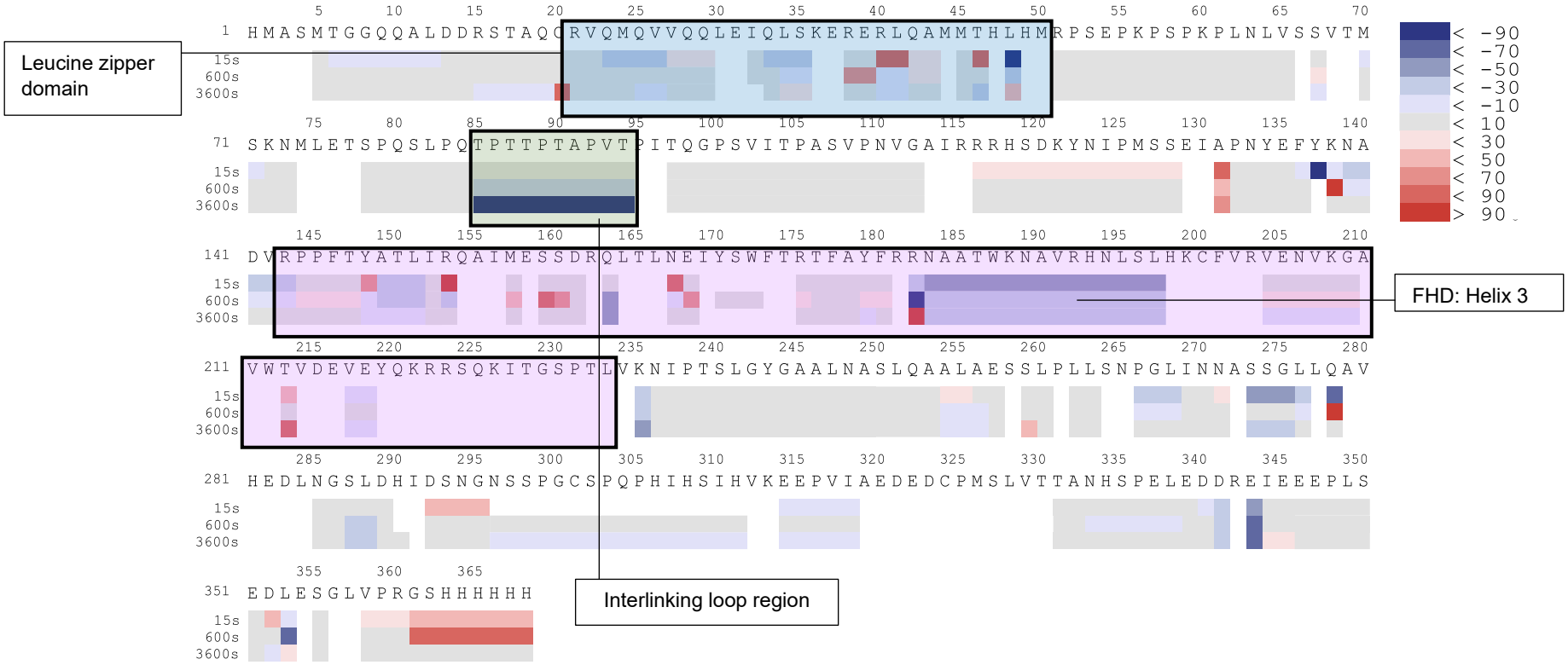


Figure 34. Difference map illustrating the backbone dynamics of the LZ-END protein in the absence and presence of DNA.

The difference map highlights the portions of the LZ-END apo construct that are protected or highly dynamic in comparison to the LZ-END: DNA complex. The colour spectrum indicates the level of isotopic exchange, i.e., regions that are highly dynamic or solvent accessible (red) and regions that are more buried or inaccessible to solvent or to exchanging (blue). The exchange is measured over 15, 600, and 3600 seconds, for the entire LZ-END construct. The leucine zipper domain is shown in blue box, the forkhead domain in a purple box and the highly protected Interlinking loop region in between both domains is shown in a green box. Helix-3 of the forkhead domain is shown to be greatly protected from exchange, whilst the leucine zipper domain displays a great level of flexibility.

B: Full LZ-END construct difference maps

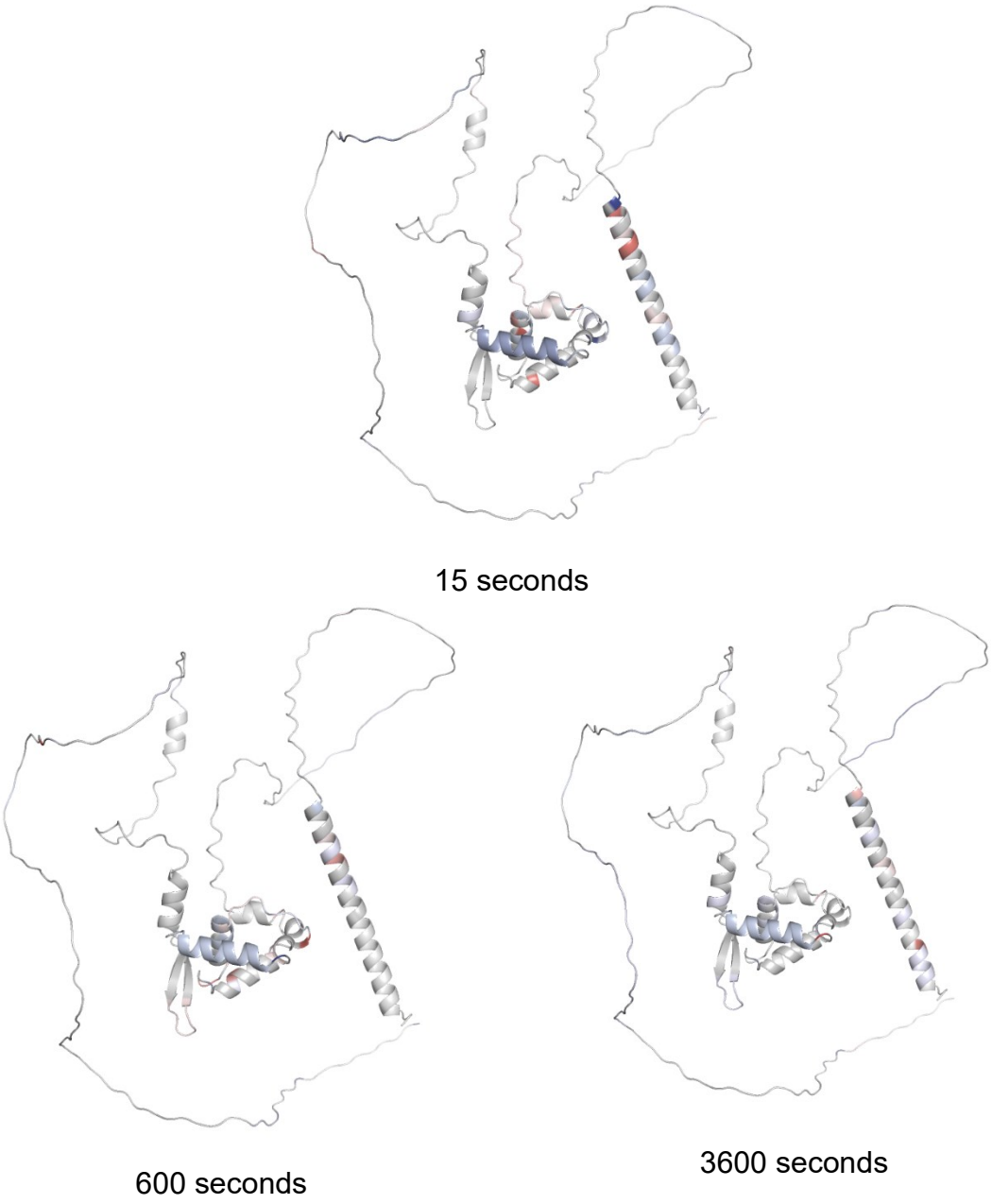


Figure 35. Comparison of the time-lapsed conformational dynamics of the full FOXP2 LZ-END construct in the absence and presence of DNA.

The difference maps were mapped to the LZ-END portion of the FOXP2 AlphaFold structure, revealing the mosaic dynamics of the leucine zipper domain, the binding of DNA to helix-3 (rendering it buried) and the dynamics of the interlinking loop region that was found to be protected from exchange.

C: Comparative deuterium uptake plots for the ILR

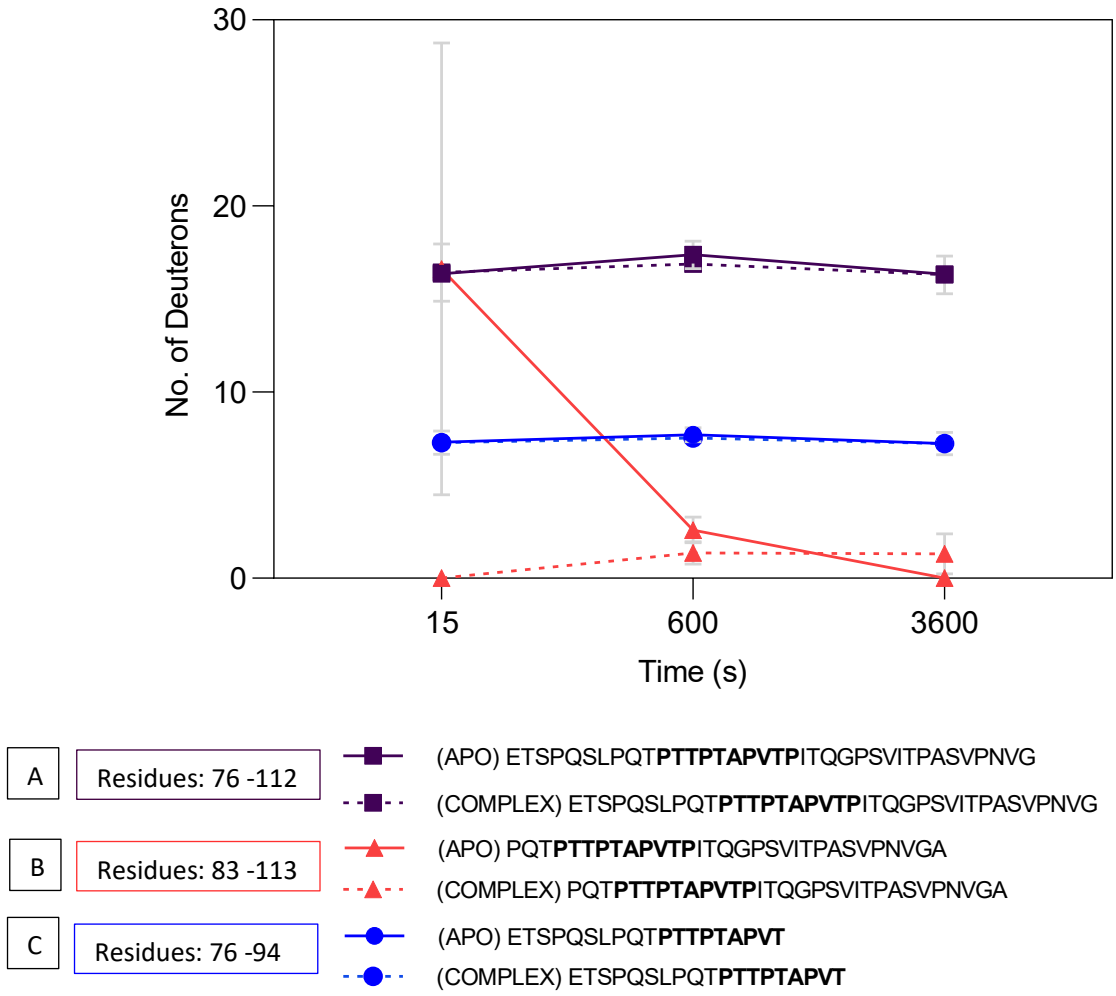
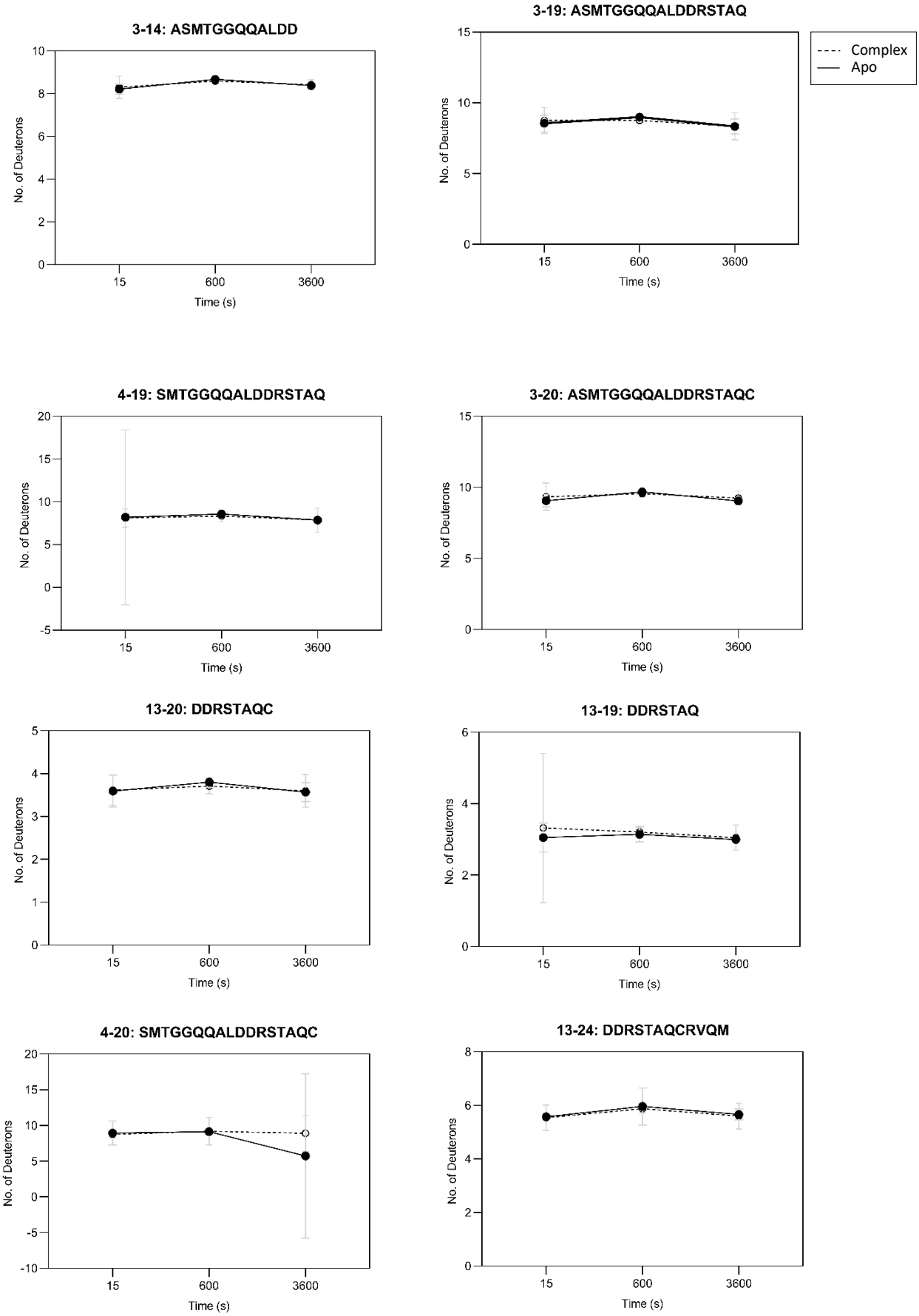
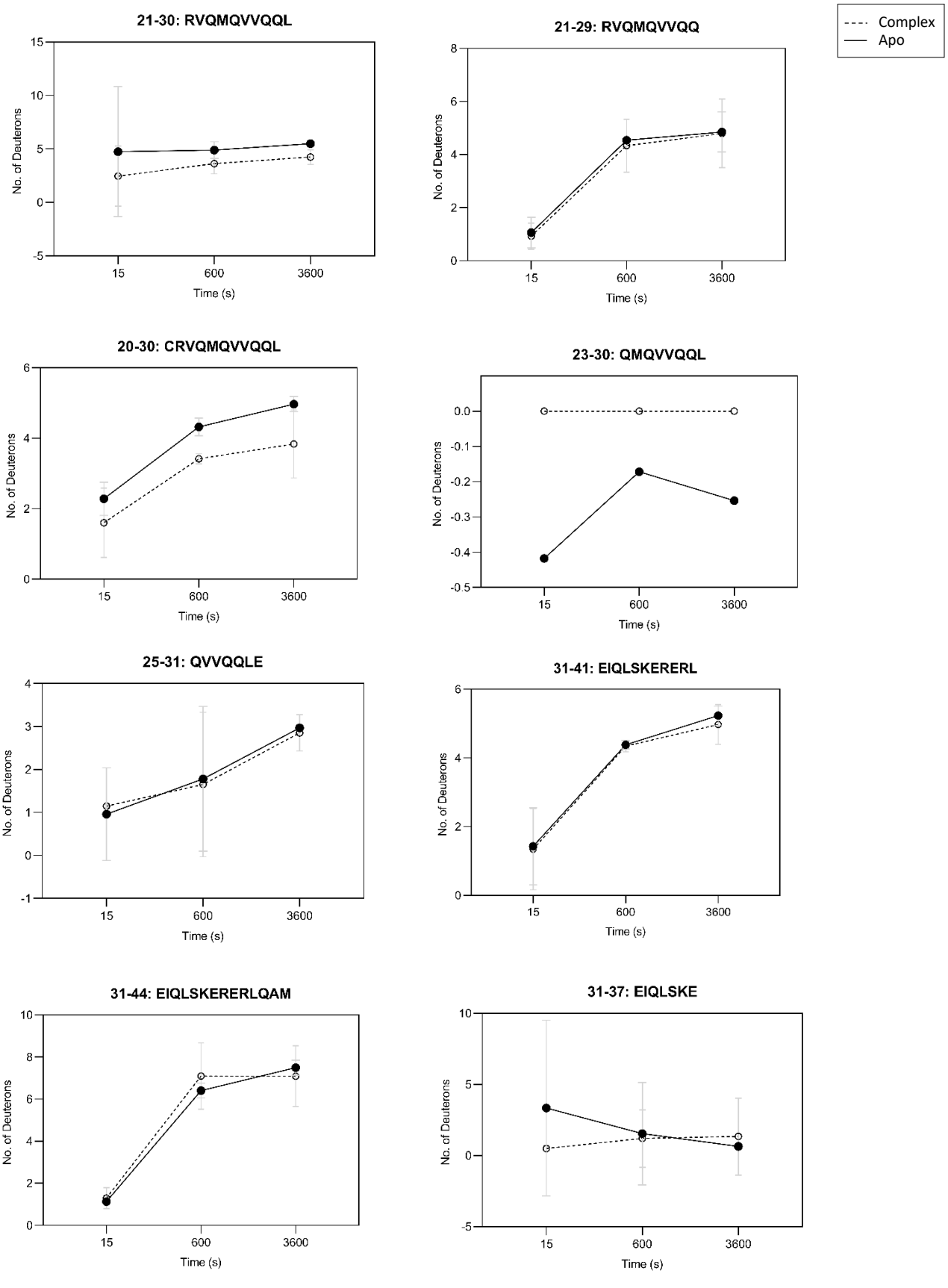


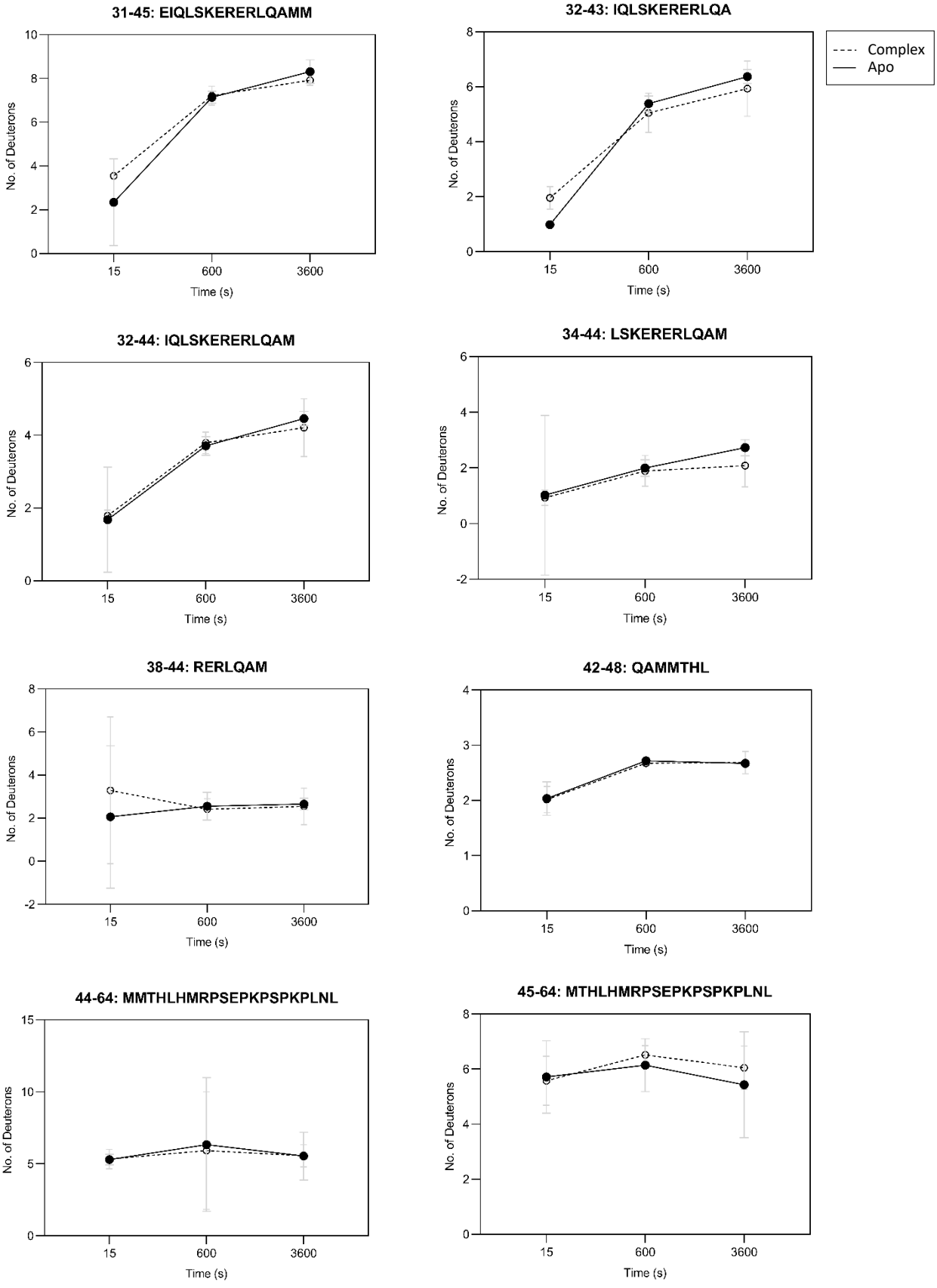
Figure 36. Deuterium uptake plots for the unstructured interlinking loop containing the *PTTPTAPVTP* sequence.

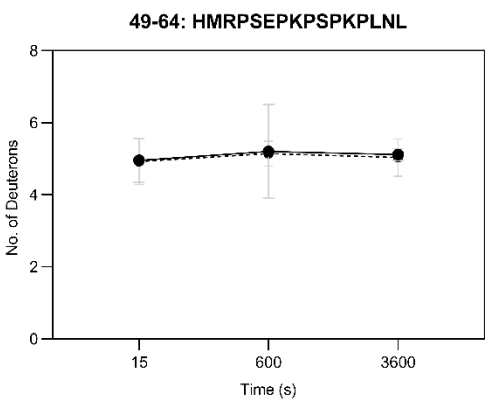
Three peptides containing the undigested *PTTPTAPVTP* sequence of the IRL were generated and the deuterium uptake for each peptide is shown above for both the apo and complex samples. The apo and complex plots for peptide A (■) hardly differ in deuterium uptake, with the complex slightly deviating from the apo at 600s. For peptide B (●) the apo and complex uptake plots can be overlaid, however it has 10 deuterons less than peptide A. In Peptide C (▲) at 15 sec the apo complex is highly dynamic (uptake of 16 deuterons), but this drastically decreases over time. In the complex form, peptide C displays very limited uptake, the most of which is seen at 3600s (1 deuteron).

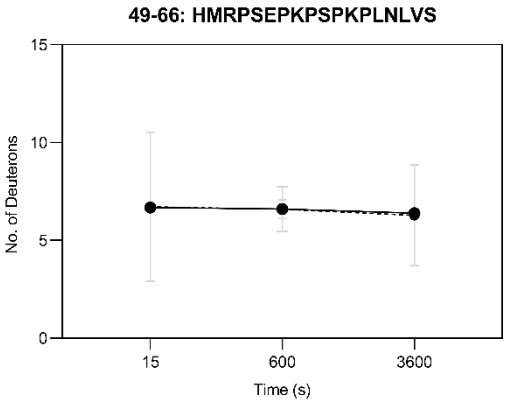
D: Deuterium uptake plots for the LZ-END in the absence and presence of DNA

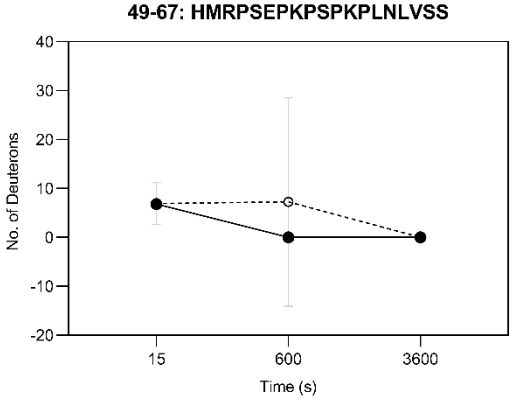


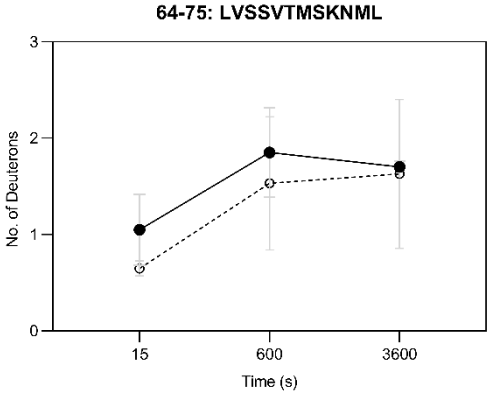

Complex
Apo
21-30: RVQMQVVQQL
21-29: RVQMQVVQQ
20-30: CRVQMQVVQQL
23-30: QMQVVQQL
25-31: QVVQQLE
31-41: EIQLSKERERL
31-44: EIQLSKERERLQAM
31-37: EIQLSKE
No. of Deuterons
Time (s)
15
600
3600

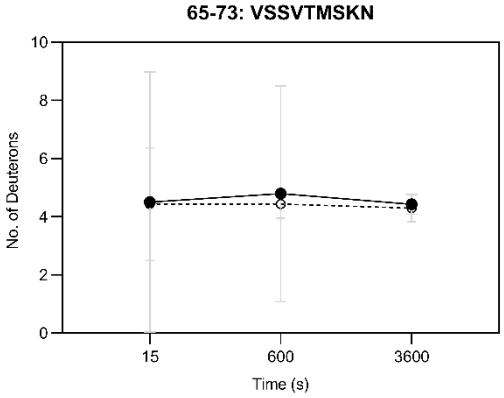

31-45: EIQLSKERERLQAMM
32-43: IQLSKERERLQA
Complex
Apo
32-44: IQLSKERERLQAM
34-44: LSKERERLQAM
38-44: RERLQAM
42-48: QAMMTHL
44-64: MMTHLHMRPSEPKPSPKPLNL
45-64: MTHLHMRPSEPKPSPKPLNL
No. of Deuterons
Time (s)
15
600
3600

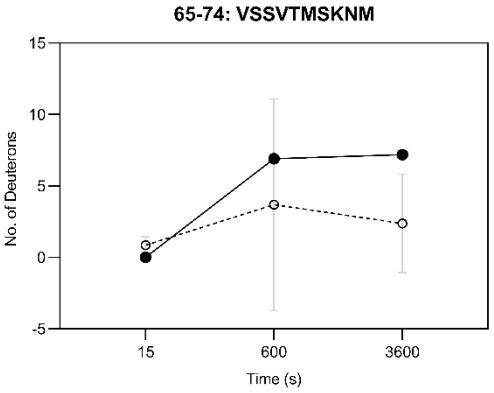

49-64: HMRPSEPKPSPKPLNL
No. of Deuterons
Time (s)
15
600
3600

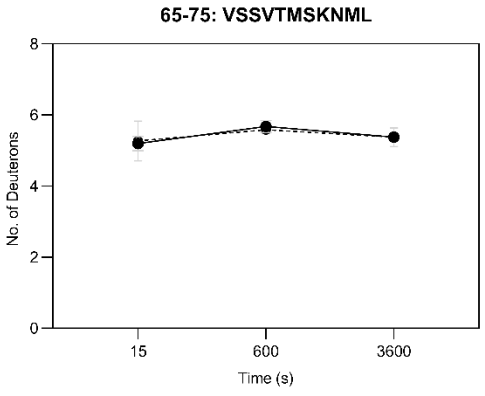

49-66: HMRPSEPKPSPKPLNLVS
No. of Deuterons
Time (s)
Complex
Apo

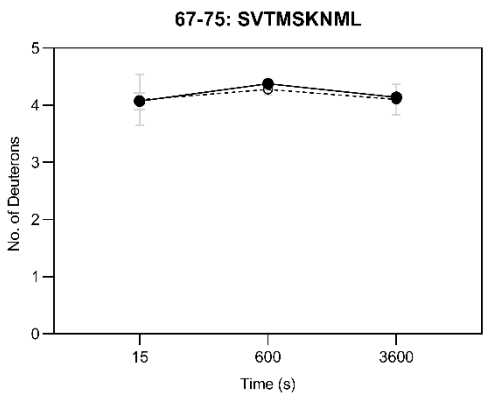

49-67: HMRPSEPKPSPKPLNLVSS
No. of Deuterons
Time (s)

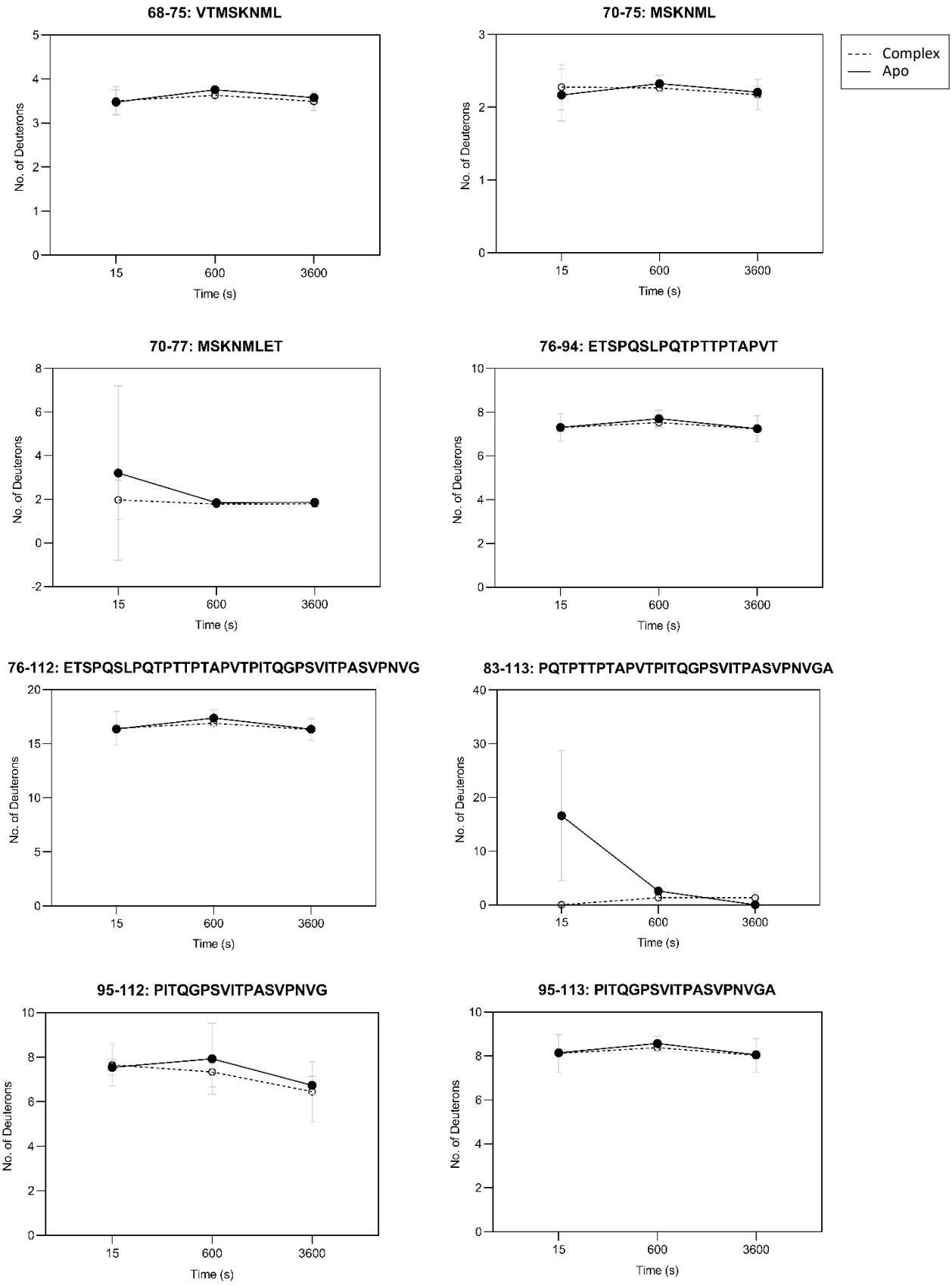

64-75: LVSSVTMSKNML
No. of Deuterons
Time (s)


65-73: VSSVTMSKN
No. of Deuterons
Time (s)

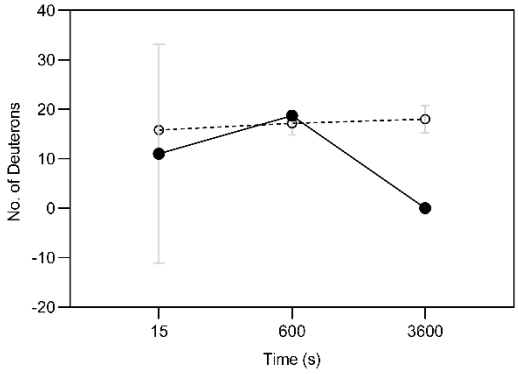

65-74: VSSVTMSKNM
No. of Deuterons
Time (s)


65-75: VSSVTMSKNML
No. of Deuterons
Time (s)

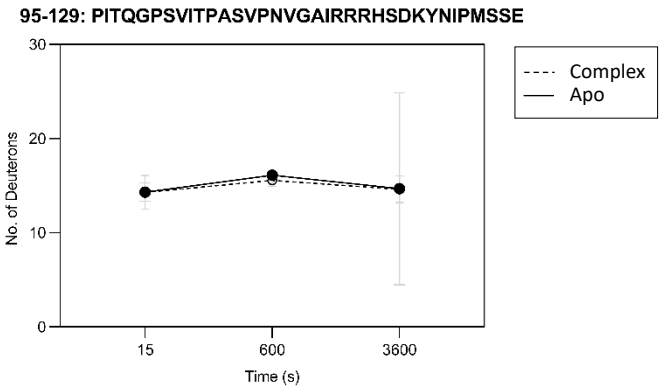

67-75: SVTMSKNML
No. of Deuterons
Time (s)


68-75: VTMSKNML
70-75: MSKNML
Complex
Apo
70-77: MSKNMLET
76-94: ETSPQSLPQTPTTPTAPVT
76-112: ETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVG
83-113: PQTPTTPTAPVTPITQGPSVITPASVPNVGA
95-112: PITQGPSVITPASVPNVG
95-113: PITQGPSVITPASVPNVGA
No. of Deuterons
Time (s)
15
600
3600

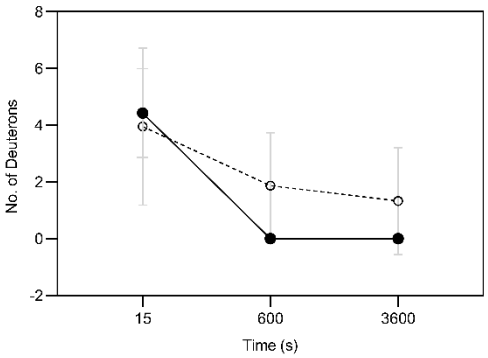
95-128: PITQGPSVITPASVPNVGAIRRRHSDKYNIPMSS



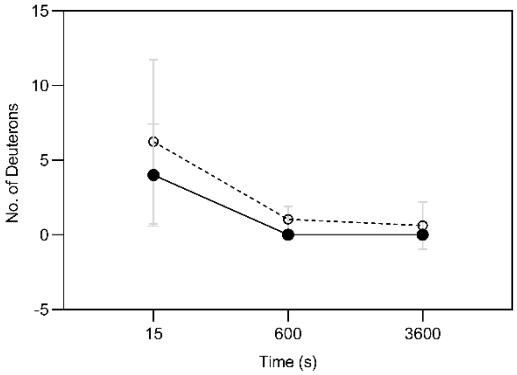
95-129: PITQGPSVITPASVPNVGAIRRRHSDKYNIPMSSE



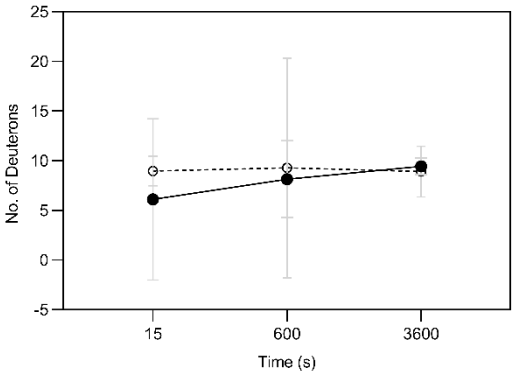
113-128: AIRRRHSDKYNIPMSS



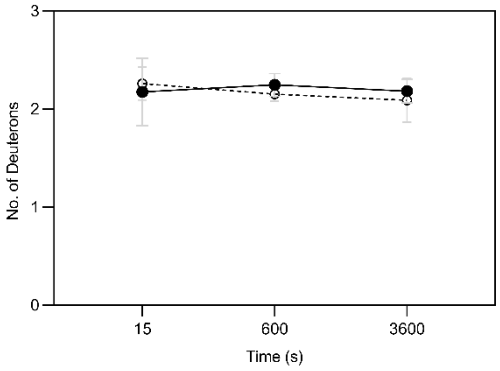
113-129: AIRRRHSDKYNIPMSSE



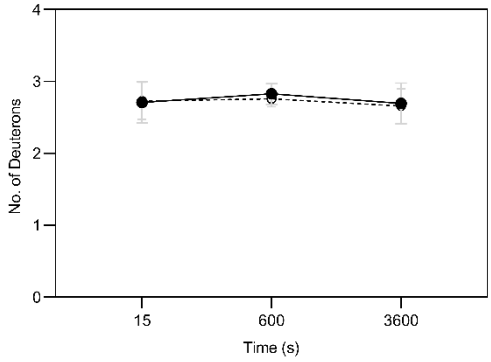
114-129: IRRRHSDKYNIPMSSE



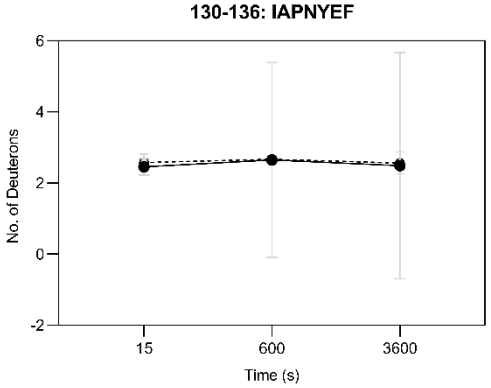
129-135: EIAPNYE

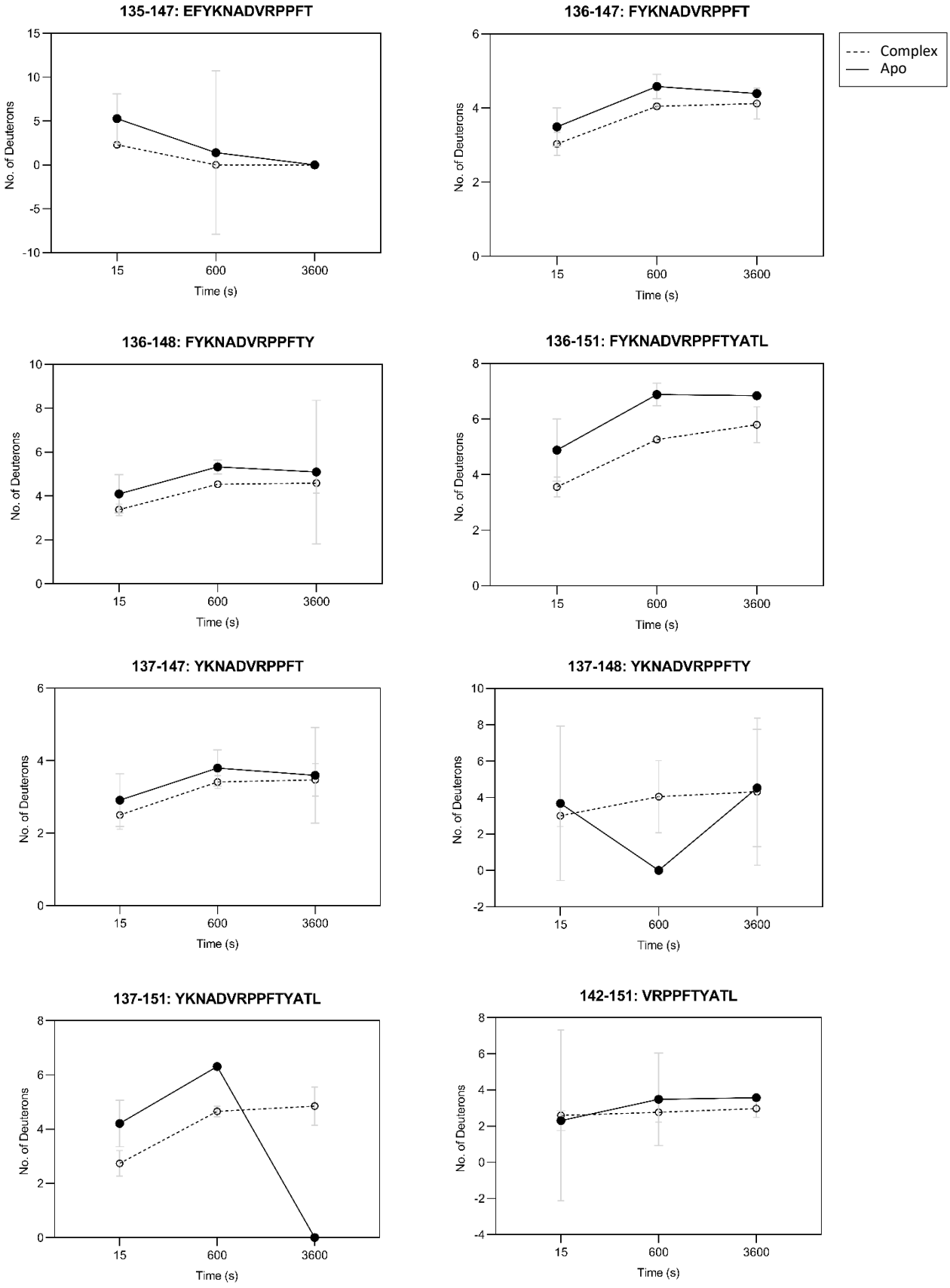


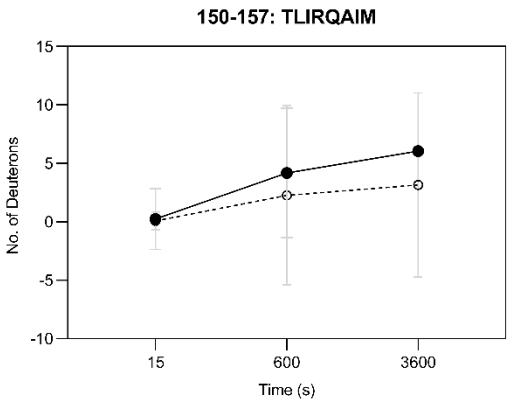
129-136: EIAPNYEF

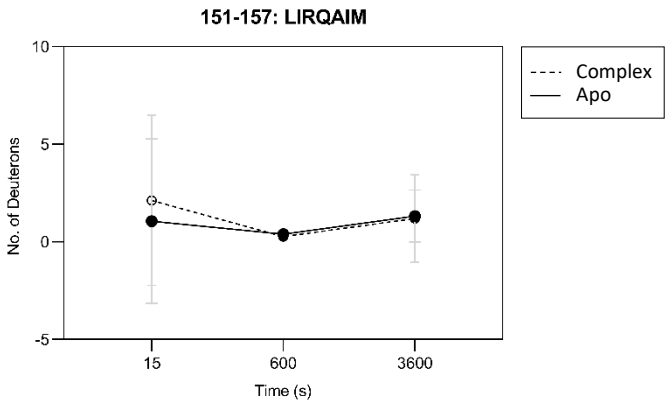


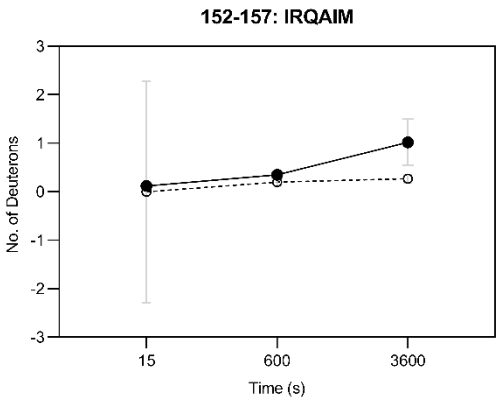
130-136: IAPNYEF

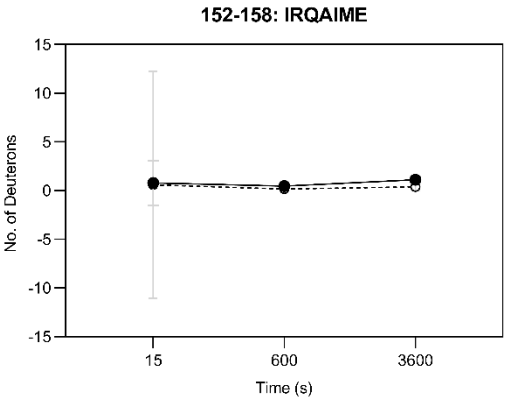


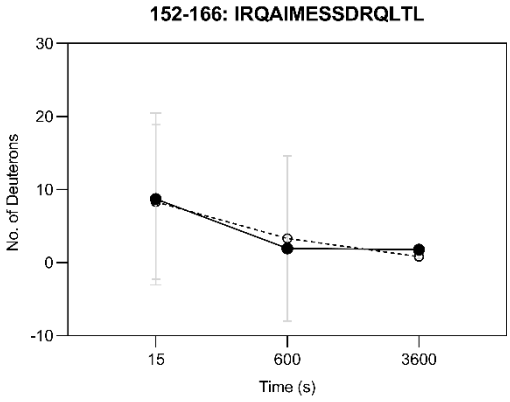

135-147: EFYKNADVRPPFT
136-147: FYKNADVRPPFT
Complex
Apo
136-148: FYKNADVRPPFTY
136-151: FYKNADVRPPFTYATL
137-147: YKNADVRPPFT
137-148: YKNADVRPPFTY
137-151: YKNADVRPPFTYATL
142-151: VRPPFTYATL
No. of Deuterons
Time (s)
15
600
3600

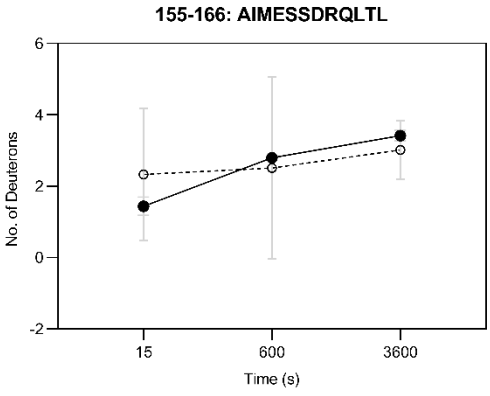

150-157: TLIRQAIM
No. of Deuterons
15
10
5
0
-5
-10
15
600
3600
Time (s)

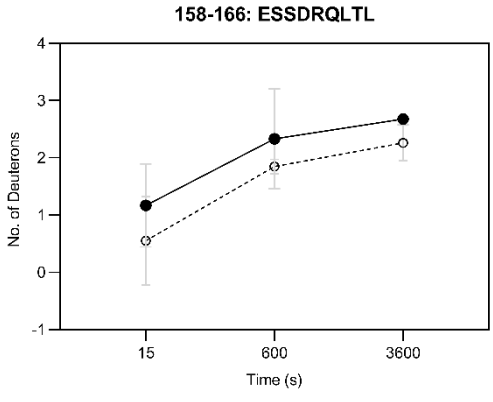

151-157: LIRQAIM
No. of Deuterons
10
5
0
-5
15
600
3600
Time (s)
Complex
Apo

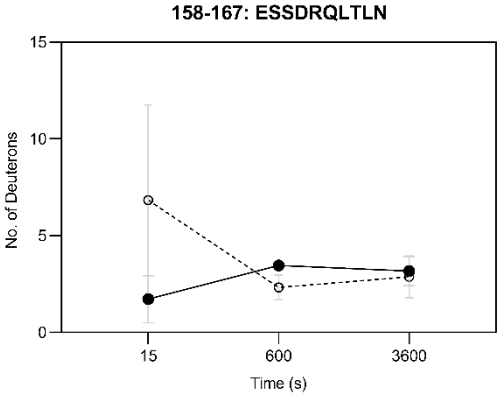

152-157: IRQAIM
No. of Deuterons
3
2
1
0
-1
-2
-3
15
600
3600
Time (s)

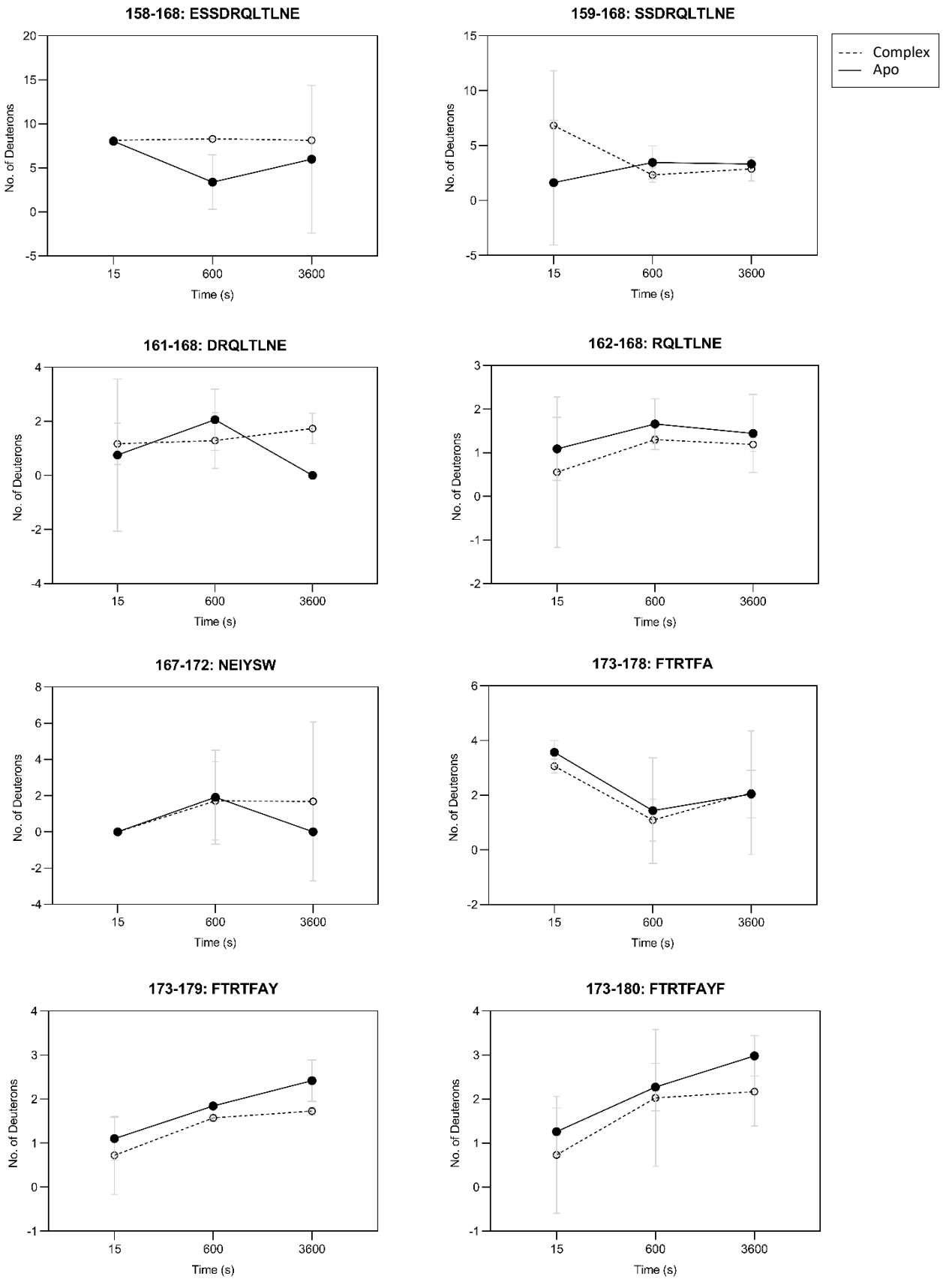

152-158: IRQAIME
No. of Deuterons
15
10
5
0
-5
-10
-15
15
600
3600
Time (s)

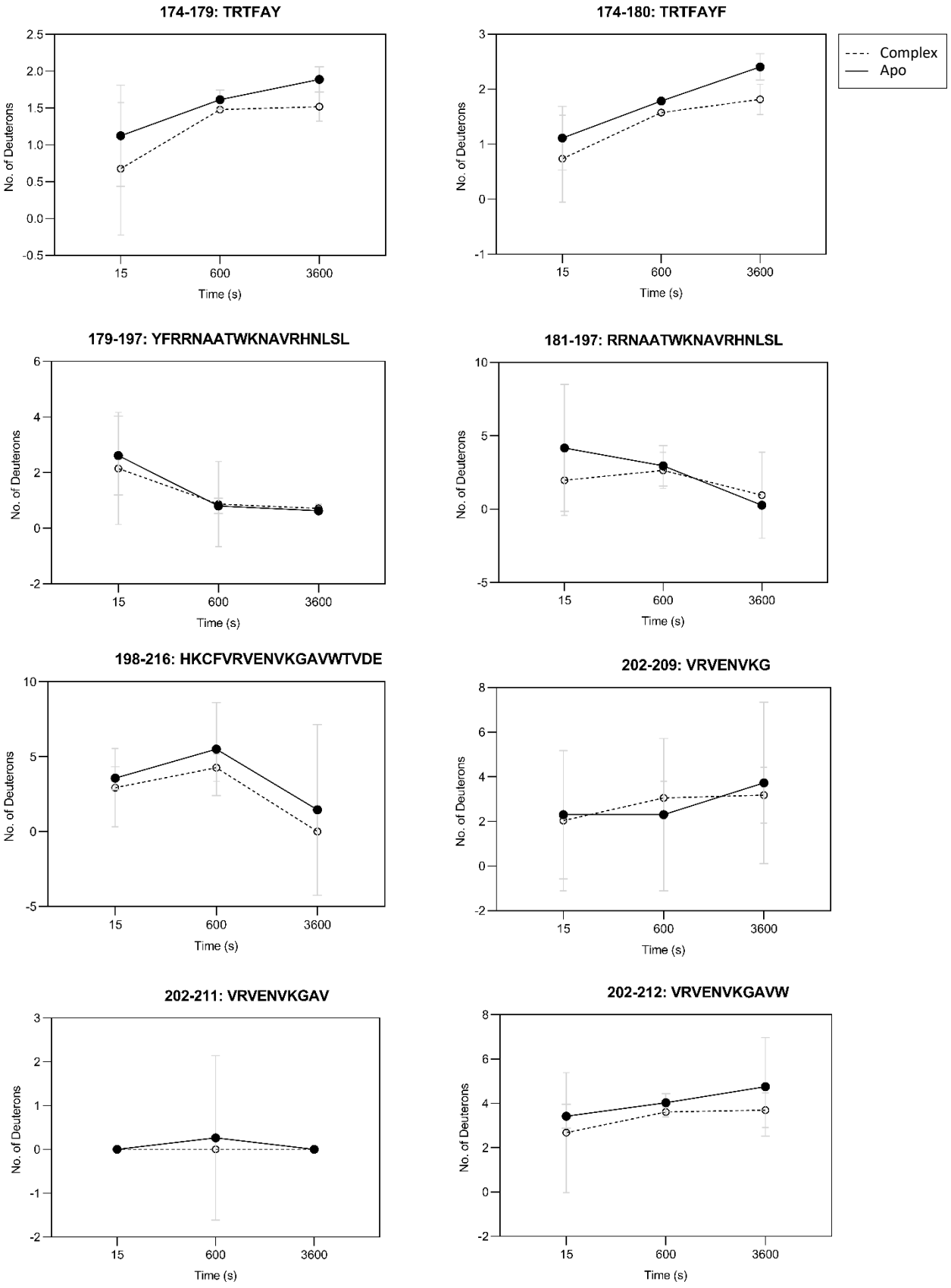

152-166: IRQAIMESSDRQLTL
No. of Deuterons
30
20
10
0
-10
15
600
3600
Time (s)

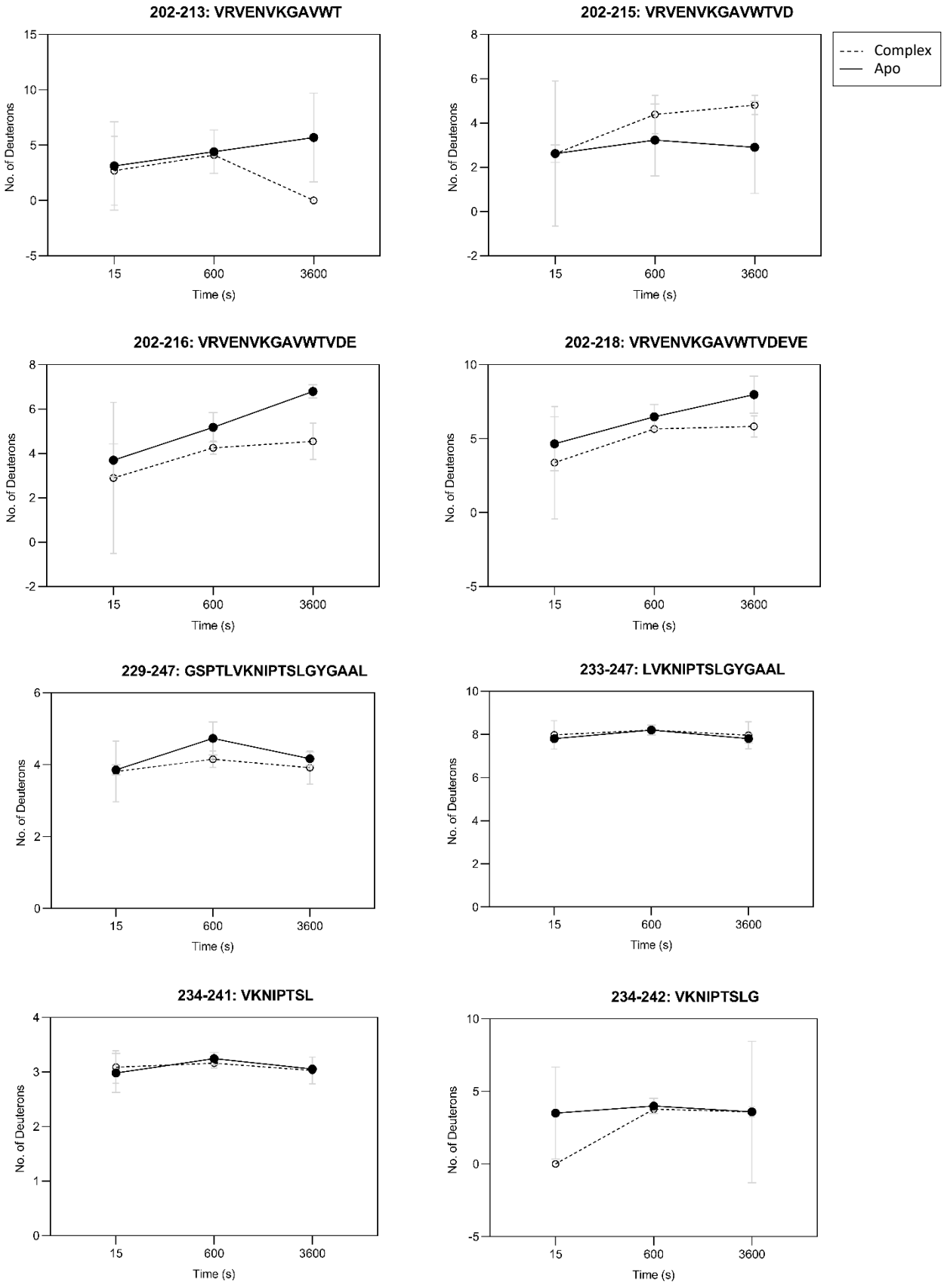

155-166: AIMESSDRQLTL
No. of Deuterons
6
4
2
0
-2
15
600
3600
Time (s)

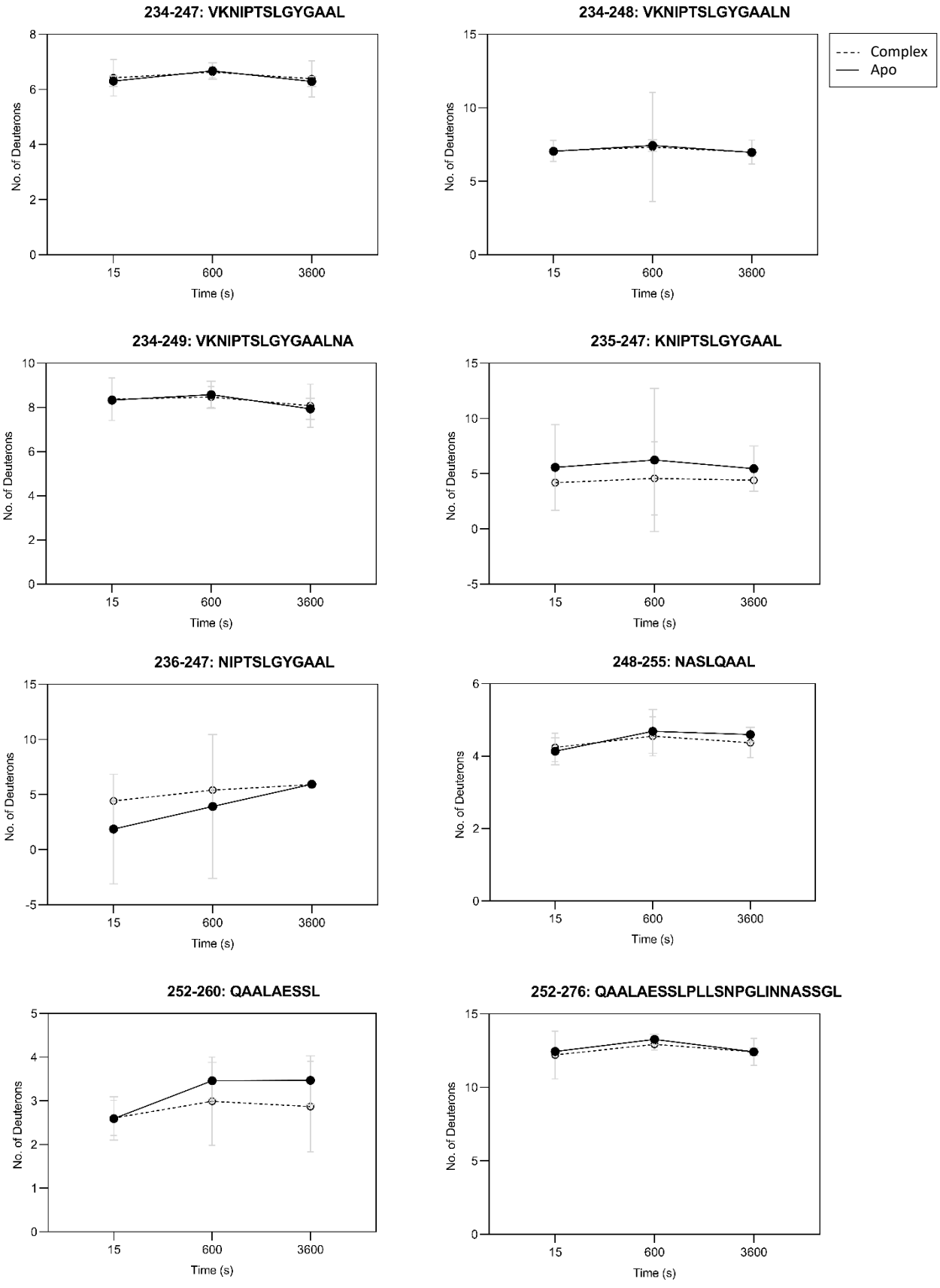

158-166: ESSDRQLTL
No. of Deuterons
4
3
2
1
0
-1
15
600
3600
Time (s)

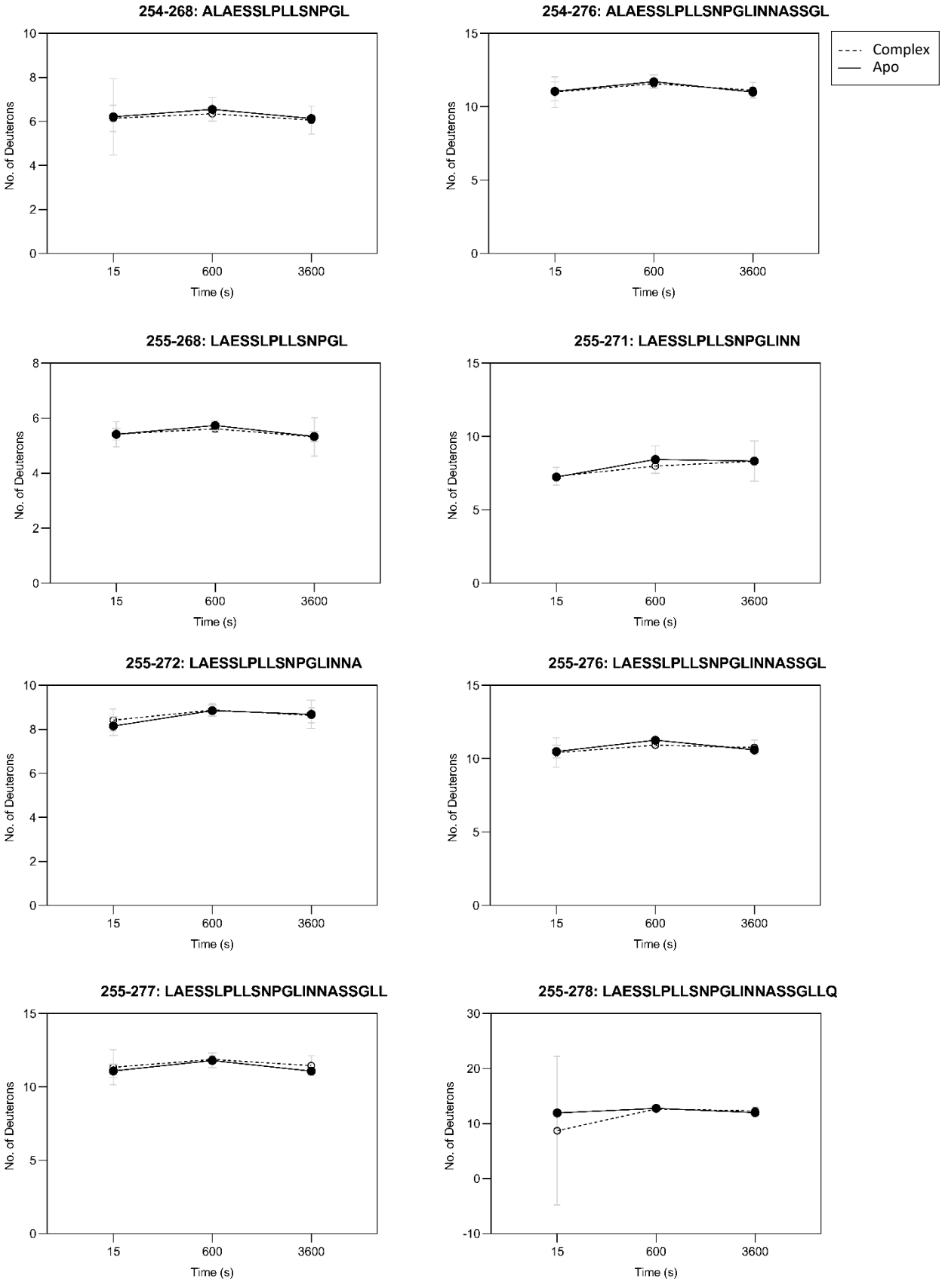

158-167: ESSDRQLTLN
No. of Deuterons
15
10
5
0
15
600
3600
Time (s)

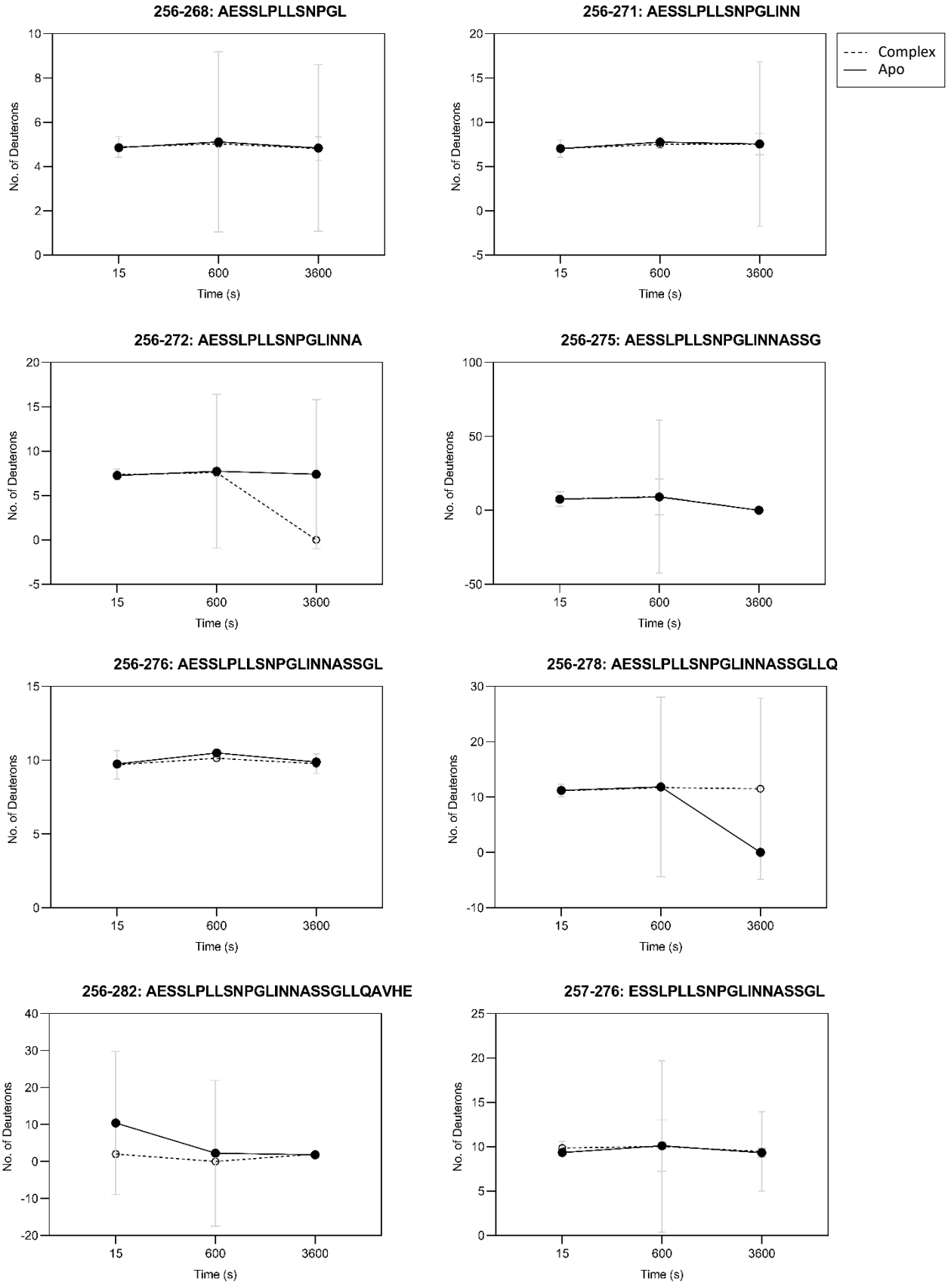

158-168: ESSDRQLTLNE
159-168: SSDRQLTLNE
Complex
Apo
161-168: DRQLTLNE
162-168: RQLTLNE
167-172: NEIYSW
173-178: FTRTFA
173-179: FTRTFAY
173-180: FTRTFAYF
No. of Deuterons
Time (s)
15
600
3600

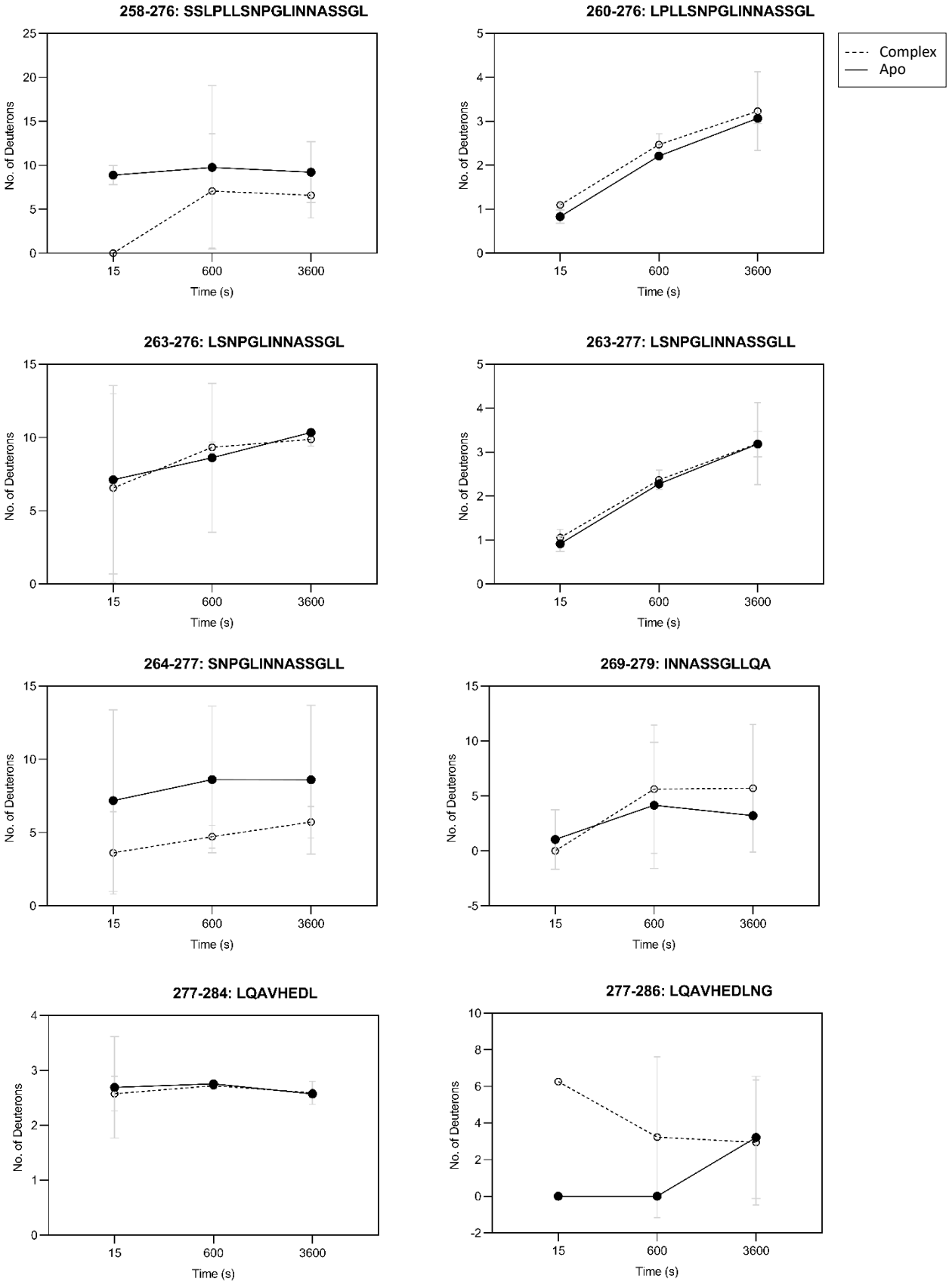

174-179: TRTFAY
174-180: TRTFAYF
Complex
Apo
179-197: YFRRNAATWKNAVRHNLSL
181-197: RRNAATWKNAVRHNLSL
198-216: HKCFVRVENVKGAVWTVDE
202-209: VRVENVKG
202-211: VRVENVKGAV
202-212: VRVENVKGAVW
No. of Deuterons
Time (s)
15
600
3600


202-213: VRVENVKGAVWT
202-215: VRVENVKGAVWTVD
202-216: VRVENVKGAVWTVDE
202-218: VRVENVKGAVWTVDEVE
229-247: GSPTLVKNIPTSLGYGAAL
233-247: LVKNIPTSLGYGAAL
234-241: VKNIPTSL
234-242: VKNIPTSLG
Complex
Apo
No. of Deuterons
Time (s)
15
600
3600

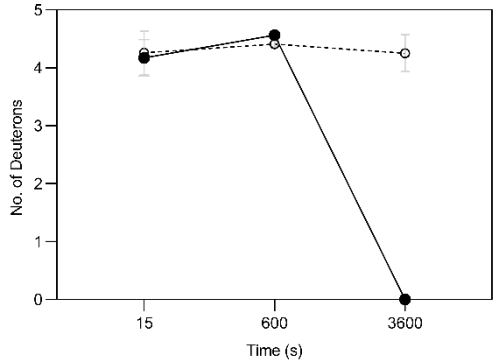

234-247: VKNIPTSLGYGAAL
234-248: VKNIPTSLGYGAALN
234-249: VKNIPTSLGYGAALNA
235-247: KNIPTSLGYGAAL
236-247: NIPTSLGYGAAL
248-255: NASLQAAL
252-260: QAALAESSL
252-276: QAALAESSLPLLSNPGLINNASSGL
Complex
Apo
No. of Deuterons
Time (s)
15
600
3600


254-268: ALAESSLPLLSNPGL
254-276: ALAESSLPLLSNPGLINNASSGL
Complex
Apo
255-268: LAESSLPLLSNPGL
255-271: LAESSLPLLSNPGLINN
255-272: LAESSLPLLSNPGLINNA
255-276: LAESSLPLLSNPGLINNASSGL
255-277: LAESSLPLLSNPGLINNASSGLL
255-278: LAESSLPLLSNPGLINNASSGLLQ
No. of Deuterons
Time (s)
15
600
3600

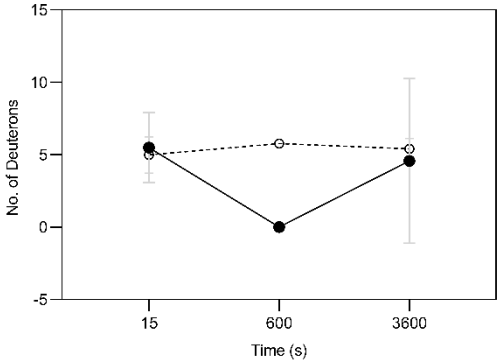

256-268: AESSLPLLSNPGL
256-271: AESSLPLLSNPGLINN
Complex
Apo
256-272: AESSLPLLSNPGLINNA
256-275: AESSLPLLSNPGLINNASSG
256-276: AESSLPLLSNPGLINNASSGL
256-278: AESSLPLLSNPGLINNASSGLLQ
256-282: AESSLPLLSNPGLINNASSGLLQAVHE
257-276: ESSLPLLSNPGLINNASSGL
No. of Deuterons
Time (s)
15
600
3600

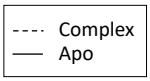

258-276: SSLPLLSNPGLINNASSGL
260-276: LPLLSNPGLINNASSGL
Complex
Apo
263-276: LSNPGLINNASSGL
263-277: LSNPGLINNASSGLL
264-277: SNPGLINNASSGLL
269-279: INNASSGLLQA
277-284: LQAVHEDL
277-286: LQAVHEDLNG
No. of Deuterons
Time (s)
15
600
3600

277-288: LQAVHEDLNGSL

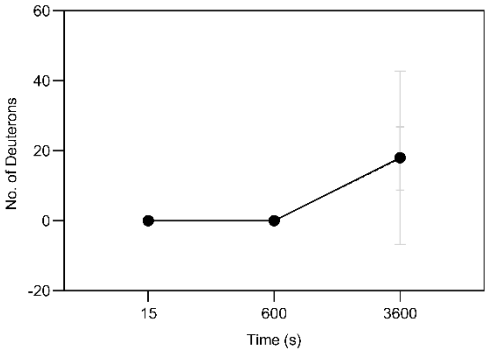


277-289: LQAVHEDLNGSLD

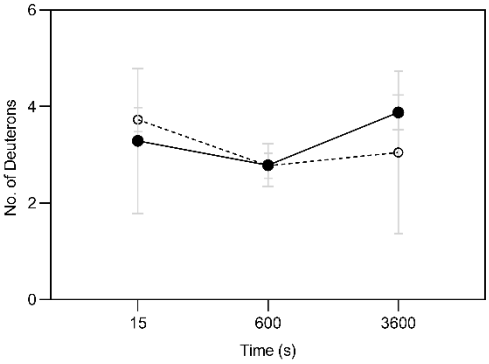




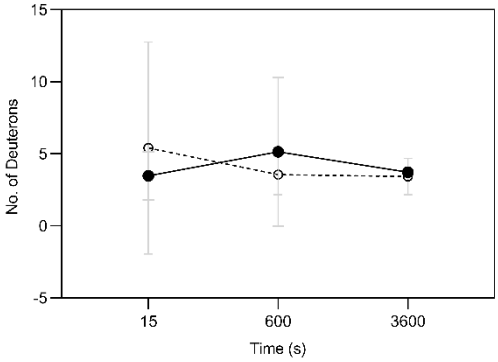
277-320: LQAVHEDLNGSLDHIDSNGNSSPGCSPQPHIHSIHVKEEPVIAE



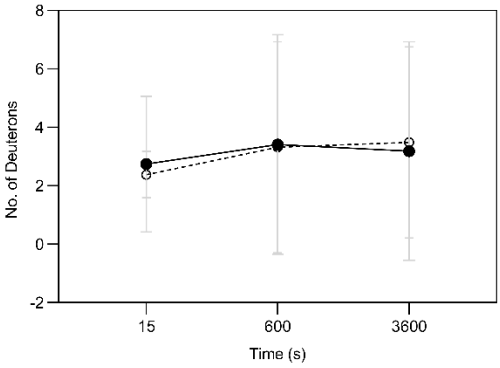
278-286: QAVHEDLNG



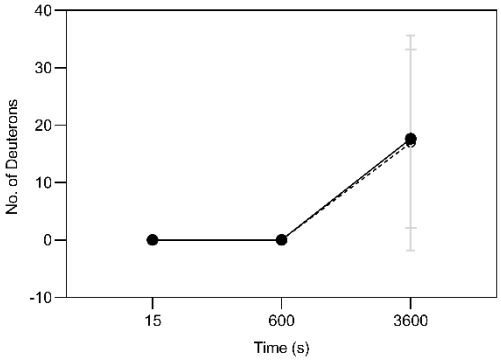
278-288: QAVHEDLNGSL



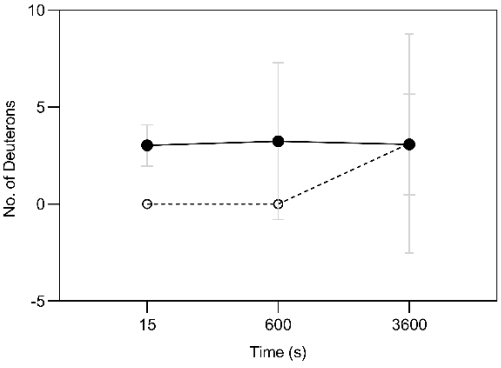
278-289: QAVHEDLNGSLD

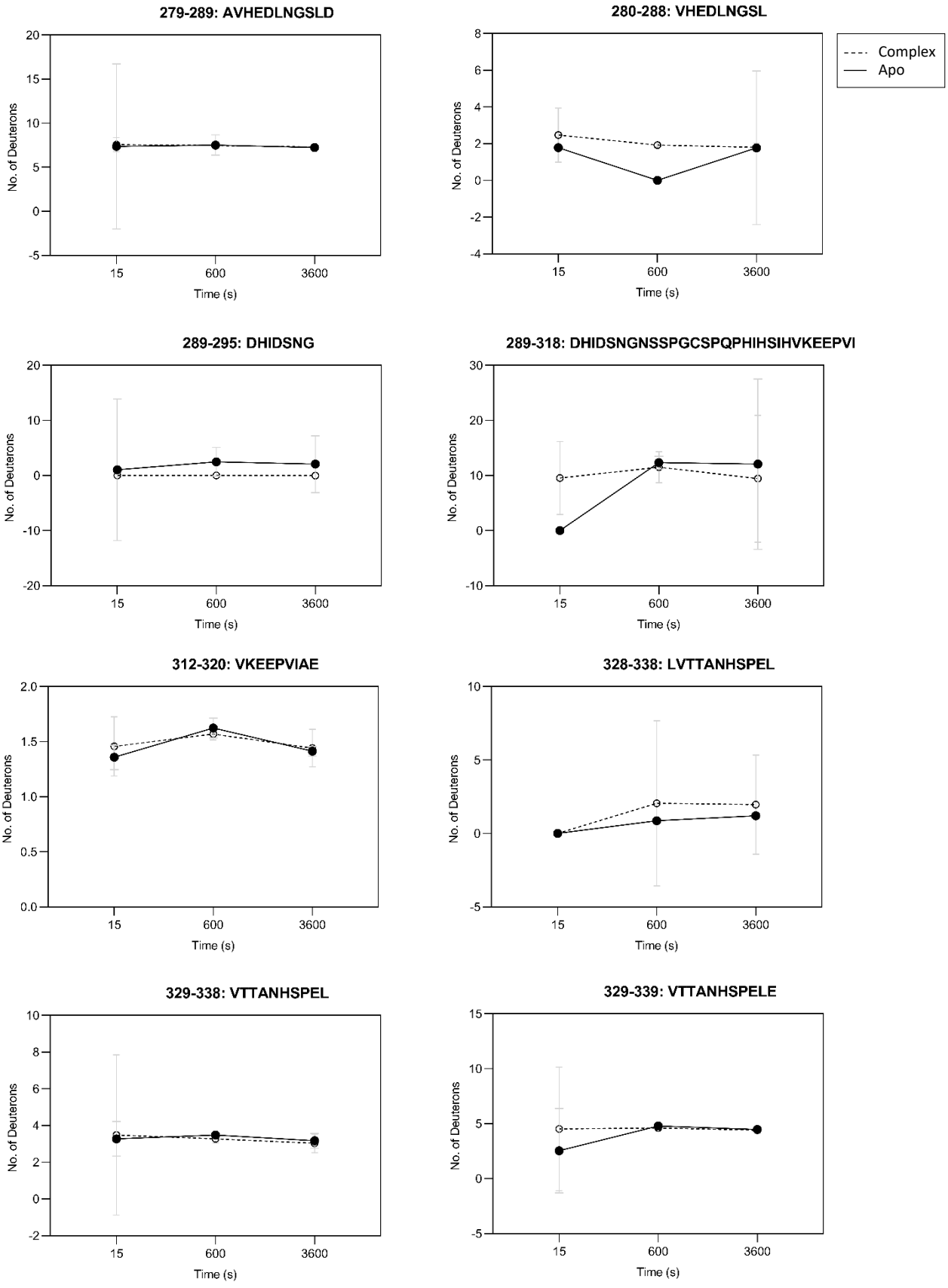


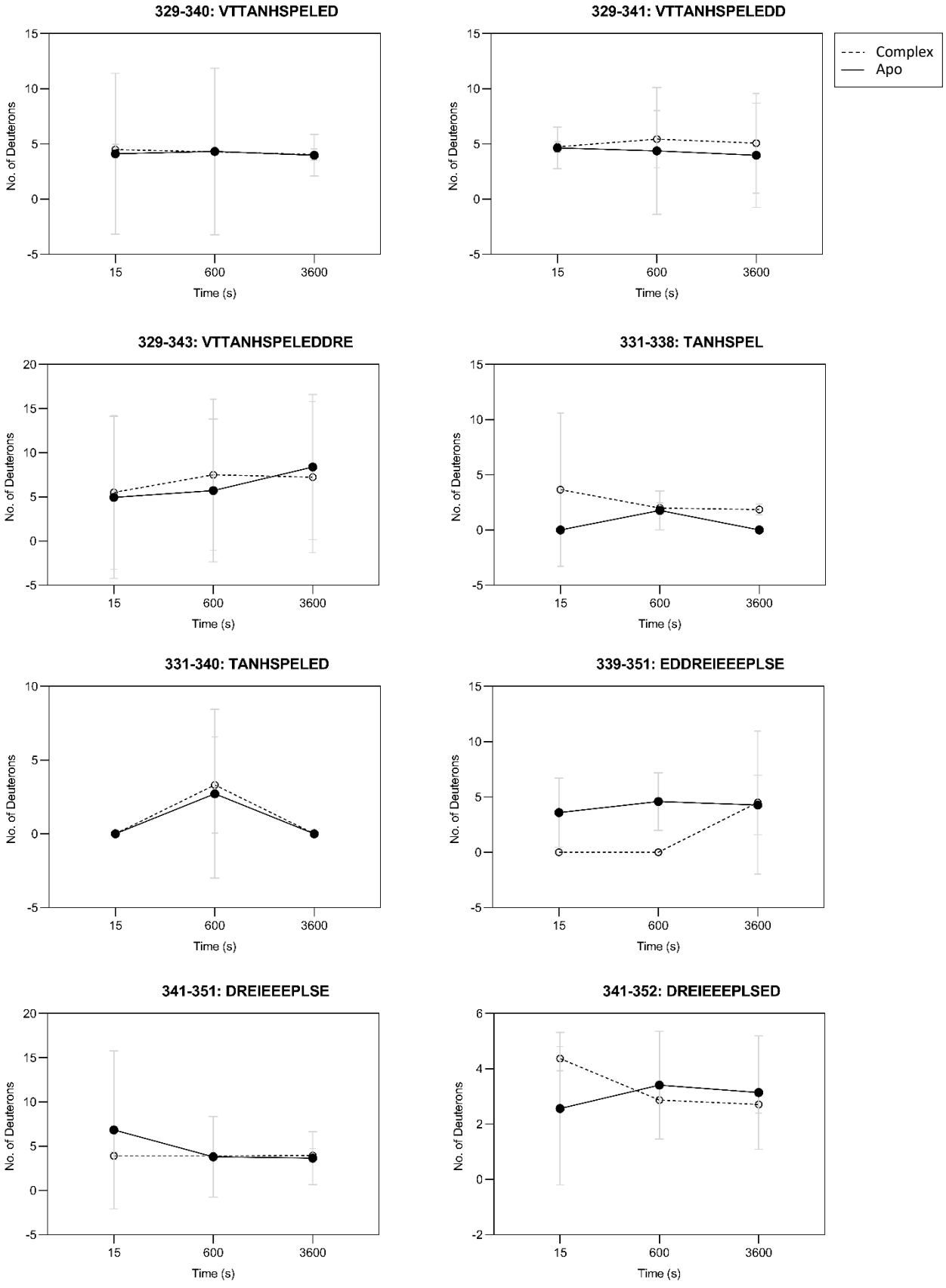
278-320: QAVHEDLNGSLDHIDSNGNSSPGCSPQPHIHSIHVKEEPVIAE

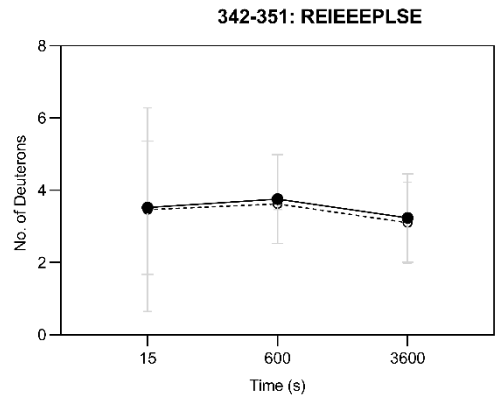


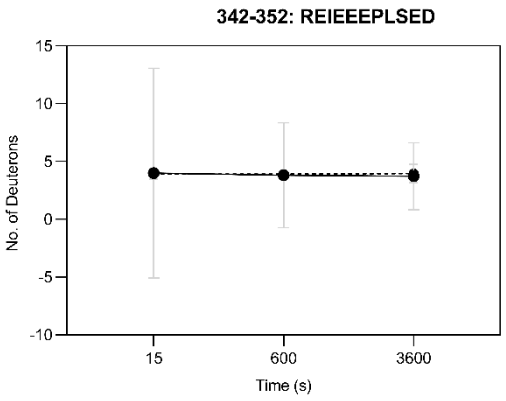
279-288: AVHEDLNGSL

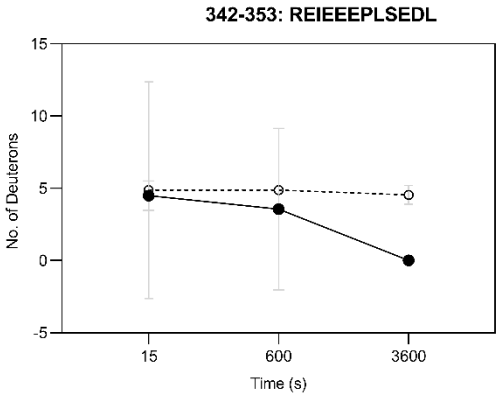


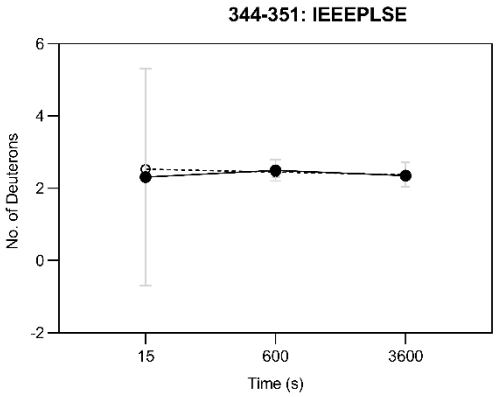

279-289: AVHEDLNGSLD
280-288: VHEDLNGSL
Complex
Apo
289-295: DHIDSNG
289-318: DHIDSNGNSSPGCSPQPHIHSIHVKEEPVI
312-320: VKEEPVIAE
328-338: LVTTANHSPEL
329-338: VTTANHSPEL
329-339: VTTANHSPELE
No. of Deuterons
Time (s)
15
600
3600

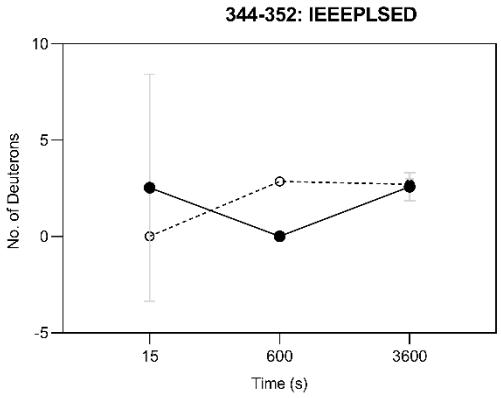

329-340: VTTANHSPELED
329-341: VTTANHSPELEDD
329-343: VTTANHSPELEDDRE
331-338: TANHSPEL
331-340: TANHSPELED
339-351: EDDREIEEEPLSE
341-351: DREIEEEPLSE
341-352: DREIEEEPLSED
No. of Deuterons
Time (s)
15
600
3600
Complex
Apo

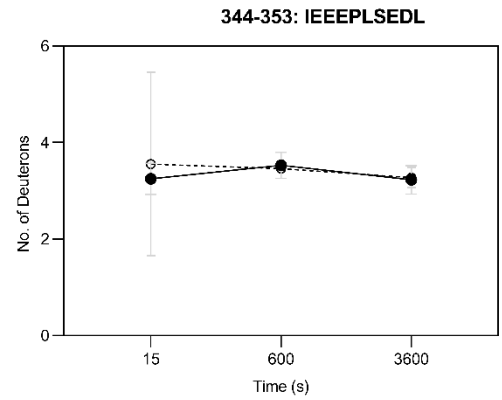

342-351: REIEEEPLSE
No. of Deuterons
8
6
4
2
0
15
600
3600
Time (s)

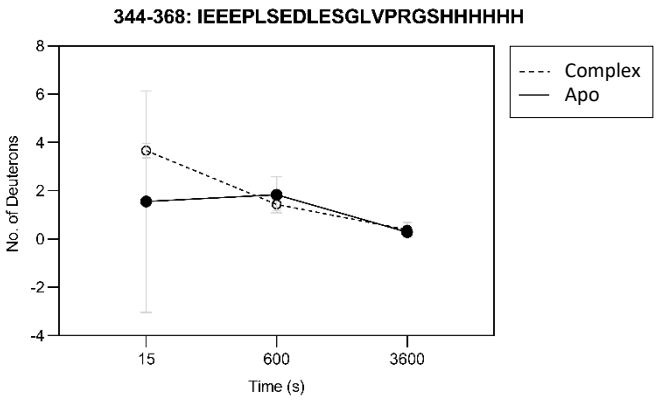

342-352: REIEEEPLSED
No. of Deuterons
15
10
5
0
-5
-10
15
600
3600
Time (s)
Complex
Apo

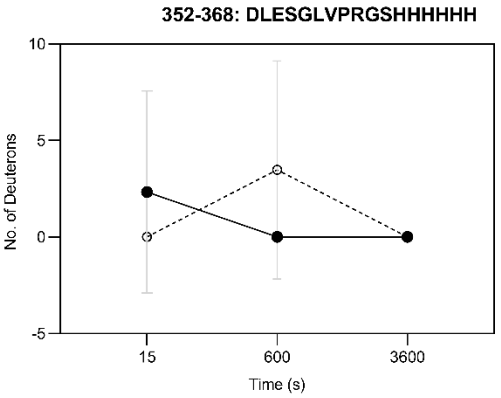

342-353: REIEEEPLSEDL
No. of Deuterons
15
10
5
0
-5
15
600
3600
Time (s)

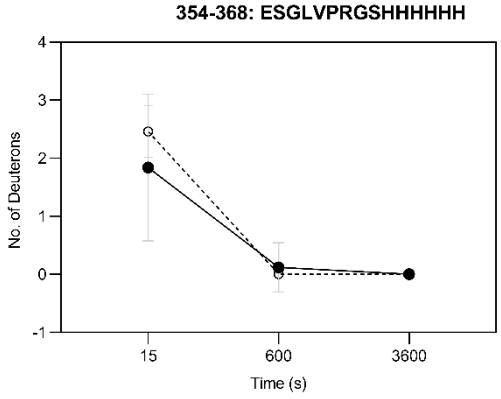

344-351: IEEEPLSE
No. of Deuterons
6
4
2
0
-2
15
600
3600
Time (s)

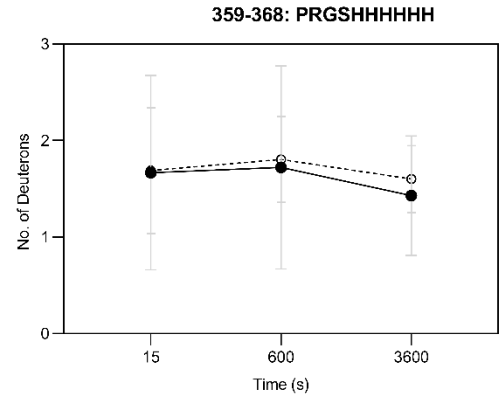

344-352: IEEEPLSED
No. of Deuterons
10
5
0
-5
15
600
3600
Time (s)


344-353: IEEEPLSEDL
No. of Deuterons
Time (s)
15
600
3600


344-368: IEEEPLSEDLESGLVPRGSHHHHHH
Complex
Apo
No. of Deuterons
Time (s)
15
600
3600


352-368: DLESGLVPRGSHHHHHH
No. of Deuterons
Time (s)
15
600
3600


354-368: ESGLVPRGSHHHHHH
No. of Deuterons
Time (s)
15
600
3600


359-368: PRGSHHHHHH
No. of Deuterons
Time (s)
15
600
3600

Akash, M.S.H. and Rehman, K., 2020. Ultraviolet-Visible (UV-VIS) Spectroscopy. In Essentials of Pharmaceutical Analysis (pp. 29-56). Springer, Singapore.

Alber, T., 1992. Structure of the leucine zipper. *Current opinion in genetics & development*, 2(2), pp.205-210.

Alberts, B., Bray, D., Hopkin, K., Johnson, A.D., Lewis, J., Raff, M., Roberts, K. and Walter, P., 2013. Essential Cell Biology. Garland Science.

Ali, M.H. and Imperiali, B., 2005. Protein oligomerization: how and why. *Bioorganic & Medicinal Chemistry*, *13*(17), pp.5013-5020.

Amoutzias, G.D., Robertson, D.L., Oliver, S.G. and Bornberg-Bauer, E., 2004. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Reports*, *5*(3), pp.274-279.

Amoutzias, G.D., Robertson, D.L., Van de Peer, Y. and Oliver, S.G., 2008. Choose your partners: dimerization in eukaryotic transcription factors. *Trends in Biochemical Sciences*, *33*(5), pp.220-229.

Anand, U., Jash, C. and Mukherjee, S., 2011. Protein unfolding and subsequent refolding: a spectroscopic investigation. *Physical Chemistry Chemical Physics*, 13(45), pp.20418-20426.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. *Science*, *181*(4096), pp.223-230.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M. and Iyer, L.M., 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiology Reviews*, *29*(2), pp.231-262.

Atchley, W.R. and Fitch, W.M., 1997. A natural classification of the basic helix–loop–helix class of transcription factors. *Proceedings of the National Academy of Sciences*, 94(10), pp.5172-5176.

Banham, A.H., Beasley, N., Campo, E., Fernandez, P.L., Fidler, C., Gatter, K., Jones, M., Mason, D.Y., Prime, J.E., Trougouboff, P. and Wood, K., 2001. The FOXP1 winged helix transcription factor is a novel candidate tumor suppressor gene on chromosome 3p. *Cancer Research*, *61*(24), pp.8820-8829.

Beamer, L.J. and Pabo, C.O., 1992. Refined 1.8 Å crystal structure of the λ repressor-operator complex. *Journal of molecular biology*, *227*(1), pp.177-196.

Belton, E., Salmond, C.H., Watkins, K.E., Vargha-Khadem, F. and Gadian, D.G., 2003. Bilateral brain abnormalities associated with dominantly inherited verbal and orofacial dyspraxia. *Human Brain Mapping*, *18*(3), pp.194-200.

Berova, N., Nakanishi, K. and Woody, R.W. eds., 2000. *Circular dichroism: principles and applications*. John Wiley & Sons.

Birnstiel, M.L., Busslinger, M. and Strub, K., 1985. Transcription termination and 3′ processing: the end is in site! *Cell*, *41*(2), pp.349-359.

Bornhorst, J.A. and Falke, J.J., 2000. [16] Purification of proteins using polyhistidine affinity tags. *Methods in Enzymology*, *326*, pp. 245-254.

Brennan, R.G. and Matthews, B.W., 1989. The helix-turn-helix DNA-binding motif. *Journal of Biological Chemistry*, *264*(4), pp.1903-1906.

Brown, R.S., 2005. Zinc finger proteins: getting a grip on RNA. *Current Opinion in Structural Biology*, *15*(1), pp.94-98.

Busch, S.J. and Sassone-Corsi, P., 1990. Dimers, leucine zippers and DNA-binding domains. *Trends in Genetics*, *6*, pp.36-40.

Caputo, S., Cuprite, J., Disband-Goulet, I., Coned, E., Lin, F., Brad, S., Gondry, M., Gilquin, B., Worman, H.J. and Zinn-Justin, S., 2006. The carboxyl-terminal nucleoplasmic region of MAN1 exhibits a DNA-binding winged helix domain. *Journal of Biological Chemistry*, *281*(26), pp.18208-18215.

Chae, W.J., Henegariu, O., Lee, S.K. and Bothwell, A.L., 2006. The mutant leucine-zipper domain impairs both dimerization and suppressive function of Foxp3 in T cells. *Proceedings of the National Academy of Sciences*, *103*(25), pp.9631-9636.

Chen, X., Vinkemeier, U., Zhao, Y., Jeruzalmi, D., Darnell Jr, J.E. and Kuriyan, J., 1998. Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell*, *93*(5), pp.827-839.

Chen, K. and Rajewsky, N., 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, *8*(2), pp.93.

Chen, X., Ji, Z., Webber, A. and Sharrocks, A.D., 2015. Genome-wide binding studies reveal DNA-binding specificity mechanisms and functional interplay amongst Forkhead transcription factors. *Nucleic acids research*, *44*(4), pp.1566-1578.

Cheow, L.F., Viswanathan, R., Chin, C.S., Jennifer, N., Jones, R.C., Guccione, E., Quake, S.R. and Burkholder, W.F., 2014. Multiplexed analysis of protein–ligand interactions by fluorescence anisotropy in a microfluidic platform. *Analytical Chemistry*, *86*(19), pp.9901-9908.

Cheung, M.S., Klimov, D. and Thirumalai, D., 2005. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proceedings of the National Academy of Sciences*, *102*(13), pp.4753-4758.

Chowdhury, S.K., Katta, V. and Chait, B.T., 1990. Probing conformational changes in proteins by mass spectrometry. *Journal of the American Chemical Society*, *112*(24), pp.9012-9013.

Christiansen, A., Wang, Q., Cheung, M.S. and Wittung-Stafshede, P., 2013. Effects of macromolecular crowding agents on protein folding in vitro and in silico. *Biophysical reviews*, *5*(2), pp.137-145.

Clark, K.L., Halay, E.D., Lai, E. and Burley, S.K., 1993. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, *364*(6436), p.412.

Coleman, R.A. and Pugh, B.F., 1995. Evidence for Functional Binding and Stable Sliding of the TATA Binding Protein on Nonspecific DNA∗. *Journal of Biological Chemistry*, *270*(23), pp.13850-13859.

Costa, R.H., Grayson, D.R. and Darnell, J.E., 1989. Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and alpha 1-antitrypsin genes. *Molecular and Cellular Biology, 9*(4), pp.1415-1425.

Cooper, G.M., Hausman, R.E. and Hausman, R.E., 2000. The cell: a molecular approach (Vol. 10). Washington, DC: ASM press.

Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R. and Kornberg, R.D., 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, *288*(5466), pp.640-649.

Crick, F., 1970. Central dogma of molecular biology. *Nature*, *227*(5258), pp.561-563.

Díaz-Villanueva, J.F., Díaz-Molina, R. and García-González, V., 2015. Protein folding and mechanisms of proteostasis. *International journal of molecular sciences*, *16*(8), pp.17193-17230.

Dayhoff, J.E., Shoemaker, B.A., Bryant, S.H. and Panchenko, A.R., 2010. Evolution of protein binding modes in homooligomers. *Journal of Molecular Biology*, *395*(4), pp.860-870.

de Folter, S., Immink, R.G., Kieffer, M., Pařenicová, L., Henz, S.R., Weigel, D., Busscher, M., Kooiker, M., Colombo, L., Kater, M.M. and Davies, B., 2005. Comprehensive interaction map of the Arabidopsis MADS box transcription factors. *The Plant Cell*, *17*(5), pp.1424-1433.

Desvergne, B., Michalik, L. and Wahli, W., 2006. Transcriptional regulation of metabolism. *Physiological Reviews*, *86*(2), pp.465-514.

Dill, K.A., 1990. Biochemistry https://doi. org/10.1021/bi00327a032 24, 1501 (1985); Google Scholar Cross ref, ISI. *KA Dill, Biochemistry https://doi. org/10.1021/bi00483a001*, *29*, p.7133.

Elrod-Erickson, M., Rould, M.A., Nekludova, L. and Pabo, C.O., 1996. Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, *4*(10), pp.1171-1180.

Elson, E.L. and Magde, D., 1974. Fluorescence correlation spectroscopy. I. Conceptual basis and theory. *Biopolymers: Original Research on Biomolecules*, *13*(1), pp.1-27.

Engen, J.R. and Smith, D.L., 2000. Investigating the higher order structure of proteins. In *Mass Spectrometry of Proteins and Peptides* (pp. 95-112). Humana Press, Totowa, NJ.

Engen, J.R. and Wales, T.E., 2015. Analytical aspects of hydrogen exchange mass spectrometry. *Annual review of analytical chemistry*, *8*, pp.127-148.

Englander, S.W., 2006. Hydrogen exchange and mass spectrometry: A historical perspective. *Journal of the American Society for Mass Spectrometry*, *17*(11), pp.1481-1489.

Enokizono, Y., Konishi, Y., Nagata, K., Ouhashi, K., Uesugi, S., Ishikawa, F. and Katahira, M., 2005. Structure of hnRNP D complexed with single-stranded telomere DNA and unfolding of the quadruplex by heterogeneous nuclear ribonucleoprotein D. *Journal of Biological Chemistry*, *280*(19), pp.18862-18870.

Erkmann, J.A. and Kutay, U., 2004. Nuclear export of mRNA: from the site of transcription to the cytoplasm. *Experimental Cell Research*, *296*(1), pp.12-20.

Failer, B.U., Aschrafi, A., Schmalzing, G. and Zimmermann, H., 2003. Determination of native oligomeric state and substrate specificity of rat NTPDase1 and NTPDase2 after heterologous expression in Xenopus oocytes. *European journal of biochemistry*, *270*(8), pp.1802-1809.

Favicchio, R., Dragan, A.I., Kneale, G.G. and Read, C.M., 2009. Fluorescence spectroscopy and anisotropy in the analysis of DNA-protein interactions. *DNA-Protein Interactions: Principles and Protocols, Third Edition*, pp.589-611.

Feuk, L., Kalervo, A., Lipsanen-Nyman, M., Skaug, J., Nakabayashi, K., Finucane, B., Hartung, D., Innes, M., Kerem, B., Nowaczyk, M.J. and Rivlin, J., 2006. Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. *The American Journal of Human Genetics*, *79*(5), pp.965-972.

Finnin, M.S., Cicero, M.P., Davies, C., Porter, S.J., White, S.W. and Kreuzer, K.N., 1997. The activation domain of the MotA transcription factor from bacteriophage T4. *The EMBO journal*, *16*(8), pp.1992-2003.

Firestein, G.S. and Manning, A.M., 1999. Signal transduction and transcription factors in rheumatic disease. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *42*(4), pp.609-621.

Fontenot, J.D., Rasmussen, J.P., Williams, L.M., Dooley, J.L., Farr, A.G. and Rudensky, A.Y., 2005. Regulatory T cell lineage specification by the forkhead transcription factor foxp3. *Immunity*, *22*(3), pp.329-341.

Förster, T., 1949. Experimentelle und theoretische Untersuchung des zwischenmolekularen Übergangs von Elektronenanregungsenergie. *Zeitschrift für naturforschung A*, *4*(5), pp.321- 327.

Fried, M.G., 1989. Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis*, *10*(5-6), pp.366-376.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M.R., Appel, R.D. and Bairoch, A., 2005. Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*, pp. 571-607.

Gajiwala, K.S. and Burley, S.K., 2000. Winged helix proteins. *Current Opinion in Structural Biology*, *10*(1), pp.110-116.

Ghisaidoobe, A.B. and Chung, S.J., 2014. Intrinsic tryptophan fluorescence in the detection and analysis of proteins: a focus on Förster resonance energy transfer techniques. *International journal of molecular sciences*, *15*(12), pp.22518-22538.

Greenfield, N.J., 2006. Using circular dichroism spectra to estimate protein secondary structure. Nature protocols, 1(6), pp.2876-2890.

Grunstein, M., 1997. Histone acetylation in chromatin structure and transcription. *Nature*, *389*(6649), pp.349.

Guttman, M., Weis, D.D., Engen, J.R. and Lee, K.K., 2013. Analysis of overlapped and noisy hydrogen/deuterium exchange mass spectra. *Journal of the American Society for Mass Spectrometry*, *24*(12), pp.1906-1912.

Ha, I.L.H.O., Roberts, S., Maldonado, E., Sun, X.I.A.O.Q.I.N.G., Kim, L.U., Green, M. and Reinberg, D., 1993. Multiple functional domains of human transcription factor IIB: distinct interactions with two general transcription factors and RNA polymerase II. *Genes & Development*, *7*(6), pp.1021-1032.

Hackett, B.P., 2002. Fox Transcription Factors. *Wiley Encyclopedia of Molecular Medicine*.

Hamilton, P.L. and Arya, D.P., 2012. Natural product DNA major groove binders. *Natural Product Reports*, *29*(2), pp.134-143.

Hannemann, F., Bera, A.K., Fischer, B., Lisurek, M., Teuchner, K. and Bernhardt, R., 2002. Unfolding and conformational studies on bovine adrenodoxin probed by engineered intrinsic tryptophan fluorescence. *Biochemistry*, *41*(36), pp.11008-11016.

Harrison, S.C., 1991. A structural taxonomy of DNA-binding domains. *Nature*, *353*(6346), pp.715.

Hashimoto, K. and Panchenko, A.R., 2010. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proceedings of the National Academy of Sciences*, *107*(47), pp.20352-20357.

Häußermann, K., Young, G., Kukura, P. and Dietz, H., 2019. Dissecting FOXP2 oligomerization and DNA-binding. *Angewandte Chemie*.

Hayat, S.M., Farahani, N., Golichenari, B. and Sahebkar, A., 2018. Recombinant Protein Expression in Escherichia coli (E. coli): What We Need to Know. *Current Pharmaceutical Design*, *24*(6), pp.718-725.

Hellman, L.M. and Fried, M.G., 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature protocols*, *2*(8), pp.1849-1861.

Heyduk, T., Ma, Y., Tang, H. and Ebright, R.H., 1996. Fluorescence anisotropy: rapid, quantitative assay for protein-DNA and protein-protein interaction. In *Methods in enzymology* (Vol. 274, pp. 492-503). Academic Press.

Hori, S., Nomura, T. and Sakaguchi, S., 2003. Control of regulatory T cell development by the transcription factor Foxp3. *Science*, *299*(5609), pp.1057-1061.

Houde, D., Berkowitz, S.A. and Engen, J.R., 2011. The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *Journal of Pharmaceutical Sciences*, *100*(6), pp.2071-2086.

Holden, N.S. and Tacon, C.E., 2011. Principles and problems of the electrophoretic mobility shift assay. *Journal of pharmacological and toxicological methods*, *63*(1), pp.7-14.

Hu, H., Wang, B., Borde, M., Nardone, J., Maika, S., Allred, L., Tucker, P.W. and Rao, A., 2006. Foxp1 is an essential transcriptional regulator of B cell development. *Nature immunology*, *7*(8), pp.819-826.

Jackson, B.C., Carpenter, C., Nebert, D.W. and Vasiliou, V., 2010. Update of human and mouse forkhead box (FOX) gene families. *Human Genomics*, *4*(5), pp.345.

Jaenicke, R., 1993. What does protein refolding in vitro tell us about protein folding in the cell? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *339*(1289), pp.287-295.

Jameson, D.M., 2014. *Introduction to fluorescence*. Taylor & Francis.

Jin, C., Marsden, I., Chen, X. and Liao, X., 1999. Dynamic DNA contacts observed in the NMR structure of winged helix protein-DNA complex. *Journal of Molecular Biology*, *289*(4), pp.683-690.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589.

Kaestner, K.H., Knöchel, W. and Martínez, D.E., 2000. Unified nomenclature for the winged helix/forkhead transcription factors. *Genes & Development*, *14*(2), pp.142-146.

Katoh, M. and Katoh, M., 2004. Human FOX gene family. *International Journal of Oncology*, *25*(5), pp.1495-1500.

Kawski, A., 1993. Fluorescence anisotropy: theory and applications of rotational depolarization. *Critical Reviews in Analytical Chemistry*, *23*(6), pp.459-529.

Keller, W., König, P. and Richmond, T.J., 1995. Crystal structure of a bZIP/DNA complex at 2.2 Å: determinants of DNA specific recognition. *Journal of Molecular Biology*, *254*(4), pp.657-667.

Kelly, S.M., Jess, T.J. and Price, N.C., 2005. How to study proteins by circular dichroism. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1751*(2), pp.119-139.

Kohler, J.J. and Schepartz, A., 2001. Kinetic studies of fos jun DNA complex formation: DNA-binding prior to dimerization. *Biochemistry*, *40*(1), pp.130-142.

Kohler, J.J. and Schepartz, A., 2001. Effects of nucleic acids and polyanions on dimer formation and DNA binding by bZIP and bHLHZip transcription factors. *Bioorganic & medicinal chemistry*, *9*(9), pp.2435-2443.

Kornberg, R.D., 2007. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences*, *104*(32), pp.12955-12961.

Krichevsky, O. and Bonnet, G., 2002. Fluorescence correlation spectroscopy: the technique and its applications. *Reports on Progress in Physics*, *65*(2), pp.251.

Krishna, R.G. and Wold, F., 1993. Post-translational modifications of proteins. *Methods in Protein Sequence Analysis,* pp. 167-172.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. and Monaco, A.P., 2001. A forkhead- domain gene is mutated in a severe speech and language disorder. *Nature*, *413*(6855), pp.519.

Lai, C.S., Gerrelli, D., Monaco, A.P., Fisher, S.E. and Copp, A.J., 2003. FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. *Brain*, *126*(11), pp.2455-2462.

Lakowicz, J.R., 1999. Fluorescence anisotropy. In *Principles of fluorescence spectroscopy* (pp. 291-319). Springer, Boston, MA.

Lakowicz, J.R. ed., 2013. *Principles of fluorescence spectroscopy*. Springer science & business media.

Lam, E.W.F., Brosens, J.J., Gomes, A.R. and Koo, C.Y., 2013. Forkhead box proteins: tuning forks for transcriptional harmony. *Nature Reviews Cancer*, *13*(7), pp.482.

Latchman, D.S., 1990. Eukaryotic transcription factors. *Biochemical Journal*, *270*(2), pp.281.

Latchman, D.S., 1993. Transcription factors: an overview. *International Journal of Experimental Pathology*, *74*(5), pp.417.

Lee, H.H. and Frasch, M., 2004. Survey of forkhead domain encoding genes in the Drosophila genome: classification and embryonic expression patterns. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, *229*(2), pp.357-366.

Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M. and Aebersold, R., 2010. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular & Cellular Proteomics*, *9*(8), pp.1634-1649.

Leng, F., Zhang, W., Ramirez, R.N., Leon, J., Zhong, Y., Hou, L., Yuki, K., van der Veeken, J., Rudensky, A.Y., Benoist, C. and Hur, S., 2022. The transcription factor FoxP3 can fold into two dimerization states with divergent implications for regulatory T cell function and immune homeostasis. *Immunity*, *55*(8), pp.1354-1369.

Lerner, E., Cordes, T., Ingargiola, A., Alhadid, Y., Chung, S., Michalet, X. and Weiss, S., 2018. Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science*, *359*(6373), pp.1133.

Levy, D.E. and Darnell Jr, J.E., 2002. Signalling: Stats: transcriptional control and biological impact. *Nature Reviews Molecular Cell Biology*, *3*(9), pp.651.

Levy, E.D., Erba, E.B., Robinson, C.V. and Teichmann, S.A., 2008. Assembly reflects evolution of protein complexes. *Nature*, *453*(7199), pp.1262.

Lichtman, J.W. and Conchello, J.A., 2005. Fluorescence microscopy. *Nature methods*, *2*(12), pp.910-919.

Levinthal, C., 1969. How to fold graciously. *Mossbauer spectroscopy in biological systems*, *67*, pp.22-24.

Lewin, B. and Dover, G., 1994. *Genes v* (Vol. 1110). Oxford: Oxford University Press.

Li, S., Weidenfeld, J. and Morrisey, E.E., 2004. Transcriptional and DNA-binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. *Molecular and Cellular Biology*, *24*(2), pp.809-822.

LiCata, V.J. and Wowor, A.J., 2008. Applications of fluorescence anisotropy to the study of protein–DNA interactions. *Methods in cell biology*, *84*, pp.243-262.

Luscombe, N.M., Laskowski, R.A. and Thornton, J.M., 2001. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Research*, *29*(13), pp.2860-2874.

Lu, M.M., Li, S., Yang, H. and Morrisey, E.E., 2002. Foxp4: a novel member of the Foxp subfamily of winged-helix genes co-expressed with Foxp1 and Foxp2 in pulmonary and gut tissues. *Mechanisms of Development*, *2*, pp.197–202.

Ma, P.C., Rould, M.A., Weintraub, H. and Pabo, C.O., 1994. Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*, *77*(3), pp.451-459.

Maiolica, A., Cittaro, D., Borsotti, D., Sennels, L., Ciferri, C., Tarricone, C., Musacchio, A. and Rappsilber, J., 2007. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Molecular & Cellular Proteomics*, *6*(12), pp.2200-2211.

Marklund, E.G., Mahmutovic, A., Berg, O.G., Hammar, P., van der Spoel, D., Fange, D. and Elf, J., 2013. Transcription-factor binding and sliding on DNA studied using micro-and macroscopic models. *Proceedings of the National Academy of Sciences*, *110*(49), pp.19796-19801.

Mayr, F., Schütz, A., Döge, N. and Heinemann, U., 2012. The Lin28 cold-shock domain remodels pre-let-7 microRNA. *Nucleic acids research*, *40*(15), pp.7492-7506.

McKee, T. and McKee, J.R., 2003. *Biochemistry: the molecular basis of life* (Vol. 3). New York: McGraw-Hill.

Miller, Y., Ma, B. and Nussinov, R., 2010. Zinc ions promote Alzheimer Aβ aggregation via population shift of polymorphic states. *Proceedings of the National Academy of Sciences*, *107*(21), pp.9490-9495.

Minton, A.P., 2006. Macromolecular crowding. *Current Biology*, *16*(8), pp.R269-R271.

Minton, A.P., 2000. Implications of macromolecular crowding for protein assembly. *Current opinion in structural biology*, *10*(1), pp.34-39.

Minton, A.P., 2006. Macromolecular crowding. *Current Biology*, *16*(8), pp.R269-R271.

Mitchell, P.J. and Tjian, R., 1989. Transcriptional regulation in mammalian cells by sequence- specific DNA-binding proteins. *Science*, *245*(4916), pp.371-378.

Moerner, W.E. and Fromm, D.P., 2003. Methods of single-molecule fluorescence spectroscopy and microscopy. *Review of Scientific Instruments*, *74*(8), pp.3597-3619.

Morris, G. and Fanucchi, S., 2016. A key evolutionary mutation enhances DNA-binding of the FOXP2 forkhead domain. *Biochemistry*, *55*(13), pp.1959-1967.

Morris, G., Stoychev, S., Naicker, P., Dirr, H.W. and Fanucchi, S., 2018. The forkhead domain hinge-loop plays a pivotal role in DNA-binding and transcriptional activity of FOXP2. *Biological Chemistry*, *399*(8), pp.881-893.

Muers, M., 2012. Transcription factors slide to find binding sites in vivo. *Nature Reviews Genetics*, *13*(8), pp.518-518.

Nelson, C.S., Fuller, C.K., Fordyce, P.M., Greninger, A.L., Li, H. and DeRisi, J.L., 2013. Microfluidic affinity and ChIP-seq analyses converge on a conserved FOXP2-binding motif in chimp and human, which enables the detection of evolutionarily novel targets. *Nucleic Acids Research*, *41*(12), pp.5991-6004.

Nick Pace, C., Alston, R.W. and Shaw, K.L., 2000. Charge–charge interactions influence the denatured state ensemble and contribute to protein stability. *Protein Science*, *9*(7), pp.1395-1398.

Nussinov, R., Xu, D. and Tsai, C.J., 1998. Mechanism and evolution of protein dimerization. *Protein Science*, *7*(3), pp.533-544.

Oakley, R.H. and Cidlowski, J.A., 2013. The biology of the glucocorticoid receptor: new signaling mechanisms in health and disease. *Journal of Allergy and Clinical Immunology*, *132*(5), pp.1033-1044.

Octobre, G., Lemercier, C., Khochbin, S., Robert-Nicoud, M. and Souchier, C., 2005. Monitoring the interaction between DNA and a transcription factor (MEF2A) using fluorescence correlation spectroscopy. *Comptes Rendus Biologies*, *328*(12), pp.1033-1040.

Olefsky, J.M., 2001. Nuclear receptor minireview series-Introduction. *Journal of Biological Chemistry.*

O'Reilly, F.J. and Rappsilber, J., 2018. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nature structural & molecular biology*, *25*(11), pp.1000-1008.

Ouellet, J., Schorr, S., Iqbal, A., Wilson, T.J. and Lilley, D.M., 2011. Orientation of cyanine fluorophores terminally attached to DNA via long, flexible tethers. *Biophysical journal*, *101*(5), pp.1148-1154.

Pabo, C.O. and Sauer, R.T., 1992. Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, *61*(1), pp.1053-1095.

Pace, C.N., 1986. Determination and analysis of urea and guanidine hydrochloride denaturation curves in Methods in Enzymology (Hirs, CHW and Timasheff, SN, eds.) Vol. 131.

Pace, C.N. and Scholtz, J.M., 1997. Measuring the conformational stability of a protein. *Protein structure: A practical approach*, *2*, pp.299-321.

Pelton, J.T. and McLean, L.R., 2000. Spectroscopic methods for analysis of protein secondary structure. *Analytical biochemistry*, *277*(2), pp.167-176.

Pereira-Leal, J.B., Levy, E.D., Kamp, C. and Teichmann, S.A., 2007. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biology*, *8*(4), pp. R51.

Pereira-Leal, J.B. and Teichmann, S.A., 2005. Novel specificities emerge by stepwise duplication of functional modules. *Genome Research*, *15*(4), pp.552-559.

Persson, F., Lindén, M., Unoson, C. and Elf, J., 2013. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nature Methods*, *10*(3), pp.265.

Peters, K. and Richards, F.M., 1977. Chemical cross-linking: reagents and problems in studies of membrane structure. *Annual review of biochemistry*, *46*(1), pp.523-551.

Porath, J. and Olin, B., 1983. Immobilized metal affinity adsorption and immobilized metal affinity chromatography of biomaterials. Serum protein affinities for gel-immobilized iron and nickel ions. *Biochemistry*, *22*(7), pp.1621-1630.

Rand, K.D., Zehl, M. and Jørgensen, T.J., 2014. Measuring the hydrogen/deuterium exchange of proteins at high spatial resolution by mass spectrometry: overcoming gas-phase hydrogen/deuterium scrambling. *Accounts of chemical research*, *47*(10), pp.3018-3027.

Reiter, F., Wienerroither, S. and Stark, A., 2017. Combinatorial function of transcription factors and cofactors. *Current opinion in genetics & development*, *43*, pp.73-81.

Rogers, D.M., Jasim, S.B., Dyer, N.T., Auvray, F., Réfrégiers, M. and Hirst, J.D., 2019. Electronic circular dichroism spectroscopy of proteins. *Chem*, *5*(11), pp.2751-2774.

Rodger, A. and Nordén, B., 1997. *Circular dichroism and linear dichroism* (Vol. 1). Oxford University Press, USA.

Rosano, G.L. and Ceccarelli, E.A., 2014. Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology*, *5*, pp.172.

Roy, R., Hohng, S. and Ha, T., 2008. A practical guide to single-molecule FRET. *Nature Methods*, *5*(6), pp.507.

Sali, A., Shakhnovich, E. and Karplus, M., 1994. How does a protein fold? *nature*, *369*(6477), pp.248-251.

Sanders, M. and Bowman, J., 2019. Genetic Analysis: An Integrated Approach, 3rd ed. *Pearson*, pp.276-304.

Selkoe, D.J., 2003. Folding proteins in fatal ways. *Nature*, *426*(6968), pp.900-904.

Schön, C., Wochnik, A., Röβner, A., Donow, C. and Knöchel, W., 2006. The FoxP subclass in Xenopus laevis development. *Development Genes and Evolution*, *216*(10), pp.641-646.

Schumacher, M.A., Goodman, R.H. and Brennan, R.G., 2000. The structure of a CREB bZIP· somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *Journal of Biological Chemistry*, *275*(45), pp.35242-35247.

Sevenich, F.W., Langowski, J., Rippe, K. and Weiss, V., 1998. DNA-binding and oligomerization of NtrC studied by fluorescence anisotropy and fluorescence correlation spectroscopy. *Nucleic Acids Research*, *26*(6), pp.1373-1381.

Shirley, B.A. ed., 1995. *Protein stability and folding: Theory and practice* (Vol. 21). New York:: Humana Press.

Shu, W., Yang, H., Zhang, L., Lu, M.M. and Morrisey, E.E., 2001. Characterization of a new subfamily of winged helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. *Journal of Biological Chemistry*, *276*(29), pp.27488-27497.

Sin, C., Li, H. and Crawford, D.A., 2015. Transcriptional regulation by FOXP1, FOXP2, and FOXP4 dimerization. *Journal of Molecular Neuroscience*, *55*(2), pp.437-448.

Sinz, A., 2006. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass spectrometry reviews*, *25*(4), pp.663-682.

Song, X., Li, B., Xiao, Y., Chen, C., Wang, Q., Liu, Y., Berezov, A., Xu, C., Gao, Y., Li, Z. and Wu, S.L., 2012. Structural and biological features of FOXP3 dimerization relevant to regulatory T cell function. *Cell Reports*, *1*(6), pp.665-675.

Spitz, F. and Furlong, E.E., 2012. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, *13*(9), pp.613-626.

Stroud, J.C., Wu, Y., Bates, D.L., Han, A., Nowick, K., Paabo, S., Tong, H. and Chen, L., 2006. Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, *14*(1), pp.159-166.

Suter, D.M., 2020. Transcription factors and DNA play hide and seek. *Trends in cell biology*, *30*(6), pp.491-500.

Swift, J. and Coruzzi, G.M., 2017. A matter of time—how transient transcription factor interactions create dynamic gene regulatory networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1860*(1), pp.75-83.

Tarazona, N.A., Maestro, B., Revelles, O., Sanz, J.M. and Prieto, M.A., 2019. Role of leucine zipper-like motifs in the oligomerization of Pseudomonas putida phasins. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1863*(2), pp.362-370.

Thulo, M., Rabie, M.A., Pahad, N., Donald, H.L., Blane, A.A., Perumal, C.M., Penedo, J.C. and Fanucchi, S., 2021. The influence of various regions of the FOXP2 sequence on its structure and DNA-binding function. *Bioscience Reports*, *41*(1).

Tian, Y., Martinez, M.M. and Pappas, D., 2011. Fluorescence correlation spectroscopy: a review of biochemical and microfluidic applications. *Applied Spectroscopy*, *65*(4), pp.115A- 124A.

Topal, M.D. and Fresco, J.R., 1976. Complementary base pairing and the origin of substitution mutations. *Nature*, *263*(5575), pp.285.

Tropsha, A., Bowen, J.P., Brown, F.K. and Kizer, J.S., 1991. Do interhelical side chain-backbone hydrogen bonds participate in formation of leucine zipper coiled coils? *Proceedings of the National Academy of Sciences*, *88*(21), pp.9488-9492.

Tsai, K.L., Sun, Y.J., Huang, C.Y., Yang, J.Y., Hung, M.C. and Hsiao, C.D., 2007. Crystal structure of the human FOXO3a-DBD/DNA complex suggests the effects of post-translational modification. *Nucleic Acids Research*, *35*(20), pp.6984-6994.

Turjanski, A.G., Gutkind, J.S., Best, R.B. and Hummer, G., 2008. Binding-induced folding of a natively unstructured transcription factor. *PLoS Comput Biol*, *4*(4), p.e1000060.

Uhlenbeck, G.E. and Ornstein, L.S., 1930. On the theory of the Brownian motion. *Physical review*, *36*(5), p.823.

Valenti, L.E., De Pauli, C.P. and Giacomelli, C.E., 2006. The binding of Ni (II) ions to hexahistidine as a model system of the interaction between nickel and His-tagged proteins. *Journal of inorganic biochemistry*, *100*(2), pp.192-200.

Valeur, B., 2003. Molecular fluorescence. *Digital Encyclopedia of Applied Physics*, pp.477-531.

van Erp, P.B., Patterson, A., Kant, R., Berry, L., Golden, S.M., Forsman, B.L., Carter, J., Jackson, R.N., Bothner, B. and Wiedenheft, B., 2017. Conformational dynamics of DNA-binding and Cas3 recruitment by the CRISPR RNA-guided Cascade complex. *ACS chemical biology*, *13*(2), pp.481-490.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. and Luscombe, N.M., 2009. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics, 10*(4), pp.252.

Vernes, S.C. and Fisher, S.E., 2013. Genetic pathways implicated in speech and language. In *Animal Models of Speech and Language Disorders,* pp. 13-40.

Vinson, C., Acharya, A. and Taparowsky, E.J., 2006. Deciphering B-ZIP transcription factor interactions in vitro and in vivo. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, *1759*(1-2), pp.4-12.

van Zon, J.S., Morelli, M.J., Tănase-Nicola, S. and ten Wolde, P.R., 2006. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysical journal*, *91*(12), pp.4350-4367.

Wang, B., Lin, D., Li, C. and Tucker, P., 2003. Multiple domains define the expression and regulatory properties of Foxp1 forkhead transcriptional repressors. *Journal of Biological Chemistry*, *278*(27), pp.24259-24268.

Warren, A.J., 2002. Eukaryotic transcription factors. *Current Opinion in Structural Biology*, *12*(1), pp.107-114.

Watson, J.D. and Crick, F.H., 1953. Molecular structure of nucleic acids. *Nature*, *171*(4356), pp.737-738.

Webb, H., Steeb, O., Blane, A., Rotherham, L., Aron, S., Machanick, P., Dirr, H. and Fanucchi, S., 2017. The FOXP2 forkhead domain binds to a variety of DNA sequences with different rates and affinities. *The Journal of Biochemistry*, *162*(1), pp.45-54.

Weigel, D., Jürgens, G., Küttner, F., Seifert, E. and Jäckle, H., 1989. The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the Drosophila embryo. *Cell*, *57*(4), pp.645-658.

Weiss, S., 1999. Fluorescence spectroscopy of single biomolecules. *Science*, *283*(5408), pp.1676-1683.

West, A.E., Griffith, E.C. and Greenberg, M.E., 2002. Regulation of transcription factors by neuronal activity. *Nature Reviews Neuroscience*, *3*(12), pp.921.

West, A.G. and Sharrocks, A.D., 1999. MADS-box transcription factors adopt alternative mechanisms for bending DNA. *Journal of Molecular Biology*, *286*(5), pp.1311-1323.

Wiehagen, K.R., Corbo-Rodgers, E., Li, S., Staub, E.S., Hunter, C.A., Morrisey, E.E. and Maltzman, J.S., 2012. Foxp4 is dispensable for T cell development but required for robust recall responses. *PLoS One*, *7*(8), pp. e42273.

Whitmore, L. and Wallace, B.A., 2004. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic acids research*, *32*(suppl_2), pp.W668-W673.

Yesudhas, D., Batool, M., Anwar, M.A., Panneerselvam, S. and Choi, S., 2017. Proteins recognizing DNA: Structural uniqueness and versatility of DNA-binding domains in stem cell transcription factors. *Genes*, *8*(8), p.192.

Zawel, L. and Reinberg, D., 1993. Initiation of transcription by RNA polymerase II: a multi- step process. *Progress in Nucleic Acid Research and Molecular Biology, 44*, pp. 67-108.

Zwanzig, R., Szabo, A. and Bagchi, B., 1992. Levinthal's paradox. *Proceedings of the National Academy of Sciences*, *89*(1), pp.20-22.