Unravelling the molecular dynamics of c-MYC's TAD domain: a journey from simulation optimisation to drug discovery

Sandra Sobral Sullivan

Department of Life Sciences

Imperial College London, London

Doctoral thesis

2021

Submitted in fulfilment of the requirement for the degree of Doctor of Philosophy of Imperial College London and the Diploma of Imperial College London

Abstract

c-MYC, part of the MYC family of transcription factors, is often deregulated in cancer, and since the early 1980's has been identified as a prime oncogenic factor. Despite much research interest, c-MYC's structural dynamics remain largely uncharted due to its intrinsic structural disorder. Disordered proteins are challenging to study using solely structural experimental methods, thus lately attention has turned towards the development of reliable *in-silico* methods to get an accurate molecular description. Molecular Dynamics simulations, commonly and successfully used to study globular proteins, can also be optimised to correctly reproduce natural protein disorder. The simulation results were assessed for convergence and conformational equilibrium, achieved by comparing the c-MYC's Molecular Dynamics conformational landscape to similar data derived from an abundantly sampled probabilistic distribution. After the preparatory and validation work, the efforts turned to the appraisal of c-MYC's first 88 amino acids. The revelation of its conformational states and structural dynamics opened the door for drug discovery and proof-of-concept that c-MYC should not be considered 'undruggable'. Further exploration into the protein first 150 residues, corresponding to its transactivation domain, uncovered important structural dynamics controlled by key phosphodegron residues. Phosphorylation and mutagenesis studies demonstrated how these control mechanisms, which serve to modulate accessibility to crucial regions, are facilitated by isomerisation events within the phosphodegron. Overarchingly, this study substantiates the robustness of well-parameterised computational simulations, and machine learning methods, in uncovering the workings of otherwise difficult to study disordered proteins.

Declaration of Originality

The content of this thesis is original and the outcome of my own research work, unless acknowledged otherwise. No part of this work has been previously used in pursuit of any other higher education degree. All third-party work has been appropriately referenced.

Copyright Declaration

Preface and acknowledgments

There are so many people that directly and indirectly contributed, helped, nurtured, and supported my work that it is impossible to do them all justice. My heartfelt thank you to all my friends, family, co-workers, and fellow students.

A special thank you to my supervisor for his unwavering help and for allowing me the freedom to work independently and pursue my research interests.

I would like to thank Drs Song and Chen for supplying the parameter set and the Perl script used to implement the ff14IDPs and the ff14IDPSFF force fields. And, likewise, Drs Skepö and Henriques for providing the Histatin 5 SAXS data.

Lastly, I would like to thank my husband for his love, care, and steadfast support – none of this would have been possible without his help.

Publication

The work in Chapters I, II and III has been published under the reference:

S. Sullivan, and R. Weinzierl. (2020) **Optimization of Molecular Dynamics Simulations of c-MYC$_{1-88}$ – An intrinsically disordered system**. *Life* 10 (7), 99-115. Doi: https://doi.org/10.3390/life10070109

Contents

Abbreviations

aMD: accelerated molecular dynamics

bHLHZ: basic-helix-loop-helix-zipper

CD: circular dichroism

cMD: conventional molecular dynamics

GB: Generalised Born

GPU: graphics processing units

Hbonds: hydrogen bonds

IC: independent component

IDP: Intrinsically disordered protein

ILE130: c-MYC residue isoleucine 130

KDE: kernel density estimation

MB: MYC boxes

MB0: MYC box 0

MBI: MYC box I

MBII: MYC box II

MC: Monte Carlo

MCMC: Markon-Chain Monte Carlo

MD: Molecular Dynamics

MSM: Markov State Model

MYC150: The first 150 residues of the c-MYC protein

MYC352: c-MYC's first 352 residues which corresponds to the protein minus the bHLHZ domain

MYC88: the first 88 residues of the c-MYC protein

NMR: Nuclear Magnetic Resonance

PC: principal components

PCA: principal component analysis

PCCA++: Perron-Cluster Cluster Analysis

PIN1: prolyl isomerase

pSER62: phosphorylated SER62

pTHR58: phosphorylated THR58

Rg: radius of gyration

RMSD: root mean square deviation

RMSF: root mean square fluctuations

S62E: Mutation of serine 62 for glutamic acid

SASA: solvent-accessible surface area

SAXS: Small-angle X-ray scattering

SER62: c-MYC residue serine 62

SSE: sum of squared errors

SSP: secondary structure propensities

T58E: Mutation of threonine 58 for glutamic acid

THR58: c-MYC residue threonine 58

TICA: time-lagged independent component analysis

Table of Figures

Table of Supplementary Figures

Table of Tables

Introduction

1. c-MYC's research interest timeline

It was the discovery of tumour-inducing viruses, and the seminal work of virologist Peyton Rous, which led to the discovery of oncogenes, including *MYC.* Well before the advent of genetic material isolation, Peyton showed that malignant cellular transformation could be transmitted between animals through cell-free chicken sarcoma filtrates (Rous, 1911). This spurred intense research into the identification of retroviruses, the understanding of retroviral replication mechanism and towards the description of the transforming genetic sequences (Varmus, 1984). One of the identified sequences was MC29, an avian leukosis virus, which aberrantly transforms myeloid cells and causes myelocytomatosis – a pathology which later gave *MYC* its name (Mladenov et al., 1967). The advances in molecular biology, including the technique of hybridisation in solution, allowed researchers to identify the *v-MYC* gene (the viral homolog of *c-MYC*), and demonstrate that *v-MYC* is the transforming oncogene in MC29 (D Sheiness, L Fanshier & J M Bishop, 1978; Pamela Mellon et al., 1978). When, in 1982, *c-MYC* was finally cloned and isolated (B Vennstrom et al., 1982) this achievement propelled a tidal wave of scientific interest into c-*MYC*-driven tumorigenesis (**Figure 1**).

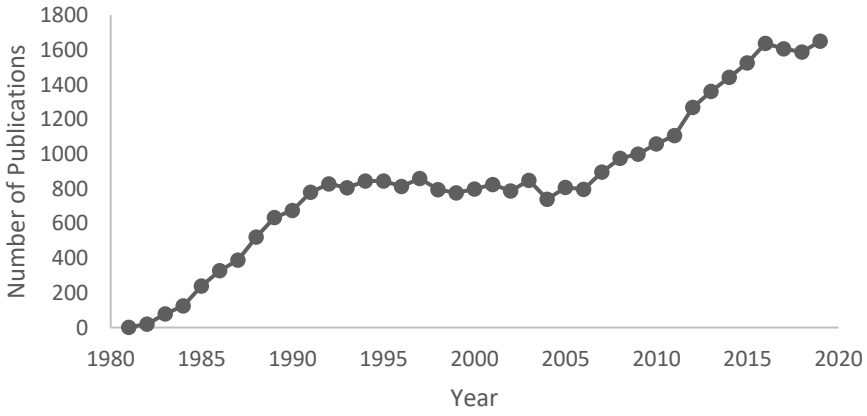

Figure 1 – *c-MYC* timeline. The figure shows the number of publications by year, listed on the NCBI's PubMed database, which used 'c-MYC' as an indexing keyword - from 1980 to 2019.

The increased research efforts, which started in the early 1980's, succeeded in identifying *c-MYC* as a key player in most human cancers. The subsequent 2006 'boom' in MYC interest was caused by the discovery of *c-MYC* as an essential factor, alongside *Sox2*, *Oct4*, and *KLF4*, in reprogramming differentiated cells and inducing them back to a pluripotent stem-cell state (iPS) (Takahashi & Yamanaka, 2006). Since then, interest in *c-MYC* has been steadily increasing and shows little signs of slowing down.

2. The MYC family of proteins

c-MYC is a notable member of the MYC basic-helix-loop-helix-zipper (bHLHZ) transcription factor family, which also includes N-MYC and L-MYC. The role of the MYC family in tissue development and maintenance is well established. Research involving c-MYC knockout (KO) mice has shown that c-MYC KO is lethal after 10.5 days of embryonic growth and generates specimens of abnormal small size afflicted by severe developmental delays (Baudino et al., 2002; Davis et al., 1993). The N-MYC KO mice tend to die a day later, at 11.5 days of embryogenesis whilst the L-MYC KO does not seem to compromise mice viability. This suggests that L-MYC can be substituted by other MYC family proteins and/or is non-essential for embryonic progression (Charron et al., 1992; K S Hatton et al., 1996; Stanton et al., 1992).

The three MYC family proteins, despite different lengths due to differences in the non-conserved regions - 439 amino acids (c-MYC), 464 amino acids (N-MYC), and 364 amino acids (L-MYC) - display high-structural homology. They are all contain an N-terminus transactivation domain, a central section containing a nuclear localization sequence (NLS) and a DNA binding domain at the C-terminus. Also, all three MYC proteins contain similar stretches of highly conserved sequences, termed MYC boxes (MB), which display up to 95% homology (**Figure 2**) (DePinho et al., 1986; Jacob Sarid et al., 1987; Lawrence W. Stanton, Manfred Schwab & J. Michael Bishop, 1986; Legouy et al., 1987).

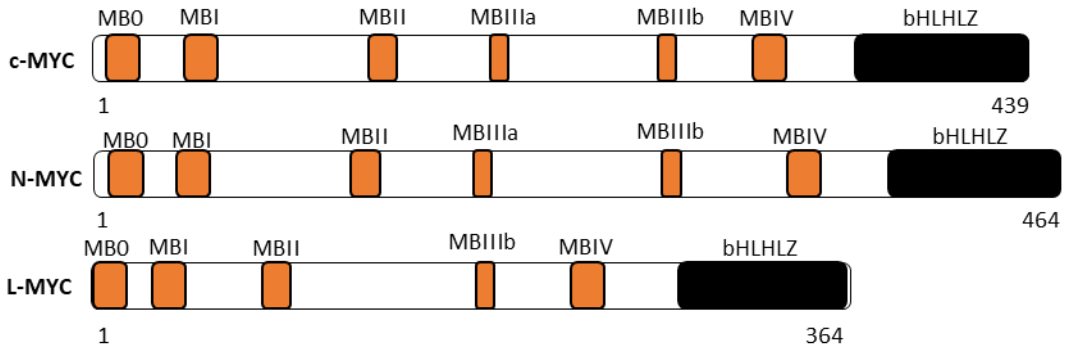

Figure 2 – The MYC family of proteins. It shows the location of the MYC boxes (MB) and the basic helix-loop-helix leucine zipper region (bHLHLZ) in the different MYC proteins.

Given the high degree of structural homology, it is unsurprising that all MYC proteins share the same basic function of transcriptional regulation. Each MYC protein, however, has a unique expression pattern and is dysregulated in different types of cancer. N-MYC is mainly expressed during embryogenesis in the neural tissues of the growing brain, fore and hindbrain. After the embryonic stage, N-MYC's expression is strictly restricted, with little to no N-MYC being expressed in fully differentiated tissues (Minna *et al*., 1986). Consequently, when present in cancers, N-MYC's overexpression is most frequently observed in tumours of neural origin, such as gliomas and neuroblastomas (Schwab, 2004). L-MYC, the least studied of the MYC proteins, is known to be expressed in both embryonic and mature lung tissue and most markedly dysregulated in small cell lung carcinomas (Bertness et al., 1985). Of the three transcriptor factors, c-MYC is undoubtedly the most ubiquitous and well expressed in any proliferative tissue, both during embryonic/foetal development and in adult animals. Its ubiquity implicates c-MYC in the pathogenesis of a great number of malignancies, up to ~60-70% of solid and hematopoietic tumours (García-Gutiérrez, Delgado & León, 2019). Unquestionably, this makes c-MYC a coveted target for structural and molecular research, drug discovery and the main subject matter of this thesis.

3. c-MYC's function

Under normal circumstances, c-MYC is deemed a "master regulator" and controls the expression of a staggering number of genes. These genes encode for proteins involved in

crucial cellular functions including (1) cell growth and cell cycle control; (2) cell metabolism and mitochondrial biogenesis; and (3) ribosome biogenesis and protein synthesis (Wahlström & Arsenian Henriksson, 2015). Recent studies have suggested that c-MYC binds to every active gene in any given cell type and increases universal transcription by promoting RNA Polymerase II pause release at the *loci* of already actively transcribed genes (Lin et al., 2012; Nie et al., 2012; Rahl et al., 2010). According to this model, c-MYC acts as an amplifier of pre-existing gene transcription programmes, rather than establishing a new one as a sequence-specific transcriptional activator. This is still subject of much debate because it fails to account for c-MYC's role as a repressor, or if c-MYC-bound genes are responding directly to c-MYC's activity or to a global transcriptional amplification of which c-MYC is only a part (William P. Tansey, 2014).

c-MYC expression is tightly regulated and increases in response to growth factors, nutrient abundance, and mitotic stimuli, including signalling from epidermal growth factor receptors and platelet-derived growth factor receptors (Oster et al., 2002), and c-MYC rapidly rebounds to basal levels as the cell progresses through the cell cycle (Kelly et al., 1983). c-MYC plays multiple roles in promoting cell cycle progression: (1) it stimulates the transcription of cell cycle inducers; (2) represses the activity of cell cycle inhibitors by either blocking their transcription (e.g. p21) or enhancing their degradation (e.g. p27); (3) and, together with E2F transcription factors, c-MYC upregulates the synthesis of proteins involved in replication initiation (Bretones, Delgado & León, 2015).

In general, c-MYC's function is achieved by direct transcriptional activation, or repression, of its target genes and is modulated by changes in c-MYC expression levels (Walz et al., 2014). c-MYC can additionally enhance mRNA 5' end capping by the enzymes RNGTT and RNMT-RAM and protect mRNA transcripts from degradation, thus promoting increased rates of translation (Dunn & Cowling, 2015). c-MYC is also known to bind ribosomal DNA, activate rDNA transcription, regulate the nuclear RNA polymerases' activity, especially I and

III, and upregulate ribosome biogenesis in order to stimulate cell growth (Eisenman et al., 2003; Galloway et al., 2005; Ridderstråle et al., 2005).

Recently, c-MYC has been implicated in a partnership with mTOR kinase, the mammalian target of rapamycin, to enhance protein synthesis in the cell. c-MYC facilitates mTOR's direct phosphorylation of the tumour suppressor eukaryotic translation initiation factor 4E (eIF4E) binding protein 1 (4EBP1). The phosphorylation negatively affects 4EBP1's capacity to downregulate the translation initiation factor eIF4E, consequently promoting eIF4E's ability to build up the engagement between mRNAs 5′-cap and the 40S ribosome subunit (Michael Pourdehnad et al., 2013).

Moreover, c-MYC is a general chromatin regulator, able to regulate the expression of, and directly bind to, chromatin-modifying complexes (Cheng et al., 2006). c-MYC binds the TRRAP coactivator, which subsequently recruits the histone acetyltransferase GCN5 to promote the histone acetylation and modulate accessibility to target DNA E-box binding sites (Orian et al., 2003).

Somewhat counterintuitive to c-MYC's role in cell proliferation is its activity as a pro-apoptotic factor (Askew et al., 1991; Evan et al., 1992; Shi, Y. et al., 1992). Paradoxically, in non-transformed cells c-MYC's role in promoting apoptosis prevents tumorigenesis (Nilsson & Cleveland, 2003), which suggests that neoplasia can only occur when c-MYC's pro-apoptotic function is abrogated by loss of tumour suppressor pathways (e.g. p53) or overexpression of anti-apoptotic factors (Fanidi, Harrington & Evan, 1992). Ultimately, c-MYC's activity is so broad, it initiates such a complex and multi-layered expansion programme that it is crucial for the cell to be able to strictly control it at every step.

4. c-MYC's tight regulation in cells

c-MYC's levels are tightly controlled at both transcription and translation stages, chiefly by modulating its stability. c-MYC's expression regulation can be achieved via microRNAs (miRNA) and long noncoding RNA control (lncRNA). Twenty-five miRNAs have been identified

as c-MYC regulators with most of them binding c-MYC transcripts directly (Swier, Lotteke J. Y. M et al., 2019). Other miRNAs, such as miR-24-3p, target c-MYC in a more indirect manner: by supressing c-MYC's *O*-GlcNAcylation via *O*-GlcNAc transferase (OGT) and causing loss of c-MYC's stability (Liu, Yubo et al., 2017).

Several lncRNAs, including CCAT1-L, have also been shown to upregulate c-MYC's transcription in *cis* by controlling chromatin interactions with the locus of c-MYC transcription (Xiang et al., 2014). Whilst other lncRNAs, such as IGF2BPs, can bind to and enhance c-MYC's mRNA stability, thereby promoting its translation efficiency (Huang, Huilin et al., 2018). Yet other lncRNA's can directly affect c-MYC protein stability by interfering with its degradation. One example is PVT1, a lncRNA that prevents the phosphorylation of c-MYC's threonine 58 and keeps the protein from being targeting for destruction (Tseng et al., 2014); and LINC01638, which prevents c-MYC from binding to the E3 ubiquitin ligase adapter SPOP (Luo et al., 2018).

Additional checkpoints for c-MYC include the tight regulation of c-MYC's transcript transport to the cytoplasm by the eukaryotic translation initiation factor (eIF4E) (Biljana Culjkovic et al., 2006); its translational supression via the activity of RNA-binding proteins; and its own very short mRNA half-life (Farrell & Sears, 2014).

Post-translationally, c-MYC is also under intense scrutiny. This type of control is often achieved by triggering the protein degradation system which relies on a complex network of protein interactions mediated by c-MYC's patterns of posttranslational modifications including phosphorylation, acetylation, glycosylation, and ubiquitylation (Hann, 2006). These modifications regulate much of c-MYC's activity in terms of transcriptional activation, repression, and protein destruction. In tandem, the activity of proapoptotic and tumour suppressor proteins such as p53, BIM, ARF and PTEN act as further barriers which keep c-MYC in balance (Stine et al., 2015). c-MYC's regulatory mechanisms are so crucial that any failure invariably leads to c-MYC dysregulation and, with it, the destructive c-MYC-induced oncogenic progression.

5. c-MYC and cancer

Given c-MYC's broad influence, even small changes in expression levels can cause large-scale abnormalities. These effects include c-MYC's oncogenic activation and are influenced by the epigenetic pattern specific to the cell type (Beer et al., 2004). c-MYC neoplasia frequently involves cooperation with other oncogenic agents driving loss of regulatory checkpoints, tumour suppressor proteins and feedback loops, which keep c-MYC dependent on mitotic stimuli (Gabay, Li & Felsher, 2014). When c-MYC reaches oncogenic levels, its far reaching interactome helps the cancer cell survive, proliferate, invade, and thrive in a hypoxic environment, by reprogramming the cancer cell's various functional pathways to support it (Dang, Chi V, 2012).

In terms of metabolic alterations, c-MYC stimulates the expression of genes involved in glycolysis and glucose uptake. c-MYC regulates the activity of the enzyme lactate dehydrogenase A (LDH-A), which converts pyruvate to lactate in the glycolytic cycle (Hyunsuk Shim et al., 1997). This has been established *in vivo*, with mice overexpressing c-MYC in liver cells also demonstrating increased hepatic glycolysis and lactate production (Valera et al., 1995). Moreover, c-MYC controls genes involved in glucose metabolism including glucose transporter GLUT1, hexokinase 2 (HK2), phosphofructokinase (PFKM), and enolase 1 (ENO1) and through their activity promotes the Warburg effect in cancer cells (Chi V. Dang, Anne Le & Ping Gao, 2009), which can be so significant that some cancer cells become addicted and undergo apoptosis if glucose deprived (Hyunsuk Shim et al., 1998).

Equally important in c-MYC-driven cancer is its dependence on glutamine as an energy source for growth and proliferation (Gao et al., 2009). This is induced by c-MYC's upregulation of key glutamine metabolism enzymes, such as the glutamine importer ASCT2 (David R. Wise et al., 2008); and c-MYC transcriptional repression of miRNAs, miR-23a and miR-23b80, leading to increased expression of mitochondrial glutaminase (GLS), which further upregulates the glutamine metabolism (Miller et al., 2012). This is an important pathway for the stressed, nutrient and oxygen deprived cancer cell, as glutamine can be used a source of

energy, nitrogen, and carbon substrate for cancer cellular anabolism (Miller et al., 2012). The (Le et al., 2012) study ties together the importance of both glycolysis and glutamine metabolism for cancer cells, by showing that c-MYC's overexpression leads to both the transformation of glucose into lactic acid, and glutamine oxidation through the tricarboxylic acid (TCA) cycle. Furthermore, it showed that in hypoxia and glucose-depleted conditions the glutamine metabolism prevails, maintaining cell survival and viability. In that inhospitable environment, glutamine was also used by the transformed cells to synthesize glutathione, a reducing agent capable of shielding the mitochondria from the accumulation of harmful reactive oxygen species (ROS). As proof-of-concept, (Le et al., 2012) study demonstrated that the inhibition of glutaminase had an apoptotic effect on tumour cells. c-MYC also plays a vital role in amplifying mitochondrial biogenesis - its nuclear genes are prime c-MYC targets (Antje Menssen & Heiko Hermeking, 2002; Guo, Q. M. et al., 2000; Hilary A. Coller et al., 2000), as well as genes coding for proteins involved in mitochondrial activity (Antje Menssen & Heiko Hermeking, 2002; Fernandez et al., 2003). This directly implicates c-MYC overexpression with an increase in mitochondrial-derived ROS and DNA oxidative damage, which compromises the cell's genomic integrity often observed in c-MYC-driven neoplasia (Dang, Chi V., Li & Lee, 2005). On the other hand, c-MYC is known to promote the overexpression of mitochondrial serine hydroxymethyltransferase (SHMT), notably SHMT2, in order to protect cancer cells from oxidative damage in hypoxia conditions (Ye et al., 2014). Moreover, c-MYC modulates mitochondrial function by promoting the synthesis of acetyl-CoA which is then used to promote lipid biosynthesis and histone acetylation (Morrish et al., 2010).

Among c-MYC's many functions in cancer, it has been suggested that c-MYC directly promotes transcription of genes responsible for *de novo* nucleotide synthesis proteins (PRSP2, inosine monophosphate dehydrogenase and thymidylate synthase) and enhances serine and glycine synthesis from glycolytic intermediates (Liu, Yen-Chun et al., 2008; Vazquez, Markert & Oltvai, 2011). Furthermore, it increases cell membrane synthesis by upregulating genes encoding for *de novo* fatty acid biosynthesis (ACC, FASN and SCD), whilst

keeping them from being used as energy sources (Hsieh et al., 2015; Karen I. Zeller et al., 2006). And, also known to upregulate cholesterol synthesis by promoting the expression of the enzyme hydroxymethylglutaryl coenzyme A reductase (HMG-CoA reductase) (Zhong et al., 2014). Lastly, recent research has uncovered c-MYC's role in influencing the tumour environment by promoting angiogenesis and metastasis. In hypoxic conditions, c-MYC was found to inhibit the microRNA tumour suppressor cluster miR-15-16 by upregulating HIF-2α and causing the loss of posttranscriptional repression of the angiogenic growth factor FGF2 (Xue et al., 2015). Although c-MYC's broad influence in cancer pathogenesis is becoming clearer, much remains to be explored. This is especially true when it comes to c-MYC's own structural dynamics, the relationship between structure and function and the rationale behind c-MYC's interactions decisions.

6. c-MYC's architecture and interactome

c-MYC contains a DNA binding domain that is followed by a helix–loop–helix/leucine zipper (bHLH-LZ) motif at the C-terminal. The DNA binding domain is the *locus* of c-MYC's heterodimerisation with its partner molecule Max (**Figure 3**). This c-MYC/Max partnership enables the molecular complex to recognise and bind to DNA E-box sequences and recruit the transcriptional machinery required to activate specific target genes (Thomas et al., 2015).

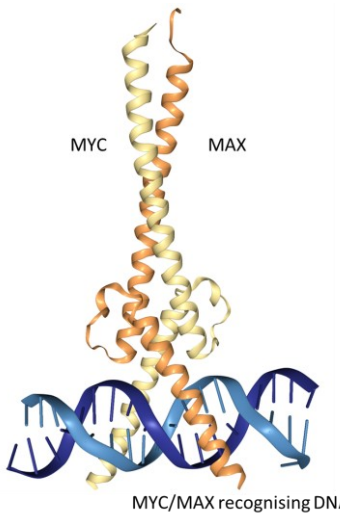

Figure 3 – Figure depicts the crystal structure of the c-MYC-MAX heterodimer binding a DNA molecule (PDB# 1NKP acquired from the Protein Data Bank).

The human c-MYC protein contains several highly conserved regions termed MYC boxes (MB). These MBs, as well the bHLH-LZ DNA binding domain, facilitate interactions between c-MYC and its molecular partners (**Figure 4**).

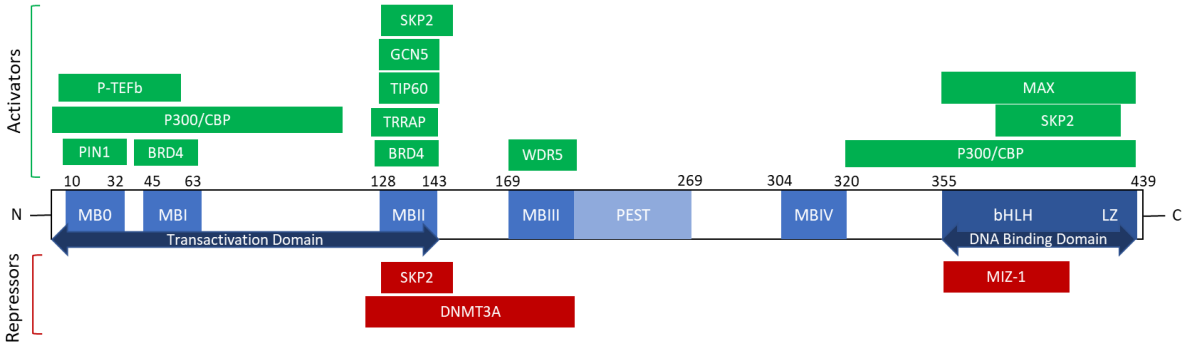

Figure 4 – This diagram depicts c-MYC's main domains and the MYC boxes location as well as the binding sites of some of the co-factors, transcriptional activators and repressors, that are known to interact with c-MYC.

The first three MB are located within c-MYC's transcriptional activation domain (TAD) which spans residues 1 to 143. MB0, comprised of residues 10 to 32, has been identified as the binding site for the prolyl isomerase PIN1 (Helander et al., 2015). It has been proposed that PIN1 regulates c-MYC activity by allosterically modulating proline isomerisation in MBI's phosphodegron (Helander et al., 2015). MBI, spans residues 45 to 63 and contains within it a phosphodegron with two phosphorylation sites - THR58 and SER62. The regulation of the MBI's phosphodegron phosphorylation is especially important since it directly dictates c-MYC's fate (Helander et al., 2015).

Figure 5, which summarises the signalling pathways promoting c-MYC's transcriptional activation and degradation. In response to growth factors, c-MYC is phosphorylated at serine 62 via mitogenic-stimulated kinases, such as extracellular signal-regulated kinases (ERKs) and CDKs (Sears, 2004).

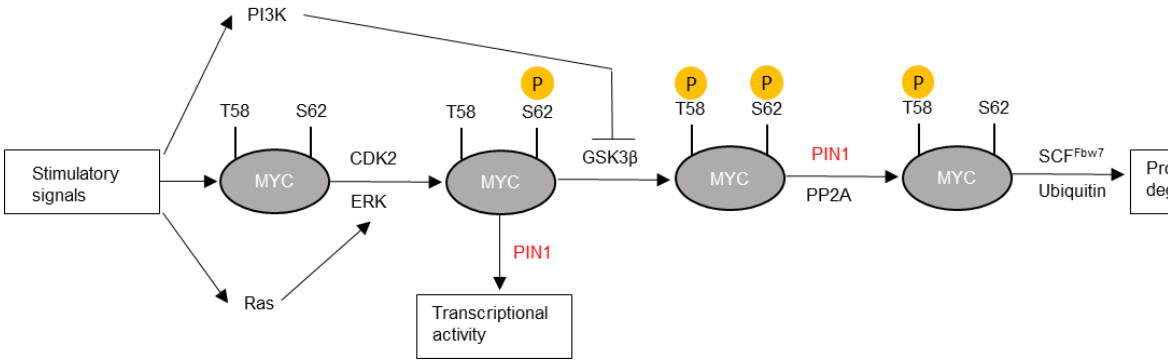

Figure 5 – This diagram, adapted from the Helander et al. (2015) paper, highlights the transcriptional activation and proteasomal degradation pathways for c-MYC. These pathways involve PIN1-mediated isomerisation control of successive phospho and dephosphorylation events in MBI phosphodegron residues (T58 and S62).

Upon serine 62 phosphorylation PIN1 isomerisation activates a stabilised c-MYC for transcriptional activity (Amy S. Farrell et al., 2013). Serine 62 phosphorylation also promotes the subsequent threonine 58 phosphorylation via the activity of glycogen synthase kinase 3-β (GSK3β) (Mark A. Gregory, Ying Qi & Stephen R. Hann, 2003). c-MYC now displaying threonine 58 and serine 62 phosphorylation, encourages a second PIN1 isomerisation which induces PP2A phosphatase to dephosphorylate serine 62 (Arnold & Sears, 2006). With only threonine 58 phosphorylation remaining, c-MYC is now recognised by the Skp1-Cullin-F-box protein-Fbxw7 (SCFFbxw7) ubiquitin ligase which mediates ubiquitylation and subsequent proteasomal degradation by the 26S proteasome (Ishida et al., 2004; Markus Welcker et al., 2004).

c-MYC can be degraded via alternative pathways, notably by the action of SKP2 - a ubiquitin kinase that, somewhat paradoxically, can simultaneously act as a c-MYC transcriptional activator as well as an inhibitor (Kim et al., 2003).

MBII spans residues 128 to 143 and is a critical region for the recruitment of multiple key c-MYC co-factors and interactors. Of these, the transformation/transcription domain-associated protein, or TRRAP, is perhaps one of the most important. TRRAP is an adaptor protein which provides scaffolding for the assembly of protein complexes with other co-factors. It forms STAGA complexes which possess histone acetyltransferase (HAT) activity (Zhang,

N. et al., 2014). These HAT-containing complexes include the General Control of Amino Acid Synthesis Protein 5-Like 2 (GCN5) and the 60kDa Tat Interacting Protein (TIP60), and mediate histone acetylation and gene activation on target promoters (Jagruti H. Patel et al., 2004; Steven B. McMahon, Marcelo A. Wood & Michael D. Cole, 2000). The p300 and cyclic AMP response element-binding protein (CBP), which are also transcriptional co-factors and histone acetyltransferases themselves, are known to bind c-MYC at two locations. The first binding site, spanning residues 1 to 110, mediates c-MYC's transactivating activity and acts as a c-MYC regulator by promoting direct lysine acetylation and protein instability (Francesco Faiola et al., 2005). The second binding site at the C-terminus is reported to function as an activator (Austen et al., 2003). Additionally, the MB0-MBII TAD region interacts with other regulatory c-MYC partners: the BET bromodomain protein BRD4, to promote transcriptional regulation at the initiation and elongation steps and the P-TEFb which controls the transcriptional pause-release at active genes (Rahl & Young, 2014). The MBIII is another important locus of interactions, particularly with the WD repeat-containing protein 5, or WDR5. The c-MYC-WDR5 complex enhances chromatin-binding at target genes and play a crucial role in gene recognition (Thomas et al., 2015). Downstream of the WDR5's binding site, c-MYC contains a PEST sequence, spanning amino acids 207 to 269, very common in short lived proteins (Rechsteiner & Rogers, 1996). The PEST region is implicated in efficient c-MYC proteolysis and c-MYC mutants deprived of its PEST sequence become much stabilised (Gregory and Hann, 2000). Additionally, c-MYC has been found to contain, at residue lysine 298, a calpain-sensitive cleavage site. The proteolytic cleavage at this site creates a cytoplasmic c-MYC product termed 'MYC-nick'. MYC-nick was found to promote acetylation of microtubules, involving to recruitment of the TRRAP-GCN5 complex, with an important role in myogenic differentiation (Conacci-Sorrell, Ngouenet & Eisenman, 2010; Mousavi & Sartorelli, 2010).

Lastly, c-MYC is known to sequester the zinc finger protein MIZ-1. This interaction transforms c-MYC from a transcriptional activator to a repressor. It enables it to recruit DNA methylases (including DNA methyltransferase DMNT3A), histone deacetylases and polycomb

proteins to repress transcription (Corvetta et al., 2013; Kouzarides et al., 2005; Licchesi et al., 2010; Zhang, X. et al., 2012). The c-MYC-MAX-MIZ-1 complex is involved in repression of cell cycle inhibitors such as CDK inhibitor p15Ink4b (CDKN2B) and is implicated in preventing cell growth inhibition (Seoane et al., 2001).

Overall, research into c-MYC's MBs and its interactors emphasizes how crucial these regions are for gene transactivation and transrepression. However, despite its importance, the exact mechanism(s) behind c-MYC's interaction decisions with such a variety of co-factors is still scantily known (Tu et al., 2015). Questions such as how much each interactor contributes to c-MYC's role in tumorigenesis, have no clear answer. c-MYC's promiscuous behaviour, establishing multiple partnerships with a growing list of co-factors, firmly plants c-MYC at the centre of many signalling pathways - each if targeted, could potentially halt the oncogenic progression. However, this would require detailed information into c-MYC's structural dynamics and how these relate to c-MYC's interactions – such endeavour has remained elusive. Indeed, without a clearer understanding of the protein's structure and its conformational dynamics, drug discovery or any attempts to target the c-MYC transcriptional cascade are severely challenged. The main reason for the lack of structural insights is c-MYC's intrinsic disorder. Intrinsically disordered proteins (IDPs) structural dynamics differ from canonical proteins. Instead of folding predictably, according to it amino acid sequence, IDPs exist in a rapidly changing ensemble of configurations. This structural diversity allows them to easily bind multiple partners, co-factors and be at the centre of important cellular pathways (Levine & Shea, 2017). IDPs behaviour occurs due to their peculiar amino acid composition, an enrichment in polar amino acids coupled with depletion of hydrophobic residues hinders the formation of hydrophobic cores, causes destabilisation of the protein fold and allows IDPs to sample a wide range of alternate conformations (Babu, 2016). The greatest challenge when studying IDPs is finding methods which preserve and accurately describe their rich dynamic disorder. Conventionally applied structural methods, such as x-ray crystallography, are often unsuitable to assess IDPs – these proteins do not form crystals and inhabit a vast number of

conformations which cannot be described by a few high-resolution models. Even Nuclear Magnetic Resonance (NMR) studies are not a standalone option. The NMR ensemble-averages provide insufficient data to create a robust protein model accounting for state transitions, or pocket formations - helpful in the drug discovery process. Recently, *in silico* methods such as molecular dynamics (MD) simulations have been proposed as promising alternative methods for IDP exploration. MD simulations can reveal with atomistic detail the structural dynamics of the protein over time, its intra and interprotein interactions and target it with drug discovery protocols (Chong, Chatterjee & Ham, 2017).

7. Using Molecular Dynamics Simulations to study IDPs

MD simulations started with the simulation of simple gasses in the 1950's (Alder & Wainwright, 1957), and progressed to biological systems in the 1970's with the MD simulation of the bovine pancreatic trypsin inhibitor (Gelin, Karplus & McCammon, 1977). Since then interest in MD simulations has soared, owing to an exponential increase in molecular biology papers, and associate disciplines, presenting MD results to help interpret and guide experimental work (**Figure 6**).

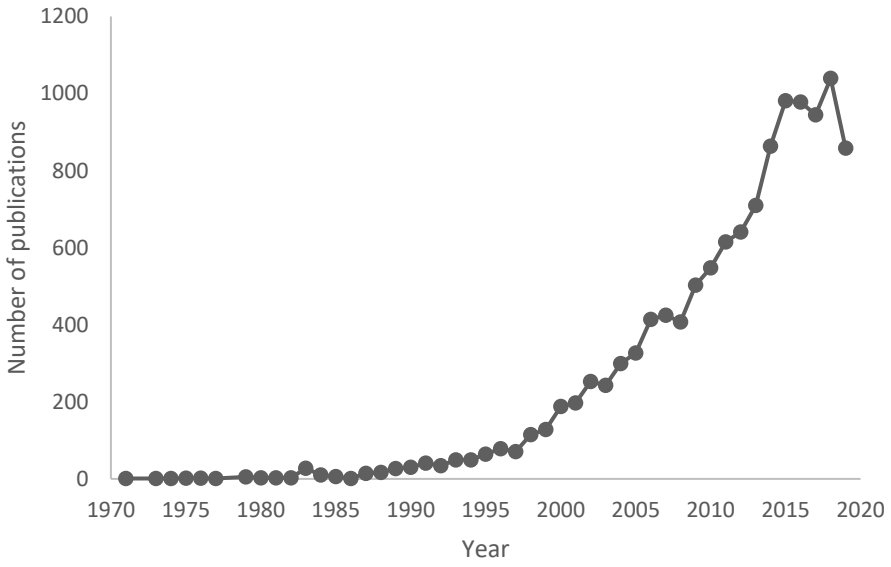

Figure 6 – **Molecular dynamics simulations publications timeline**. The figure shows the number of publications by year, listed on the NCBI's PubMed database, that include the term "molecular dynamics" in either the title, abstract or keywords. - from 1970 to 2019.

The increased interest in MD simulations was facilitated by the rise in efficient and accessible computational resources, particularly since the early 2000's. Nowadays, MD simulations no longer require Anton-like supercomputers but can be run on regular computer hardware using graphics processing units (GPUs), which permits low cost simulation parallelisation (John E Stone et al., Jan 1, 2016; Salomon-Ferrer et al., 2013). Over time, the MD simulation software packages have become more user-friendly and the physical models, underpinning the biological systems simulation estimates, notably increased in accuracy (Hollingsworth & Dror, 2018).

MD simulations are established as the evolution of cartesian coordinates for every atom in a system, using a general physics model governing particle interaction (McCammon & Karplus, 2002). Given a starting set of atomic coordinates it is possible to calculate the velocity, directionality and spatial location of each atom based on the force exerted on it by the remainder atoms in the system. These calculations over time are done by numerically solving Newton's laws of motion and result in a trajectory of atoms. This trajectory enables the study of a variety of processes, including structural and conformational changes, protein-protein interactions, ligand binding, drug discovery, protein folding and energetics and the system's response to perturbations, such as mutations and posttranslational modifications (Hollingsworth & Dror, 2018). MD simulations allows research that would otherwise be impossible, such as determining the exact position of every atom in a molecule at any point in time. It also affords absolute control over the experiment's conditions – from the initial conformation of the system, to the temperature, presence of ligands, presence of interactors, protonation state, solvent and ions, pH, etc. Modifying and comparing across different parameters makes it is possible to investigate a very wide scope of research questions. Once the parameters are decided and the system is created, the forces exerted on each atom are calculated by deriving equations, known as force-fields, where the potential energy is extracted directly from the molecular structure. Force fields are empirical potential energy functions derived from the molecule's bonded interactions (bonds, angles, and dihedrals

potentials) and the nonbonded interatomic electrostatic interactions (van der Waals and Coulomb potentials) between every atom, to calculate the total energy the system (Hospital et al., 2015).

Although performing MD simulations is now relatively straightforward and the computational resources commonly accessible, there is still much to be explored in terms of implementing adequate methods for studying different biological systems. There is much to be optimised regarding the *in silico* experimental design, comparing experimental data to simulation data and interpreting noisy MD results to gain valid biological insights.

8. Aims of the project

This work focuses on exploring c-MYC's transactivation domain, for which there is little information concerning its structural dynamics. This lack of knowledge has impaired any solid understanding of c-MYC's intra and intermolecular interactions and has dampened attempts for drug discovery. Given the importance of intrinsically disordered transcription factors, such as c-MYC, in molecular biology and cancer research, it is vital to find new ways to study these systems. This project addresses these questions by using a combination of MD simulations, experimental data, and machine learning analysis to shed light on c-MYC, which for far too long has been deemed an 'undruggable black box'.

Materials and Methods

1. MD simulations setup

The MD simulations were created using the AMBER16 MD simulation package (Case et al., 2016). To study the force fields, the only difference in the simulation preparation was the force field used: the conventional AMBER ff14SB force field (Maier et al., 2015); the ff14IDPs force field (Song et al., 2017); and the updated ff14IDPSFF (Song, Luo & Chen, 2017). All other aspects of simulation parameterisation remained the same:

- LEaP (part of AmberTools 16 analysis suite) was used to parameterise and prepare the structures for simulation (Case et al., 2016) . TIP3P water model was used to solvate the system for all simulations. A periodic water box was created with a 15 Å distance between the Histatin 5 molecule and the limits of the box. The solvation environment was enriched with Na+ and Cl- ions to a final concentration of 150 mM NaCl. LEaP's output files (.prmtop and .inpcrd files) were used to run the conventional molecular dynamics (cMD) simulation. An example of the system preparation file can be found in the supplementary information section **suppl. Figure S8**.
- The cMD consisted of two successive minimisations (min1 and min2). The minimisations were followed by two molecular dynamic stages (md1 and md2). Files with input examples can be found in **suppl. Figure S9**.
    - Min1 consisted of solvent minimisation run with the protein fixed, 10000 maximum cycles and 5000 ncycles of steepest descent.
    - Min2 consisted to a total system minimisation with 2500 maximum cycles and 1000 ncycles of steepest descent.
    - Md1 entailed 100 ps of MD with weak restraints on the protein to reach a temperature of 310K. It used Langevin dynamics for temperature control. A total 50000 MD-steps were performed at a 0.002 ps time step.

- Md2 created a 1000 ns simulation of the whole unrestrained system at a temperature of 310K. It used Langevin dynamics for temperature control. A total 500000000 MD-steps were performed at a 0.002 ps time step.
- Simulations were executed via the pmemd.cuda (Amber 16), relying on exclusive GPU usage, to obtain total potential (EPTOT) and dihedral (DIHED) energy values used for setting up the accelerated molecular dynamics (aMD) (Salomon-Ferrer et al., 2013).

- The EPTOT and DIHED energy values obtained from the cMD were used to calculate thresholds and to run the aMD also on pmemd.cuda. Each aMD created a 1000 ns simulation of the system at 310K also using Langevin dynamics as a thermostat. An example file for aMD setup can be found in **suppl. Figure S10**.

The explicit water models' simulations were setup using the exact same protocol both the TIP3P simulations and the TIP4-D water model simulations, as developed by (Piana et al., 2015).

For the implicit solvent simulations, the starting structure was prepared with LEaP to using the modified ff14SBonlysc as a force field without any water box parameters. Several GB implementations were tested but only the GB8 model for solvation was found to accurately describe the protein and was used to execute the simulations. Since the simulations did not contain water molecules, only a short total system minimisation was performed consisting of 5000 maximum cycles and 2500 ncycles of steepest descent. After the minimisation, the simulations were executed for a total of 1000 ns per run at a temperature of 310K, using Langevin dynamics as a thermostat and a time step of 0.002 ps. The protocols used to prepare the Histatin 5 simulations were replicated for the MYC88 and MYC150 simulations. The MYC88 refers to c-MYC's first 88 amino acids and MYC150 to the first 150 amino, which correspond to the entire transactivation domain. The starting structure for all Histatin 5 simulations was created unstructured to avoid conformational bias using PyMOL v1.8.6.0 (Schrodinger, 2010). The starting structures for both MYC88 and MYC150 simulations were

created using QUARK *ab initio* protein structure prediction software (Xu & Zhang, 2012; Xu & Zhang, 2013). MCMC representative structures obtained via K-means clustering, using the cluster centres representative structures were also used as input coordinates. The starting structures for the phosphorylated pTHR58 and pSER62 were prepared completely extended using LEaP (Case et al., 2016), based only on their amino acid sequence.

2. MCMC simulations

The Markov chain Monte Carlo (MCMC) simulations used the PHAISTOS programme package for protein structure inference (Boomsma et al., 2013). An example for the MCMC simulation input file can be found in **suppl. Figure S13.** Two sets of MCMC independent simulations were set up with 25 threads each. Each thread simulated a total of 2000 structures. Therefore, a total of 100000 structures were obtained for analysis. To avoid any structural bias, the simulations were parameterised using the amino acid sequence as sole input. The backbone and sidechains were efficiently sampled using both pivot-uniform and sidechain-uniform moves. These moves are widely used in Monte Carlo simulations and produce a random, uniformly distributed rotation of the dihedral (φ, ψ angles) and sidechain torsion angles (χ angles) in single residues. The energy terms integrated the highly efficient Profasi force-field, parameterised to simulate interactions in the presence of a solvent. The Metropolis-Hastings algorithm was used as the acceptance criterion for the simulation method.

3. Trajectory analysis

The geometric measures such as RMSD, Rg, distances, solvent-accessible surface area (SASA), hydrogen bonds, dihedrals and the trajectory clustering analysis in Chapter I was calculated using CPPTRAJ (Roe & Cheatham, 2013). An example file for CPPTRAJ calculations can be found in **suppl. Figure S11**. The simulation's secondary structure propensities calculation was done using VMD's (Humphrey, Dalke & Schulten, 1996) timeline feature. The Sα assessment was conducted using the open-source PLUMED library version 2 (Tribello et al., 2014) implementing (Pietrucci & Laio, 2009) protocol. The Ramachandran and

Psi-Omega plots were calculated using the Dihedral module of the MDAnalysis Python package (Gowers et al., Sep 11, 2016; Michaud-Agrawal et al., 2011).

Plots and graphs were created using custom Python scripts and the Plotly package, GraphPad Prism version 8.3.1 for Windows (GraphPad Software, 2020) and Gnuplot 5.2 (Williams et al., 2018).

4. PCA and TICA analysis

The principal component analysis (PCA) calculation of the XYZ trajectory coordinates was obtained using the R package Bio3D (Grant et al., 2006). The dihedral PCA 'featurised' with the backbone dihedrals was calculated according to the published protocol (Mu, Nguyen & Stock, 2005; Sittel, Jain & Stock, 2014)using the GROMACS analysis tools package (Abraham et al., 2015). The PCA using the pairwise distances between alpha carbons was calculated using the Python MDtraj package (McGibbon et al., 2015). The time-lagged independent component analysis (TICA) and associated plots was produced using the Pyemma 2.5.7 package for Python (Scherer et al., 2015). The TICA was performed using the backbone torsion angles for featurisation at a lag of 20 nanoseconds projected over two independent components. The Rg of gyration of 45 structures per TICA state were considered to calculate the average. Example scripts for both the PCA and TICA analysis can be found in the supplementary information section - **suppl. Figures S12 and S15**, respectively.

5. Rg peak detection

The peak minimum and maximum analysis of the radius of gyration over time plot was done using the argrelextrema package imported from the scipy.signal module. The local peaks were found using a window of 500 to avoid the noise of neighbouring unimportant peaks. An example script for the peak analysis can be found in **suppl. Figure S14**.

6. Network analysis and contact activity

The contact analysis was done by deploying a Python custom script which used CPPTRAJ-calculated distance data between each MYC150 residue and every other residue

in the protein. The cut off distance was 10 Å, and an offset of 10 residues, allowed for local interactions to be removed to focus only on long-range interactions. The network analysis used the same CPPTRAJ distance calculations to establish the functionally important residues for the web of intramolecular contacts. This was created using the Python networkx package.

The monitoring of contact activity in the MD simulation trajectories was achieved using the python module tagging.py from the D.E. Shaw's Timescapes 1.5 suite of programs (Kovacs & Wriggers, 2016; Wriggers et al., 2009). The cutoff contact distance was set at 6.0 angstroms. The turning.py module from Timescapes 1.5 was used to map important residues based on their correlations of backbone pivot angles, which display hinge bending. The pivot angles coefficient calculations are based on the 'pseudodihedral' angles created by four consecutive α carbons.

7. Pocket prediction, docking setup and drug discovery

The identification of a druggable region entailed the deployment of several predictive methods: CASTp 3.0, a geometry-based pocket calculation (Dundas et al., 2006); FTMap for the detection of binding hotspots using organic probes (Kozakov et al., 2015); Pockdrug which calculates suitable pockets by assessing correct geometry and biochemical composition (Hussein et al., 2015); MDPocket to assess the stability of the pocket over time (Schmidtke et al., 2011).

The process of drug screening and protein-ligand docking was conducted using AutoDock Vina and the established Vina protocol described in literature (Trott & Olson, 2010) and iDock (Hongjian Li, Kwong-Sak Leung & Man-Hon Wong, May 2012). For both tools the exact same coordinates, corresponding to the calculated pocket, were used as the docking site. The compound screening searched for suitable small ligands from the ZINC the libraries 'All clean', 'Natural Products' and 'FDA approved', part of the freely available ZINC database (Irwin et al., 2012). A total of 23,221,614 compounds were screened - 23,129,083 compounds coming from the 'All clean' library and 92,531 compounds coming from the 'FDA approved' and

'Natural Products' combined libraries. Subsequently, a compilation containing the 10000 best scoring ligands, in terms of their binding affinity to MYC88, was analysed to identify drug candidates that satisfied the lead-like ligand conditions and/or the Lipinski rule. This was done using a custom R script deploying the GGPLOT2 package (Wickham, 2016).

8. The experimental methods

The Histatin 5 SAXS data used in this study was generated by (Cragnell et al., 2016) and acquired according to their published protocol. The Histatin 5 NMR data was derived from (Raj, Marcus & Sukumaran, 1998). The MYC88 NMR-derived secondary structure propensities were obtained from (Andresen et al., 2012) published work.

Regarding the MYC150 experimental CD data, the plasmid containing human c-MYC residues 1–150 with a N-terminal 6× His tag was expressed in BL21-DE3 competent cells and incubated at 37°C overnight. The cells were cultured at 37°C overnight in auto induction media (containing 0.6% Na_2HPO_4, 0.3% KH_2PO_4, 2% tryptone, 0.5% yeast extract, 0.5% NaCl, 1% of 60% v/v Glycerol, 0.5% of 10% w/v Glucose, 2.5% of w/v 8% Lactose) supplemented with ampicillin to the final concentration of 100ug/ml. Subsequently, the cells were harvested, pelleted and the pellet frozen at -80°C for 1 hour. The pellet was resuspended in buffer containing PBS, 10/% glycerol and 15mM of mercaptoethanol. The sample was then sonicated on ice 3 × 30 s and centrifuged at 10 000g for 10 min. Followed this, the pellet was resuspended in lysis buffer containing 1% lysozyme, 5% sodium deoxycholate and 0.2% EDTA and left to incubate at room temperature for 90 minutes. This was followed by a 10 000g centrifugation run for 10 min after which the sample was resuspended in lysis buffer and left overnight to incubate. The sample was then centrifuged, resuspended in buffer without lysozyme, centrifuged again in resuspended in dH2O and 15mM mercaptoethanol. The sample was then purified using ion exchange chromatography and left to dialyse overnight in dialysis buffer containing sodium fluoride. The sample was then analysed with circular dichroism (CD) using an Applied Photophysics Chirascan machine. The CD spectrum was obtained at a temperature of 25°C, wavelength range between 190 to 260 nm, at a 0.5nm

intervals, with a bandwidth of 1 nm and at 1 second per time point. Protein purity was assessed using MALDI–TOF–MS analysis.

9. Experimental data analysis

The theoretical Cα proton chemical shifts were calculated for each trajectory, sampled at every 10 ns, using SPARTA+ (Shen & Bax, 2010). Scatter Biosis software (Rambo, 2017) was used to calculate the theoretical SAXS intensities of the simulation representative structures, for comparison with the experimental data. The experimental SAXS data was analysed using GNOM and PRIMUS, part of a suite of programmes developed for small angle scattering data analysis (ATSAS data analysis software) (Konarev et al., 2003; Petoukhov et al., 2012; Svergun, 1992). The circular dichroism (CD) protein spectra analysis was performed using the Dichroweb tool (Whitmore & Wallace, 2004; Whitmore & Wallace, 2008), the SELCON 3 method (Sreerama & Woody, 2000), a scale factor of 1.5 and the reference set 7, appropriate for IDPs.

Chapter I – MD parameterisation for IDPs

1. Introduction

The proposal that IDPs can be successfully studied using MD simulation is heavily reliant on the accuracy of the calculated properties, and on the correctness of the simulation parameterisation. Amongst the most well-studied and popular force fields packages, AMBER, GROMOS and CHARMM are typically used to set up the MD simulations in molecular biology. However, commonly used force fields and solvation models are known to present with different performance biases (**Table 1**).

Table 1 – Performance comparison of commonly used force fields, and solvation methods, in the structural characterisation of different protein systems.

Force field	Package	Solvation	Performance bias	Reference
ff99	AMBER	TIP3P	Overestimates *α*-helical content.	(Hornak et al., 2006)
ff99SB	AMBER	TIP3P	Underestimates *α*-helical content but retains structure compactness.	(Best & Hummer, 2009)
ff99SB*-ILDN	AMBER	TIP3P	Produces high structure compactness displaying increased number of intrapeptide hydrogen bonding. Underestimates Rg values for IDPs. IDPs secondary structure is inconsistent with experimental values.	(Robustelli, Piana & Shaw, 2018) (Rauscher et al., 2015)
ff99SB-ILDN	AMBER	TIP4P-D	Severe destabilisation of folded proteins. Substantial underestimation of helical content for IDPs.	(Robustelli, Piana & Shaw, 2018)
ff03ws	AMBER	Modified TIP4P/2005 interactions	Severe protein destabilisation for folded systems. For IDPs it is inaccurate in describing its secondary structure content.	(Robustelli, Piana & Shaw, 2018)
ff99SB-UCB	AMBER	TIP4P-Ew with modifications	Structural deviations leading to partial or complete unfolding of ordered proteins. Considerable helicity underestimation for IDPs.	(Robustelli, Piana & Shaw, 2018)
C22*	CHARMM	TIP4P-D	C22* with TIP4P-D produces ensembles that are too expanded when compared to experimentally determined Rg.	(Rauscher et al., 2015)
C22*	CHARMM	TIP3P-CHARMM	Stable for < 60 residue proteins but leads to structural instability for larger folded proteins. Unsuitable to study IDPs by severe underestimation of Rg values and inconsistent secondary structure formation.	(Robustelli, Piana & Shaw, 2018) (Rauscher et al., 2015)
C36m	CHARMM	TIP3P-CHARMM	It overestimates Rg of small proteins of < 60 amino acids. Unstable when simulating ordered systems of > 60 residues. Produces overly collapsed IDPs with inconsistent secondary structure,	(Robustelli, Piana & Shaw, 2018) (Henriques, et al., 2018)
C36	CHARMM	TIP3P and TIP3P-CHARMM	It displays a bias towards long left-handed *α*-helices, which should be absent from structured proteins.	(Rauscher et al., 2015)
G54A7	GROMOS	SPC	Produces collapsed protein structures, biased towards displaying unreasonably high helical content.	(Henriques, Cragnell & Skepö, 2015)
G53A6	GROMOS	SPC	Simulated proteins are too collapsed, when compared to SAXS experimental averages. Heavy bias towards β-sheet content with higher prediction of β-hairpins.	(Henriques, Cragnell & Skepö, 2015) (Sun, Qian & Wei, 2016)

Whilst the widely used AMBER and CHARMM force field refinements have been found to precisely describe small globular proteins (Beauchamp et al., 2012), the older AMBER ff99 and CHARMM22/CMAP both tend to overemphasise helical content; whereas the more recent AMBER ff99SB underestimates it; and GROMOS96 displays a heavy bias towards the creation of β-sheet structures (Chong, Chatterjee & Ham, 2017; Rauscher et al., 2015). Other modified, more recent iterations of the standard force fields, such as AMBER ff99SB-ILDN, AMBER ff99SBNMR1-ILDN, GROMOS 53A6 and GROMOS 54A7 have proved to be equally unsuitable by simulating excessively collapsed proteins, failing to emulate the structural diversity associated with IDPs, and/or exhibiting considerable bias towards folded secondary structure motifs (Henriques, Joao, Cragnell & Skepö, 2015). The recent CHARMM36 force field displayed a marked bias towards left-handed α-helix oversampling (Rauscher et al., 2015), which was addressed by their latest iteration for IDP simulation - CHARMM36m (Huang, Jing et al., 2017). Nevertheless, even CHARMM36m has been subsequently found to display biases towards secondary structure motifs inconsistent with experimental data, and overly collapse the protein structures (Robustelli, Piana & Shaw, 2018). As for AMBER, even its latest force field release, ff14SB, touted to have improved the sampling accuracy of backbone and sidechains, failed when applied to IDPs. AMBER ff14SB was found to create excessively hydrophobic and overly folded structures, which display extravagant α-helices and/or β-sheet formations (Best, 2017; Piana, Klepeis & Shaw, 2014).

To correct the biases and limitations of conventional MD simulations parameters two main ways have been proposed: (1) re-design the simulation models by optimising the force fields or (2) improve the accuracy of the simulation by enhancing the solvation conditions (Piana et al., 2015).

1.1 Optimising force fields for IDP simulations

In 2017, two novel AMBER force field optimisations for IDP simulation were published: ff14IDPs (Song et al., 2017) and ff14IDPSFF (Song, Luo & Chen, 2017). These constitute re-

parameterisations of the conventional AMBER ff14SB and maintain its the main features. The AMBER ff14SB uses the following molecular mechanics (MM) potential energy function:

$$E_{total} = \sum_{\text{Bonds}} K_r\left(r - r_{eq}\right)^2 + \sum_{\text{Angles}} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{\text{Dihedrals}} \frac{V_n}{2}[1 + cos(n\emptyset - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right]$$ (Cornell et al., 1995)

Which can be summarised as:

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{non-bonded}}$$

In which:

$$E_{\text{bonded}} = E_{\text{dihedrals}} + E_{\text{angles}} + E_{\text{bonds}}$$

and

$$E_{\text{non-bonded}} = E_{\text{electrostatics}} + E_{\text{van-der-Waals}}$$

The two force field modifications implement a grid-based energy correction map (CMAP) method to optimise the dihedral energy terms (Song et al., 2017; Song, Luo & Chen, 2017), leading to following optimised energy function:

$$E_{\text{total}} = E_{\text{ff14SB}} + E_{\text{CMAP}}$$

The difference between the two force field optimisations is based on the scope of the CMAP application. The ff14IDPs force field iteration improves IDP sampling by modifying the *φ/ψ* distributions of the 8 disorder-promoting amino acids (G, A, S, P, R, Q, E, and K). These modifications were based on the statistical assessment of a total of 17 540 IDP structures obtained from the PDB database, which contained 54 838 coil fragments and a total of 346 335 pairs of backbone dihedrals. The second IDP-optimised force field modification - ff14IDPSFF expands the application of the backbone dihedral terms upgrade to all 20 naturally occurring amino acids, since IDPs contain both order and disorder-promoting residues.

A recent computational study used the ff14IDPSFF force field iteration to assess the order-disorder transition of inducible transactivation domain (KID) (Liu, Hao et al., 2018). It demonstrated that ff14IDPSFF force field might be a valid and accurate option to parameterise IDP MD simulations.

1.2 Enhancing the water model

IDPs are very susceptible to the solvation model used to generate the simulation environment due to its large solvent-exposed surfaces, making them highly responsive to the protein–water interaction forces. Incorrect solvation potentials have been recognized as the main source of inaccurate, overly stabilised, or fragmentary IDP conformational description (Levine & Shea, 2017). In the simulations of globular proteins, the conventional solvation models create water molecules with 3-site rigid pair potentials with charges and Lennard-Jones parameters assigned to each atom (Mark & Nilsson, 2001). This type of solvation, including the TIP3P and SPC water models, has been identified as major contributor to deficient IDP sampling (Best, Zheng & Mittal, 2014). Recently, a new solvation model has been developed to address this limitation - (Piana et al., 2015)'s TIP4P-D, a 4-site water model, with a similar geometry to the older TIP4P/2005 model, but with improved charges and Lennard-Jones parameters (**Table 2**).

Table 2 – Comparison of the physical properties of the commonly used solvation models. Adapted from Piana et al. (2015).

	TIP3P	**SPC/E**	**TIP4P-EW**	**TIP4P/2005**	**TIP4P-D**
μ **(D)**	2.35	2.35	2.32	2.305	2.403
hydrogen charge	0.417	0.4238	0.52422	0.5564	0.58
C6 (kcal mol^{-1} Å^{6})	595	625	653	736	900
C12 (kcal mol^{-1} Å^{12})	582000	629482	656138	731380	904657
ΔHv **(kcal mol^{-1})**	10.2	10.5	10.6	11	11.3
C_p **(kcal mol^{-1} K^{-1})**	15.2	17.3	17.6	17.8	16.8
T_{md} **(K)**	250	270	275	270	
ρ **(T_{md})(g cm^{-3})**		1.012	1	1.001	0.997
γ **(mN m^{-1})**	47.8	58.4	59.2	63.3	71.2
α_V **(10^{-4} K^{-1})**	8.5	4.7	3.1	2.7	2.6
D **(10^{-5} cm^2 s^{-1})**	5.8	2.6	2.6	2.2	2.1
$\varepsilon 0$	96(3)	72(4)	63(3)	56(2)	68(2)

The TIP4P-D increases by 50% the TIP3P R^6 dispersion coefficient, with the R^{12} parameters adjusted accordingly. These modifications optimise the potentials for dispersion interactions, reduce the protein's overall tendency for intramolecular interactions in support of protein–water interactions, thus counteracting common biases towards overestimation of intraprotein and protein–protein interactions displayed by the standard MD simulations (Henriques, João & Skepö, 2016). The TIP4P-D water model is found by Henriques & Skepö (2016) to expand the conformational diversity of the disordered states for small peptides, in agreement with experimental SAXS data. However, explicit water models such as the TIP4P-D, in which water molecules are explicitly defined to simulate the aqueous medium, are computationally costly especially for large systems. Hence, exploring the feasibility of implicit solvation methods is also worth pursuing.

Implicit solvent models represent their solvation free energy as a continuum of electrostatic approximation. At each simulation point, the solvation potential of the system is re-computed based only on the degrees of freedom of the solute's coordinates and the solvation environment instantaneously adjusts to the new solute conformation (Onufriev & Case, 2019). In this type of solvation, because there are no explicitly created water molecules, the number of atoms in the system is considerably reduced, creating a very efficient method of simulation parameterisation in terms of time and computational cost. There is no need for lengthy water equilibration steps, no water box constraints, no artifacts caused by periodic boundary conditions clashes and improved protein sampling due to lack of viscosity (Onufriev & Case, 2019). The AMBER package for MD simulations offers different 'flavours' of the Generalised Born (GB) implicit solvent model. The GB method calculates the total energy of the molecule by decomposing its electrostatic and non-electrostatic potentials:

$$\Delta G solv = \Delta G el + \Delta G nonel$$

The ΔG*nonel* value is derived from the energy required to solvate the molecule with the charges removed. It is calculated from the favourable van der Waals interactions between

solvent and solute and the unfavourable cost of disrupting the solvent around the solute. The ΔG_{el} value is calculated by removing charges in a vacuum and then adding them back in a continuum solvent environment (Case et al., 2016).

Despite the clear advantages, and a raise in research interest, continuum implicit solvents have been neglected as viable parameterisation methods for MD simulations. This is mostly out of fear that implicit solvation might improve simulation speed at the cost of biological realism and compromise the accuracy of the simulation (Beauchamp et al., 2012). However, given its multiple advantages - especially the considerable reduction in computational cost - makes it an appealing alternative worth testing against experimental data.

2. Results and discussion

2.1 Using Histatin 5 as the model protein

Histatin 5 is a 24-amino acid human salivary protein, known for its antimicrobial and antifungal role. It was chosen as a preliminary model due to its small size, which allows for efficient MD sampling. It has a completely unstructured configuration in solution which has been experimentally characterised with small-angle x-ray scattering (SAXS) (Cragnell et al., 2016). To compare the Histatin 5 SAXS data to the results obtained from the simulations, noise reduction clustering was first performed, to address the complexity of the simulation trajectory, using the k-means algorithm. This revealed the average structures for the most abundant states sampled during the simulation. **Table 3** includes the cluster representative structures' radius of gyration (Rg) for each force field.

Table 3 – Comparison of the radius of gyration for each the representative structures, per cluster and simulated force field condition, with the experimentally-determined Histatin 5 radius of gyration.

Force fields	Cluster 1	Cluster 2	Experimental
ff14SB	9.15 Å	7.71 Å	13.8 Å
ff14IDPs	7.38 Å	8.15 Å	
ff14IDPSFF	7.48 Å	9.87 Å	

The comparison between the simulation representative structures and the SAXS-determined Rg affords invaluable insight reveal how poorly the clusters of simulation conformations agree with the experimental data, for any of the tested force fields. To highlight this further, **Figure 7 (a)** presents the Kratky plot comparison between the experimental SAXS data and the most extended simulation cluster centroid structures for each force field tested. Upon assessment it is evident that none of the force fields creates structures consistent with the experimental data, producing instead collapsed and overly folded Histatin 5 configurations. Furthermore, the modified ff14IDPs force field variant performs significantly worse than the conventionally used ff14SB, with the ff14IDPSFF only marginally outperforming it.

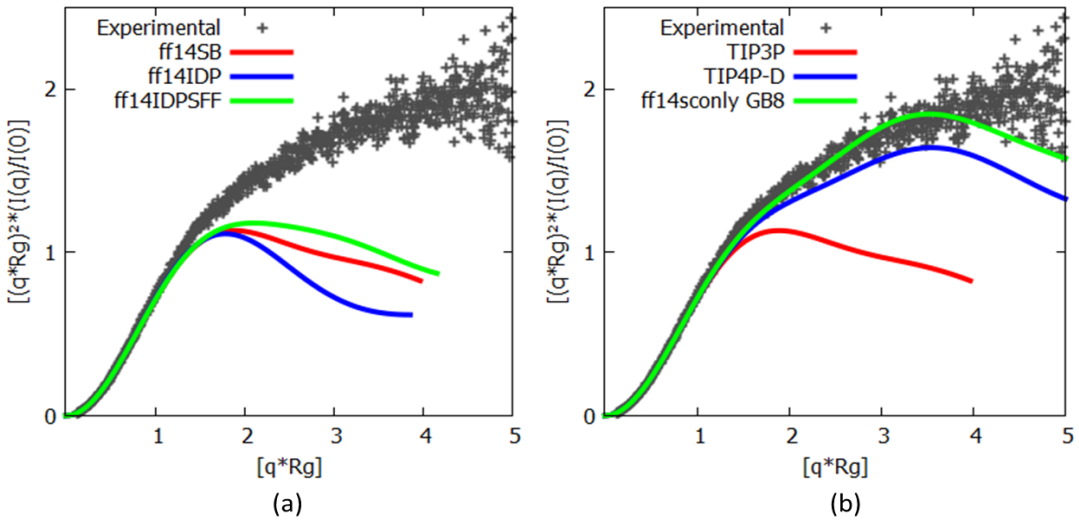

Figure 7 – Kratky plots comparing the experimental small-angle X-ray scattering (SAXS) data (in gray) to representative structures obtained from each of the (a) simulated force field conditions: ff14SB and the modified ff14IDPs and ff14IDPSFF force fields and (b) solvation conditions: TIP3P, TIP4P-D and the implicit solvent GB8.

With the different force field iterations failing to adequately describe Histatin 5, attention turns towards the water models, to have its performance benchmarked against the experimental

data. **Table 4** contains the Rg values for each cluster centroid structure and the experimentally derived Rg.

Table 4 – Comparison of the radius of gyration for each the representative structures, per cluster and simulated water model condition, with the experimentally-determined Histatin 5 radius of gyration.

Water model	Cluster 1	Cluster 2	Experimental
TIP3P	9.15 Å	7.71 Å	13.8 Å
TIP4P-D	13.47 Å	12.12 Å	
Implicit GB8	10.68 Å	14.14 Å	

Table 4 results demonstrate that the modified TIP4P-D water model and the implicit solvation based on Generalised Born (GB) 8 create average structures closer to the experimentally determined Rg value, especially when compared to the conventional TIP3P solvation. It is interesting to note that for the simulations solvated with the TIP4P-D method, the two most abundant clusters consist of extended conformations, whereas the implicitly solvated GB8 simulations creates conformations with a wider range of structure compactness, oscillating between averages of 10.68 Å and 14.14 Å. **Figure 7 (b)** reveals the Kratky experimental data plotted against the most extended representative structures for each of the three water models tested: the conventional TIP3P, the optimised TIP4P-D and the implicit solvent model GB8. **Figure 7 (b)** further emphasizes the close agreement between the optimised water models (TIP4P-D and GB8) and the experimental data, especially when comparing it to the TIP3P result. Of remarkable consistency with the experimental values is the structure derived from the GB8 simulation, which accurately agrees the experimentally determined Kratky curve.

Additionally, when considering available Histatin 5 NMR data (**Figure 8**), the results show a similar conclusion – both TIP4P-D, and especially GB8, match the experimental HA chemical shifts accurately, whilst TIP3P does not. The GB8 solvation solution displays a RMSE score of 0.12 ppm against the experimental data, closely followed by TIP4P-D with a RMSE of 0.14. Comparatively, TIP3P achieved a RMSE of 0.21 ppm, which collectively with its *p-value* of 0.0208 highlights its unsuitability to solvate disordered systems simulations.

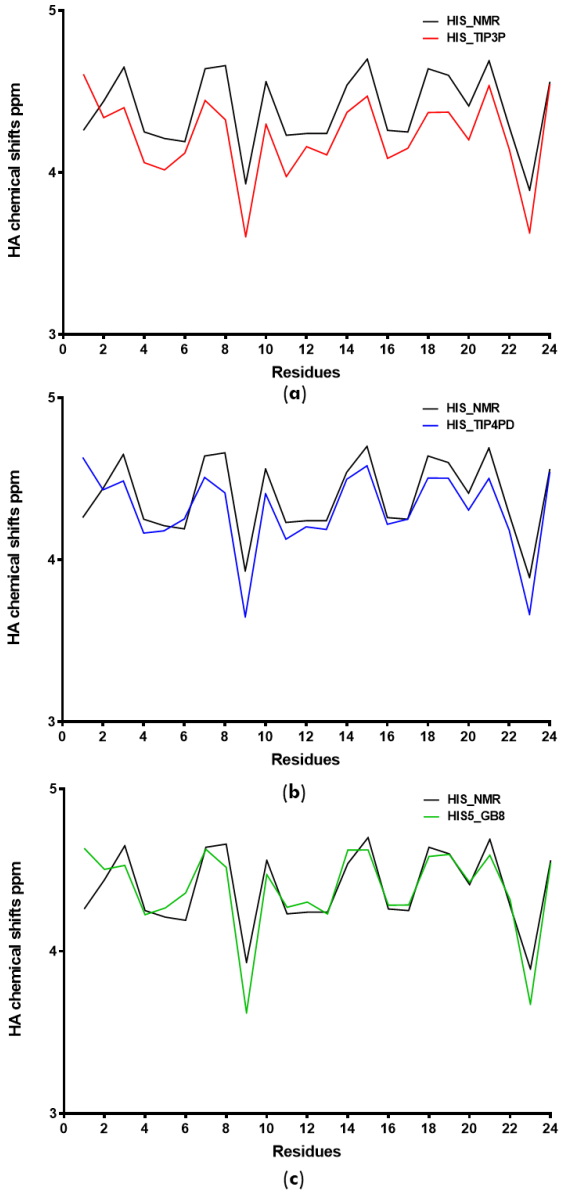

Figure 8 – Comparison of NMR-determined HA chemical shifts to calculated chemical shifts for the simulation trajectories (a) using the TIP3P solvation method (*p-value* 0.0208*), (b) using the TIP4P-D water model (*p-value* 0.1387) and (c) those obtained from the simulations using the implicit GB8 solvation method (*p-value* 0.9874).

Overall, it is clear from the preliminary tests with Histatin 5 that the adapted water models offer a solid performance in terms of recreating the conformational nature of a fully disordered protein, outperforming any of the tested force field modifications. However, further assessment is required into how well these water models perform when simulating larger protein systems. Therefore, the different solvation methods were benchmarked against NMR-derived secondary structure propensities, using MYC88 as the protein model.

2.2 Water model testing using MYC88 as the model protein

MYC88, encompassing the first 88 c-MYC amino acids, contains within it two highly conserved regions - MB0 and MBI. (Andresen et al., 2012)'s NMR study of the protein offers a glimpse into its secondary structure propensities (SSP) per residue (**Figure 9 (a)**).

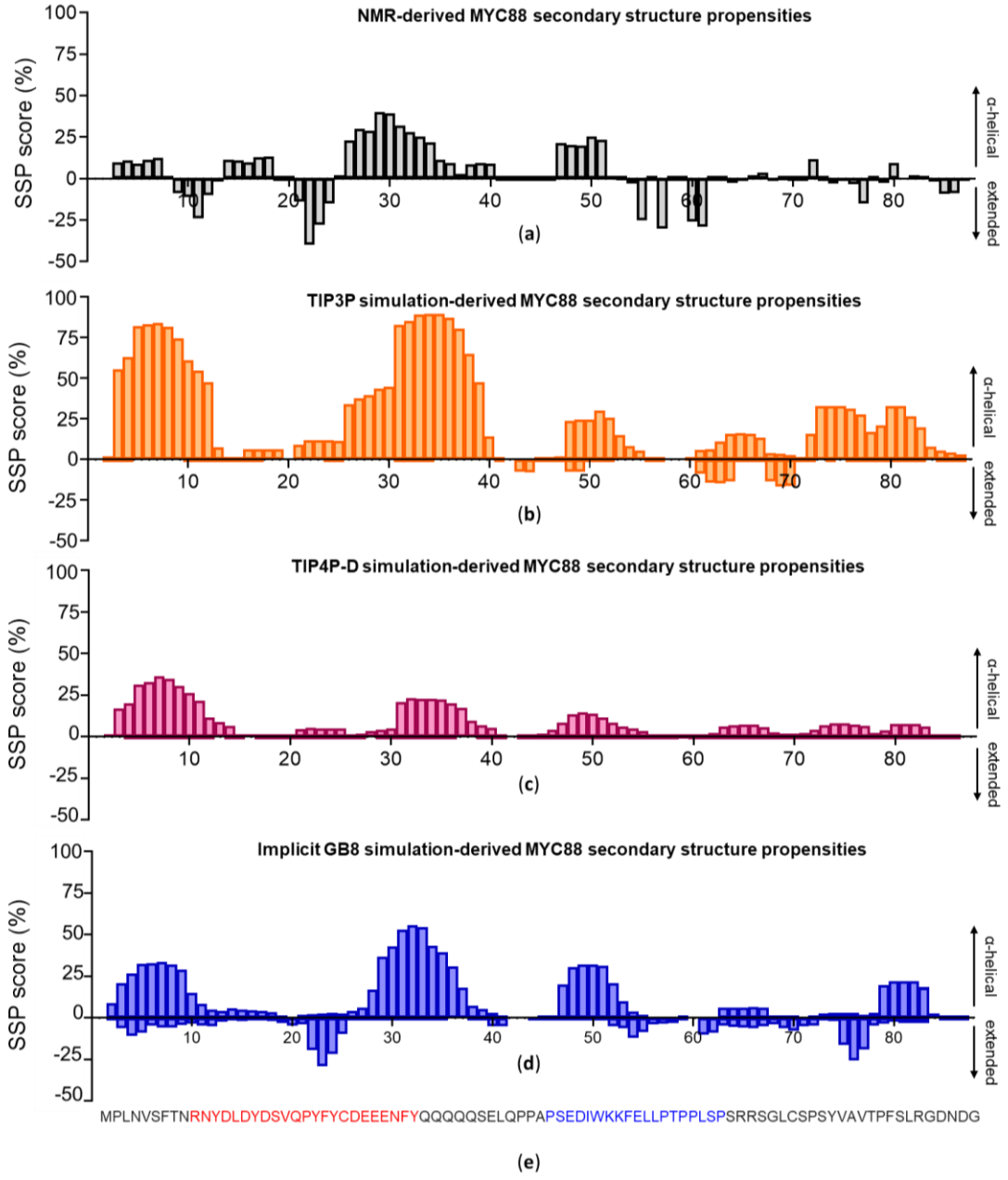

Figure 9 – Comparison of (a) NMR-determined transient secondary structure propensities of c-MYC1-88 with (b) those obtained from the molecular dynamics simulations using the TIP3P solvation method, (c) those obtained from simulations using the TIP4P-D water model and (d) those obtained from the simulations using the implicit generalized Born (GB8) solvation method. The positive values on the Y-axis correspond to regions with a tendency to form α-helices, whilst the negative Y-axis values reflect regions with propensities towards extended structure formation. (e) Sequence of MYC88 with MYC-boxes MB0 coloured in red and MBI in blue).

Andresen et al. (2012) found based on combined SSP and NOE assessment four main regions displaying transient ordered structure formation: a β-turn at residues 21 to 24; a transiently helical region comprised of residues 26 to 34; another important helical region between residue 47 and 54; followed by an extended region from residue 55 to 65. Andresen et al. (2012)'s findings can be compared to the averaged secondary structure propensity for each residue, calculated over the course of the entire MD trajectory. This avoids any potential bias introduced by the clustering and the exclusion of rarer states. **Figure 9 (b)** displays the SSP derived from the simulations solvated with the TIP3P water model. It is clear, when studying the ranges of the Y-axis for **Figure 9 (a)** and **(b)**, how biased the TIP3P system is towards heavily helical formations, especially considering the first 40 residues. For the regions between residues 1 to 12 and between residues 26 to 40 the incidence of helical motifs for the TIP3P simulations (**Figure 9 (b)**) reaches over 75%. The experimental data (**Figure 9 (a)**) indicates that no MYC88 region is predicted to display helical propensity above 50%, emphasizing the discrepancy between the simulation and the experimental results. The extended regions, particularly residues 9 to 12 and 22 to 25, predicted by NMR are also absent from the TIP3P simulation. This is consistent with the results obtained from the Histatin 5 simulations, which show that the conventional TIP3P solvation method produces overly ordered and compact structures with a severe bias towards helical motifs and is incompatible with experimental data.

Figure 9 (c) shows the average SSP data obtained from the simulations parameterised with the optimised TIP4P-D water model. One of the main findings when considering the TIP4P-D SSP distribution histogram is that this solvation method dramatically reduces the development of any ordered secondary structure. Ostensibly, the TIP4P-D simulations demonstrate a 50% reduction in the propensity for helical formations, but also completely abrogate the formation of predicted β-sheet extended motifs, which is inconsistent with experiment. Lastly, **Figure 9 (d)** displays the data obtained from the implicitly solvated simulations. The results obtained with the implicit GB8 solvation method display the highest

conformity with the experimental data (**Figure 9 (a)**). Notably, the implicitly solvated simulation consistently simulates the β-sheet extended regions between residue 21 to 25 and the α-helical regions formed by the residues 26 to 34 and 47 to 54, which are predicted by experiment. Although it creates slight overly ordered loci at the two terminal regions, it otherwise displays noteworthy consistency with the experimentally determined SSP. It should also be considered that the MYC88 NMR N-terminal oligo-histidine tag might have slightly affected the experimental results of the surrounding N-terminal helices (Andresen et al., 2012). Thus, the overall conclusion is sustained by the total values for helical and extended β-sheet content which demonstrates that both explicit water models replicate the helical content well, but only GB8 recreates the extended content predicted by experiment (**Suppl. Figure S1**).

2.3 Further testing using MYC150 as the protein model.

MYC150, spanning residues 1 to 150, contains the entire MYC88 TAD domain and its first three highly conserved regions: MB0, MBI and MBII. Since neither the modified force fields nor the TIP3P solvation model produced adequate protein structures, these conditions are discarded in favour of further assessing TIP4P-D and the implicit solvation GB8 method, which showed more promise. Circular dichroism (CD) was used to analyse MYC150 structure by estimating its helical, β -sheets and random coil ratios, to be compared with the MD simulations data. The MYC150 experimental data is directly derived from the CD spectra, whilst the secondary structure percentages from simulation are obtained by averaging the total values, for each type of secondary structure content, over the entire simulation. **Table 5** contains the secondary structure percentage for both the experimental data and the two simulation conditions: TIP4P-D and GB8 and **Figure 10** offers the graphical representation of the same data.

Table 5 – Comparison of the secondary structure ratio for the TIP4P-D, GB8 MYC150 simulation with the experimentally determined secondary structure ratio from CD analysis.

Condition	Helical (%)	β-sheets (%)	Random coil (%)
Experimental	32.5	7.2	59.5
TIP4P-D	41.6	1.26	57.08
Implicit GB8	40.83	7.56	51.61

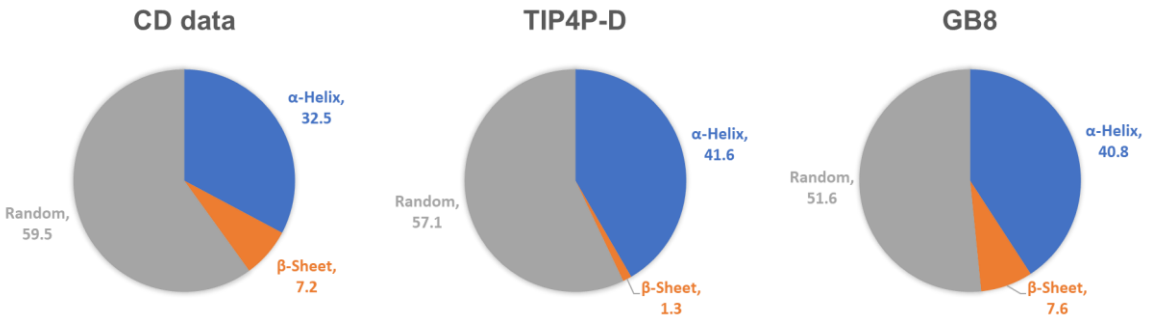

Figure 10 – Comparison of MYC150 CD-determined secondary structure ratio (in percentage) to those obtained from the TIP4P-D and the implicit GB8 solvation simulations (also in percentage).

It is clear from the results that the comparison of the simulation data to CD experiment is consistent with the earlier SAXS and NMR findings. The TIP4P-D solvation method fails to adequately describe the protein's β-sheet extended structure content despite closely emulating the MYC150's random coil score. Overall, TIP4P-D is undoubtedly a better option when compared to the standard TIP3P solvation method. However, the implicit solvation GB8 method despite marginally underestimating the random coil and overestimating the helical content – a slight bias present also in the MYC88 simulation, displays the protein description most consistent with the available experimental data. Given GB8 model's accuracy, when compared to SAXS, NMR and CD data using multiple protein models, and the advantages of the implicit water model in terms of speed and computational efficiency, makes this the prime solvation choice for c-MYC's *in silico* studies.

2.4 MYC88's GB8 MD and Markov-chain Monte Carlo simulation comparison.

The Markov-Chain Monte Carlo (MCMC) simulation produces a collection of samples from a dense stationary (π) target distribution. The Markov chain builds a detailed balanced equation in which new states are accepted or rejected based on the following probability:

$$\pi(x)P(x \rightarrow x') = \pi(x')P(x' \rightarrow x)$$

Which means that the probability of inhabiting state *x* multiplied by the probability of going from state *x* to state *x'*, is reversibly equal to the probability of moving from state x' to state *x*.

The next phase is to define two transition probability steps - the proposal probability and the acceptance-rejection probability:

$$P(x \rightarrow x') = Pp(x \rightarrow x')Pa(x \rightarrow x')$$

Where the proposal probability $Pp(x \rightarrow x')$ corresponds to the calculated probability of proposing a given state and the acceptance probability $Pa(x \rightarrow x')$ corresponds to the calculated probability of accepting the new state or rejecting it.

There are many Bayesian inference algorithms that can be implemented to practically solve for the acceptance probability equation, but of all the Metropolis-Hastings is undoubtedly the most common and well-researched (Boomsma et al., 2013). It takes the previous steps into account and aims:

$$P(x \rightarrow x') = min\left(1, \frac{\pi(x')Pp(x' \rightarrow x)}{\pi(x)Pp(x \rightarrow x')}\right)$$

Which entails, assuming unbiased transitions, that the algorithm fully accepts the new state if its probability is increased according to the target distribution (π (x') > π (x)). If the probability is lower, the acceptance will depend on how unfavourable the new conformation is. This ensures harmonious sampling and a congruent probability distribution in which the structures are sampled according to their conformation favourability. Thus, comparing the balanced and well-sampled conformational distribution approximated by the MCMC simulations to the conformational landscape derived from the MD simulations allows for insight into how extensively the MD protein's conformational ensemble is being described (Sullivan & Weinzierl, 2020).

To compare the conformational landscapes derived from MCMC and MD simulations, they are both firstly defined in terms of the RMSD and R_g of the structural ensembles generated. This allows for assessment of the compactness, flexibility, and conformational divergence of the MYC88 data. The RMSD was calculated against a completely extended

structure as a reference. Therefore, the highest RMSD and low R_g values correspond to structures in folded states and, conversely, the lower *RMSD* and high R_g values correspond to the most extended structures. The MCMC landscape and *RMSD* frequency distribution plot indicate that the most probable states, and most well-sampled conformations, occupy the highest *RMSD*. This correlates with the lowest R_g values, between 13 and 18 Å, hinting at c-MYC88 preferentially inhabiting a series of relatively compact states, whereas the MCMC probability landscape also identifies a wealth of extended MYC88 states (**Figure 11**).

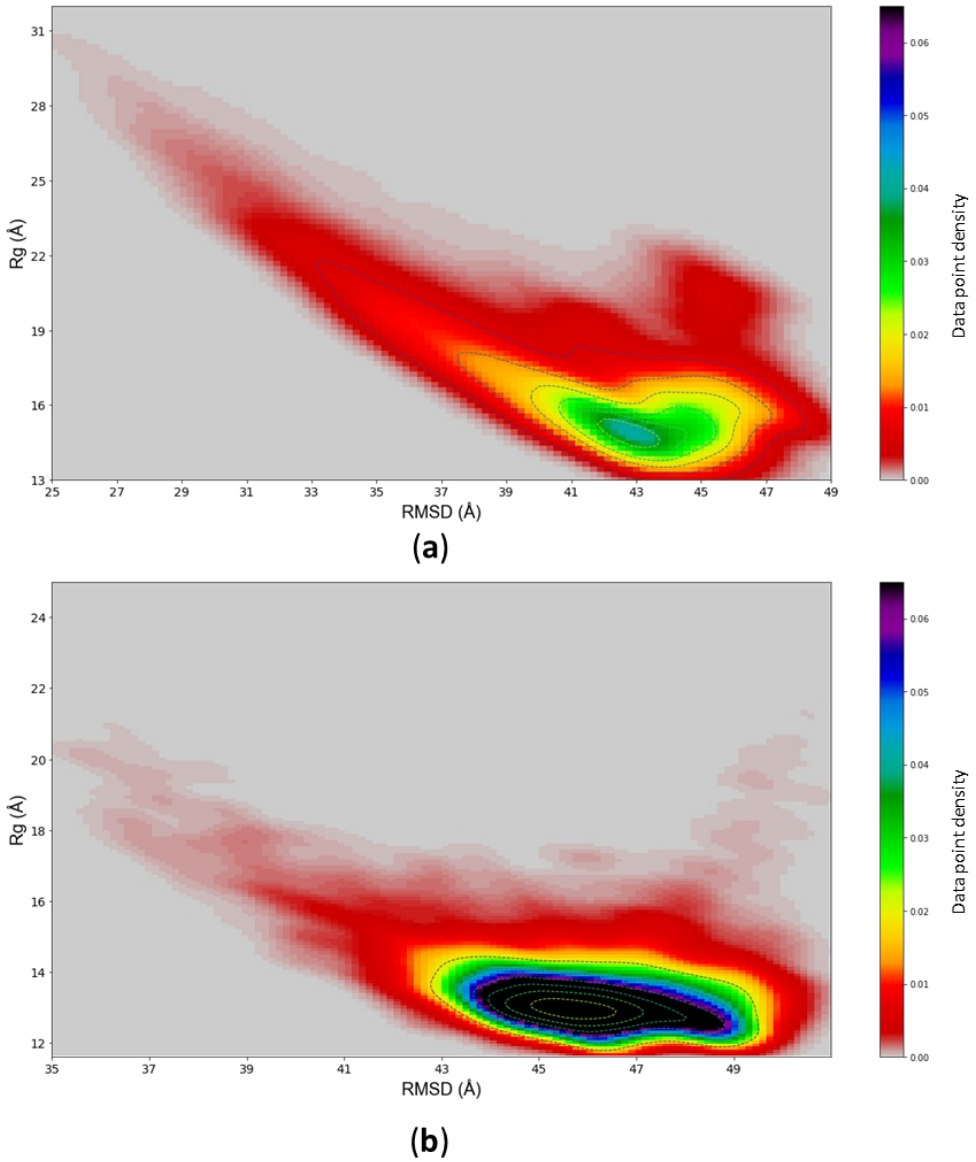

Figure 11 – (**a**) MCMC conformational landscape of MYC88 defined in terms of its RMSD and R_g. (**b**) The same for MD GB8 simulations.

The rationale for creating such a landscape is not to directly assess the protein dynamics of MYC88 from it - MCMC simulations do not reflect a temporal progression of a system, but rather give an overview of the range of possible conformational spaces available to the protein. A comparison of the MCMC and MD conformational landscapes results calculated in the same manner shows that the two methods create structures that inhabit an extensively overlapping landscape, especially when referring to the most compact conformations. Furthermore, using Sα—a metric of α-helical content similarity—as a conformational descriptor against the Rg, allows for investigation into the helical sampling of the MD landscape; the MD simulation explores a wide range of helical content, similar to the range explored by the MCMC landscape (**Figure 12**).

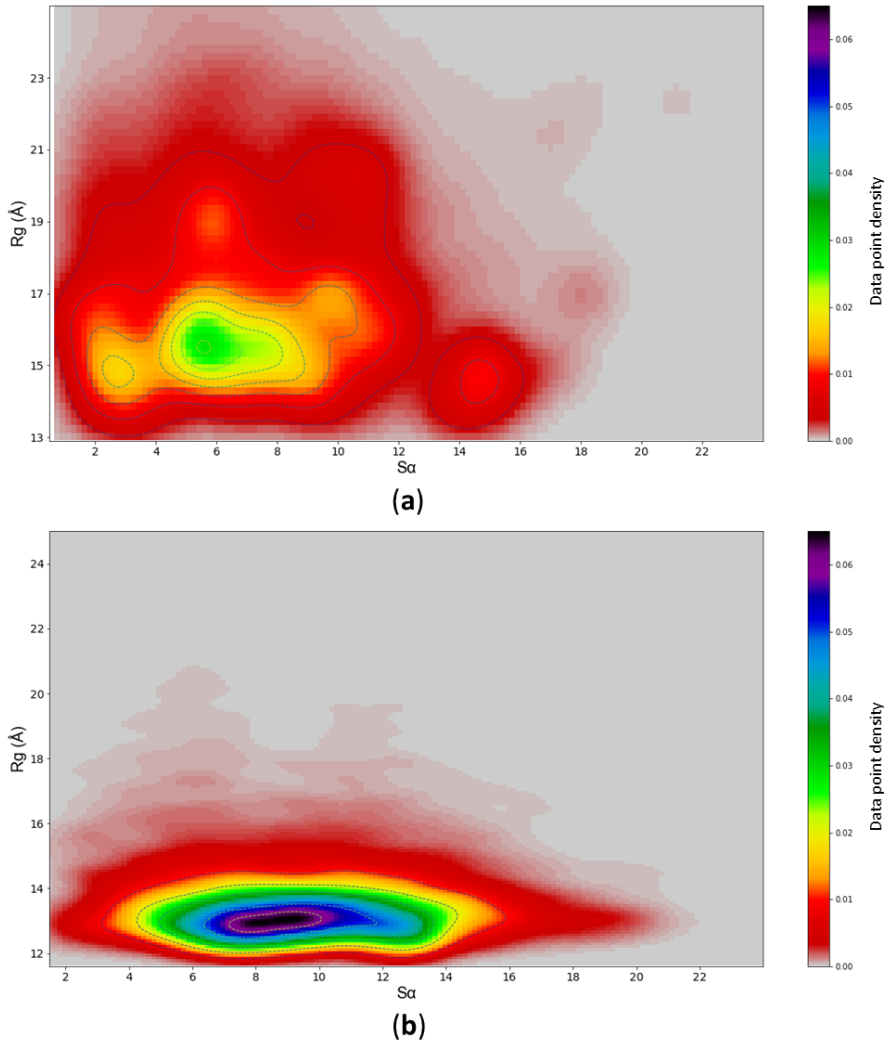

Figure 12 – (**a**) MCMC conformational landscape of MYC88 defined in terms of its Sα and Rg. (**b**) The same for MD GB8 simulations.

From these data, it is evident that the MD simulations do not become trapped in overly helical states, but sample within a wide basin of *Sα* values. Nevertheless, the MD simulations preferentially explore the lowest R_g states, which could be due to a variety of reasons. It is possible that the most extended states, as predicted by the MCMC simulation, are not very favourable energetically. Alternatively, the MD simulations may require more extensive sampling on a longer time scale. To assess the validity of these two hypotheses, MD simulations were repeated, starting with the coordinates from an extended, compacted structure or MCMC k-means clustering centroids as starting points. The results demonstrate that the choice of initial structures has no significant impact on the MD simulation conformational sampling range; all simulations converge on the same common space (**Figure 13**).

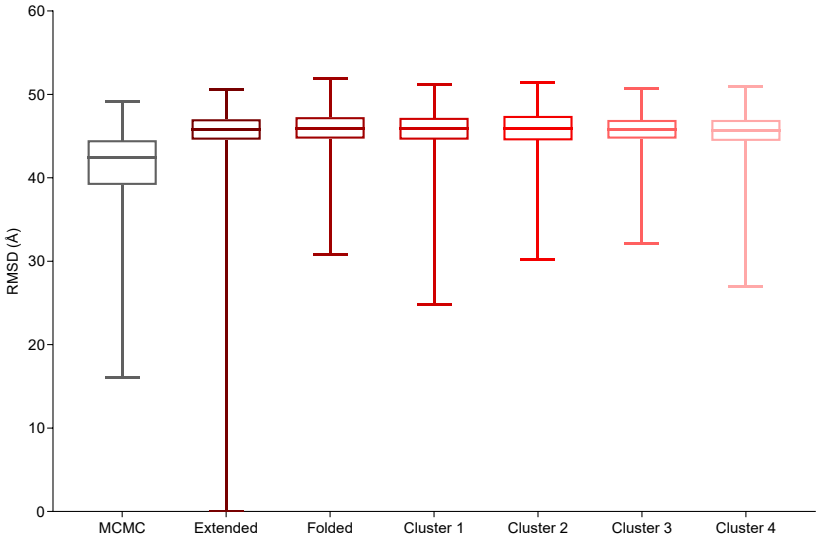

Figure 13 – Boxplots comparing the RMSD values obtained by the MCMC simulation and the MD simulations initiated with different starting structures: **Extended** corresponds to fully unstructured initial coordinates; **Folded** to a structure created with *ab initio* software; the cluster structures correspond to the different MCMC centroid structures obtained via K-means clustering.

Given that our finally selected simulation parameterization (ff14SB/GB8) agrees with experimentally determined secondary structure propensity—and the starting coordinates do not bias the simulations towards more compact structures—any differences between the MCMC and MD simulations are therefore more likely to be caused by the energy functions

used to calculate the structural properties, rather than any bias introduced by the starting point of the simulation.

3. Conclusion

Prior to drawing conclusions from computational simulation studies, it is necessary to ensure that they reflect, with high degree of accuracy, the structural features of the simulated systems. Accurate MD simulation results rely heavily on the parameters used to set up the simulation. It is known that the conventionally used force fields and solvation methods, although adequate to study globular proteins, introduce various structural biases which are incompatible with IDPs conformations. Here, several alternatives to the standard force field and solvation models were tested using Histatin 5, MYC88 and MYC150 as protein models. The modified force fields failed to adequately describe the IDPs structural characteristics when compared to experimental data, whilst the modified water models showed more promise. Surprisingly, the method displaying the highest degree of accuracy when compared to SAXS, NMR-derived and CD data was the Generalised Born 8 implicit solvation model. This parameterisation method not only produces accurate IDP structural determination, highly consistent with different experimental results, as it is also exceedingly efficient in protein sampling and computational productivity.

Being able to demonstrate that the GB8 MD simulation samples a varied conformational space instead of being trapped in some unimportant local minima is also crucial. Here a new method is proposed: comparing the MD simulation conformational ensemble to the structural landscape built from a MCMC simulation. The results show that the MCMC and MD simulations converge to the same conformational space determined by Rg and Sα metrics. The Sα results further demonstrate that MYC88 MD simulation samples a wide range of helical content, indicating that the GB8 parameterisation allows the molecule to sample diverse structural states and display suitable secondary structure content, consistent with experim ental data.

Chapter II – MYC88 structural dynamics

1. Introduction

The MD simulation data takes the form of a series of snapshots, created by moving the coordinates of the system, over time. Before gaining insight into the dynamic findings of MYC88's MD trajectory, methods that can extract the relevant features of such large, noisy datasets are urgently needed. This is especially true when dealing with IDPs, which create even noisier and more complex trajectory data.

The MD trajectory can be unpacked into a variety of basic geometrics, which includes the root mean square fluctuations (RMSF) of aligned structures, a metric that calculates the deviation between the position of aligned residues and is a measure of protein flexibility. Distance measurements are also routinely used in MD analysis to understand the dynamics between different regions, different proteins or between ligand and receptor. The calculation of dihedrals and torsions allows for enquiry into the study of conformers and conformational arrangements. Similarly, the study of bonds and contacts provides additional information regarding intra and intermolecular interactions. Other methods of analysis include the calculation of secondary structure propensities, which can be done over the course of the trajectory, or averaged per residues as seen in Chapter I.

The calculation of the MD trajectory features can be achieved using an array of different programmes commonly available, which include CPPTRAJ from the AMBER suite of MD programmes (Roe & Cheatham, 2013), MMTSB toolset from the CHARMM suite (Feig, Karanicolas & Brooks, 2004), GMX the toolbox from GROMACS suite of programmes (Abraham et al., 2015), BIO3D an R based analysis package (Grant et al., 2006), MDAnalysis (Gowers et al., Sep 11, 2016; Michaud-Agrawal et al., 2011) and MDTraj (McGibbon et al., 2015) packages that use Python Conda environment to deploy its algorithms. These constitute

some of the most well-known alternatives that are routinely used in research to perform the MD trajectory analysis. However, the calculation of simple geometric measures is frequently insufficient to describe, by itself, the protein's dynamics. The relevant protein states are often perfectly hidden in the complexity of the data (Shao et al., 2007). Therefore, noise reduction and dimensionality reduction methods are regularly used to sieve the data and reveal the underlying system motions. One of such methods is clustering, which was deployed previously as an analysis tool in Chapter I. Clustering entails grouping together protein conformations with a high degree of similarity by allocating the data points into separate sets called clusters (Shenkin & McDonald, 1994). The configurations in a cluster are structurally closer to each other than to structures from other clusters, allowing for a rapid description of the resulting conformational sets. The calculation of the cluster centroids, which correspond to the averaged representative structures, summarises the type of configurations inhabiting each set.

There is a wide variety of clustering methods and several algorithms are available in the Data Science arsenal. In Chapter I, the K-means algorithm, one of the most popular, was used (De Paris et al., 2015). The K-means clustering approach relies on two main decisions: the pre-determination of the total number of clusters (or the cluster radius) and the distance metric by which to assess similarity. In machine learning, K-means is an unsupervised learning algorithm and to group similar data points, it starts by initialising the centroids which constitute the centre of each cluster followed by the assignment of the remainder data points to their nearest centroid. The algorithm then repeatedly iterates through the data to refine the allocation of data points and stops when there are no new cluster reassignments (J. A. Hartigan & M. A. Wong, 1979). However, whilst K-means clustering is a great way to quickly summarise the data landscape, it is dependent on many factors that might introduce bias such the heuristic determination of the cluster (k) number. Additionally, noisy datasets, such as those produced by IDPs simulations, and small changes in cut off parameters makes the clustering unreliable and yield a 'unclusterable' landscape (Rajan, Freddolino & Schulten,

2010). Furthermore, the centroids and averaged molecular representations often do not accurately describe the IDP conformational range. Although helpful to give an overview of the different states of the molecule, clustering cannot give any insight into its structural transitions. These are several of the reasons why many researchers turn to Principal Component Analysis, or PCA, as a data reduction method to extract the important data features.

PCA is a multivariate statistical method that reduces the high-dimensional MD trajectory space to a smaller spatial scale. PCA applies a linear transform, to obtain the most important data elements, using a matrix created from the atomic coordinates that describe the system's main features. It assumes that each trajectory snapshot conformation comes from a well-sampled, equilibrated simulation hence capturing the protein's essential dynamics. It then decomposes the matrix and projects the data onto a set of eigenvectors (a principal component or PC) with a corresponding variance that reduces the system's degrees of freedom and explains the largest amplitude motions, typically corresponding to folding events (David & Jacobs, 2014). Whilst PCA is a robust method for feature extraction and noise reduction, it deals poorly with data that is not linearly correlated. This is because it uses a linear transformation based on covariance and projection of the data onto orthogonal eigenvectors, meaning that any not linearly related variable, very common in IDP simulations, will not be accurately described (David & Jacobs, 2014; Rajan, Freddolino & Schulten, 2010).

Ultimately, there are many methods that can be used in Data Science to analyse MD trajectories and some are well described and commonly deployed in literature. However, little is known about the performance of many of these methods in the analysis of IDP simulation data. This Chapter will address this question and aim to decipher MYC88's structural dynamics.

2. Results and Discussion

The first step in analysing the MD trajectory is to describe it in terms of its simple geometric calculations. **Figure 14** shows several descriptive landscapes obtained by plotting different

trajectory analysis metrics against the RMSD. Upon close examination, the plots suggest that the landscape does not contain any differentiated clusters and is quite homogenous, as would be expected of an IDP simulation.

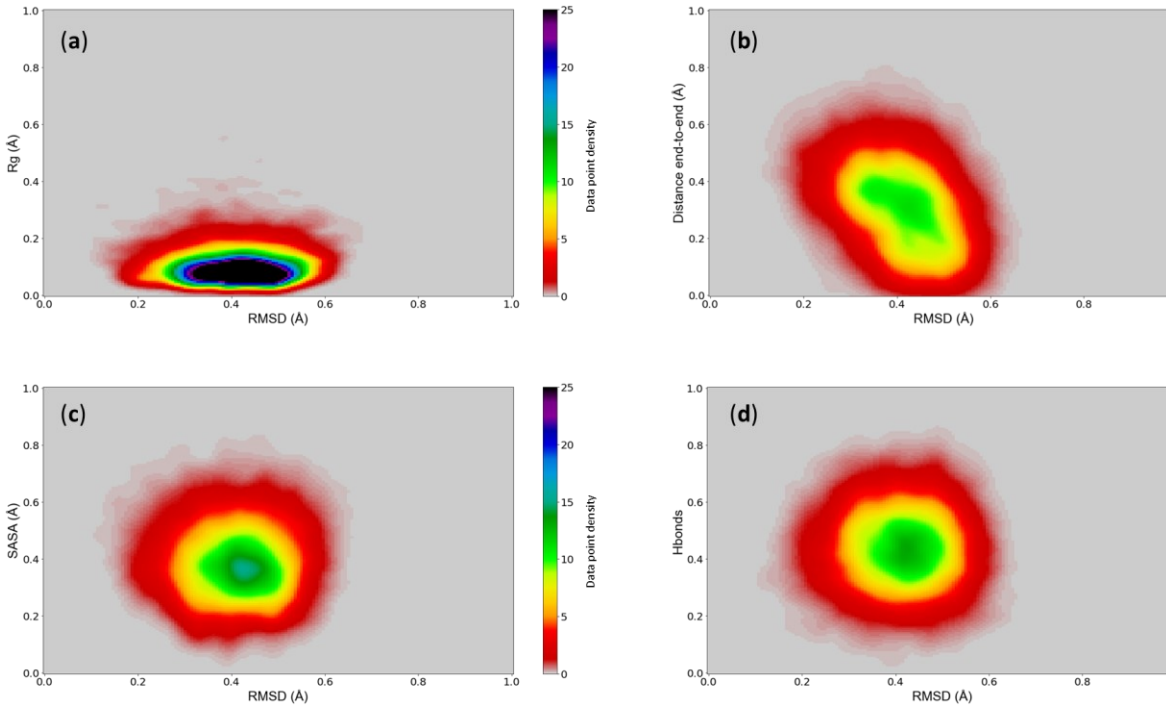

Figure 14 – Normalised MYC88 landscapes obtained by plotting RMSD values against different simple MD simulation metrics: (a) radius of gyration (Rg), (b) the molecule's distance from N-terminal to the C-terminal (Distance end-to-end), (c) solvent-accessible surface area (SASA) and (d) the number of hydrogen bonds formed (Hbonds).

The lack of discernible clusters makes it difficult to deploy clustering algorithms to gain insight into the different molecular macrostates, making direct clustering entirely unsuitable to define protein states. Therefore, analysis methods that can reduce the dimensional space and find the hidden messages in the data are necessary. PCA is the most likely candidate for the purpose. **Figure 15** depicts the PCA landscape obtained from MYC88's C-alpha atoms XYZ coordinates over the course of the trajectory.

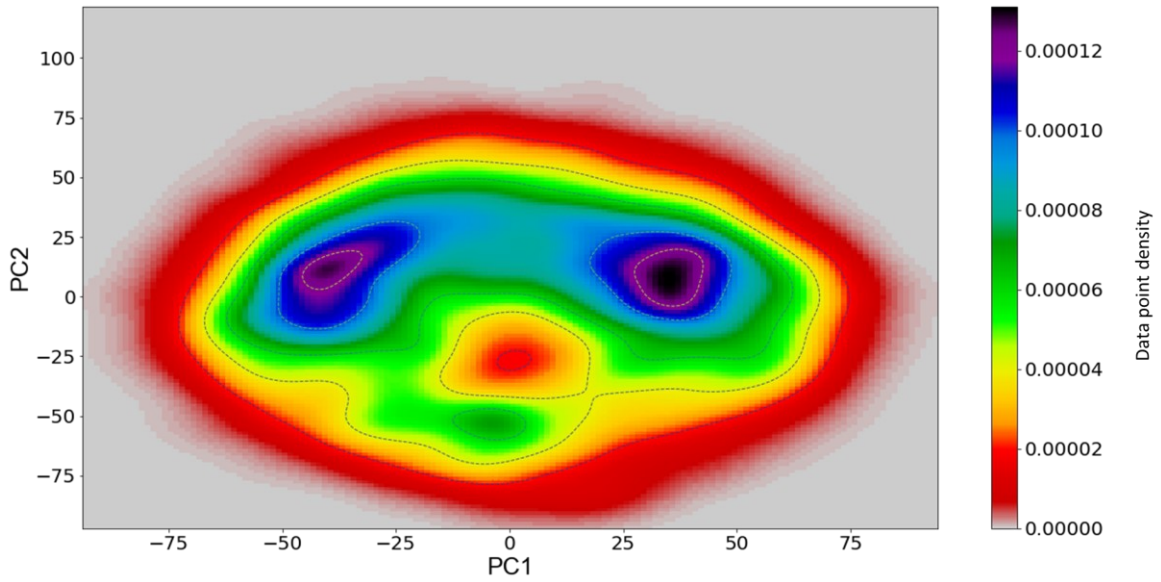

Figure 15 – PCA plot depicted with a kernel density estimation heatmap for easy detection of areas with high density of data points.

Upon inspection, the PCA plot reveals a slightly less noisy landscape when compared to the RMSD/Rg landscape and presents a potential separation of clusters, with two large clusters present and a third smaller one at the bottom. However, the problem with using the PCA to describe the protein's conformational space arises from the fact that the first two PC's only cumulatively explain ~22.5% of the data variance. (**Table 6**).

Table 6 – Eigenvalues, explained variance and cumulative explained variance for the first 6 principal components.

	Eigenvalue	Variance (%)	Cumulative (%)
PC 1	1273.756	12.328	12.328
PC 2	1042.172	10.087	22.415
PC 3	799.842	7.741	30.156
PC 4	719.519	6.964	37.120
PC 5	612.431	5.927	43.047
PC 6	513.945	4.974	48.022

Even when considering the first 6 PC's, it cumulatively explains less than 50% of the data which is clearly insufficient to derive any conclusions. When visually inspecting the highest-amplitude atomic displacement projected over the PC1, it is interesting to note that the protein

region with the highest atomic displacement exactly coincides with MBI – residues 43 to 63 (**Figure 16**).

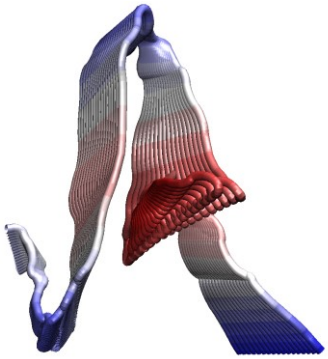

Figure 16 – Figure depicting the structures projected onto PC1. The colour scale highlights the regions with high atomic displacements (in red) and low atomic displacements (in blue).

However, this finding by itself it is unlikely to be very descriptive of an important protein motion given the low explanative power of the PCA vectors. Alternatively, internal coordinates such as backbone dihedral angles have been deemed as a viable option to resolve the PCA landscapes for flexible systems (Mu, Nguyen & Stock, 2005; Sittel, Jain & Stock, 2014). MYC88 the backbone dihedral PCA is plagued by the same issue – whilst it produces a free-energy landscape (FES) landscape with distinct clusters (**Figure 17**), the variance explained by each component is too low to be taken as a solution for dimensionality reduction (**Table 7**).

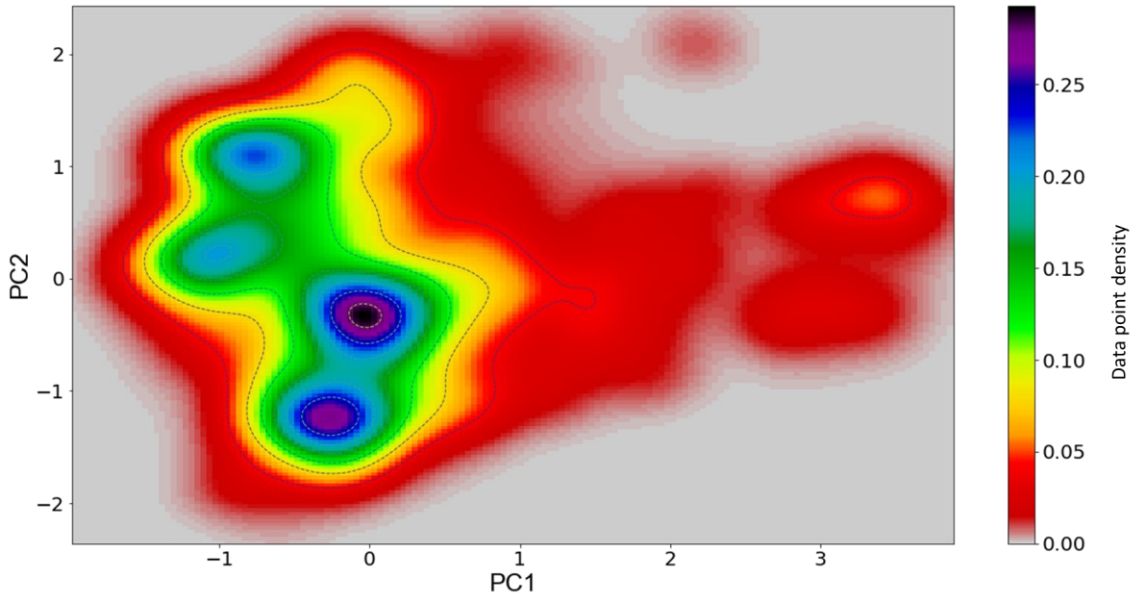

Figure 17 – Dihedral PCA plot for MYC88 MD simulation.

Table 7 – Dihedral PCA: Explained variance and cumulative explained variance for the first 6 principal components.

	Variance (%)	Cumulative (%)
PC 1	3.7	3.7
PC 2	2.9	6.6
PC 3	2.8	9.4
PC 4	2.3	11.7
PC 5	1.9	13.6
PC 6	1.9	15.5

Other metrics were also attempted - namely, PCA based on geometric descriptors (RMSD, Rg, distance end-to-end, SASA, and number of hydrogen bonds) (**suppl. Figure S2**); secondary structure content (calculated using the 'Define Secondary Structure of Proteins' or DSSP algorithm) (**suppl. Figure S3 and Table S1**); and pairwise distances between alpha-carbons (**suppl. Figure S4**). Unfortunately, PCA based on any of these features also failed to satisfactorily reduce the dimensionality of the conformational space by either producing landscapes without discernible clusters or failing to adequately explain the data variance. With the PCA being insufficient to define the MYC88's conformational landscape, another method is required to describe the structural identities hidden in the MYC88 trajectory.

Time-lagged independent component analysis (TICA) is another linear transformation method oriented towards finding coordinates of maximal correlation within a given time lag. The slowest motions, rather than maximal amplitude motions as with PCA, are tracked (Scherer et al., 2015). The main advantage of TICA over PCA is its lower dependence on the distance metrics since TICA is not so much concerned with the variance of atomic displacement but rather with the speed of temporal change. The speed is embedded into the process of structural change and is not so coordinate dependent. However, featurisation of data remains important should be carefully decided to minimise statistical error (Chodera & Noé, 2014).

With TICA analysis the data was projected over two dimensions, the first two independent components (IC), with a lag of 20 ns. The two IC's correspond to the two slowest and largest timescale transitions in the data. When plotting the two IC's directly as a free energy plot it immediately becomes apparent that TICA creates a clearer landscape with well-defined and separated minima basins, which are less prone to clustering errors (**Figure 18**).

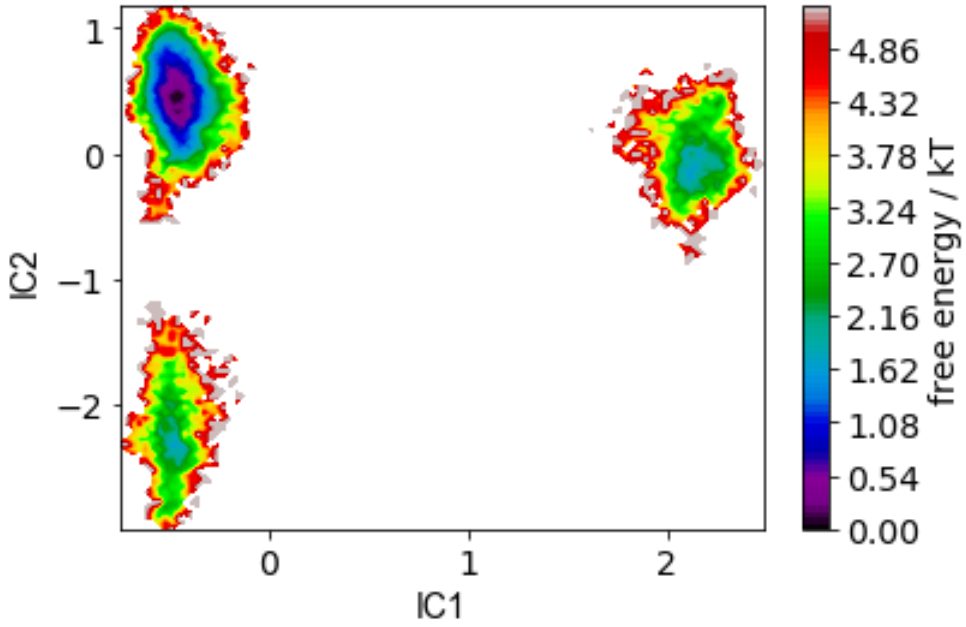

Figure 18 – Free energy plot showing the conformational basins created by the first two ICs after TICA analysis.

TICA also predicts three conformational basins, which correspond to three metastable states. Unlike PCA, the states in the TICA landscape are very well-defined making clustering more reliable. The K-means can now be deployed to discretize the IC landscape and allocate the trajectory structures to their respective clusters, ensuring that the cluster centroids align with the calculated IC landscape. **Figure 19** shows the overlapped IC landscape and the location of the calculated 200 K-means cluster centres. The centroids are very well-distributed within the TICA landscape and match the predicted conformational space.

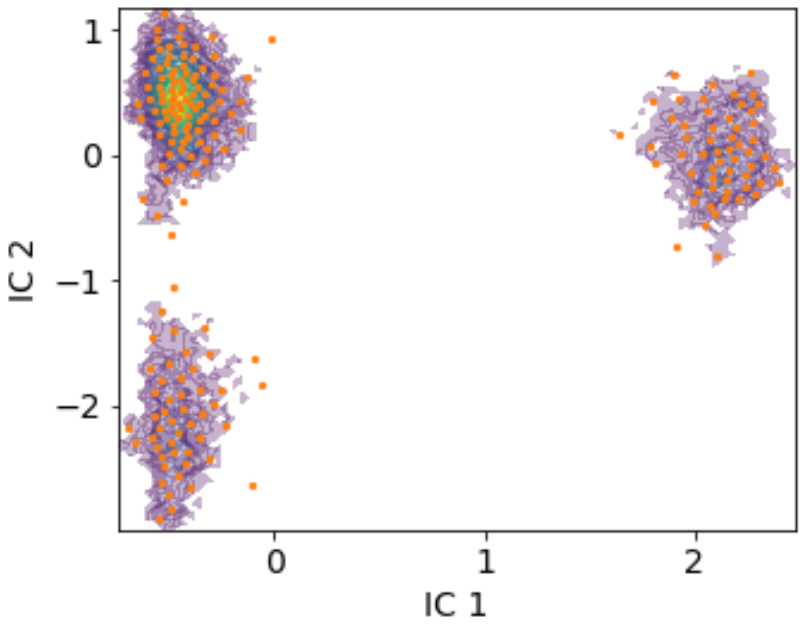

Figure 19 – Free energy plot showing the conformational basins with the K-means overlapped cluster centres (in orange).

With the discretization completed, the data can now be used to build a matrix and calculate the dynamic transitions between each state using a Markov State Model (MSM) to predict the protein's kinetics. The meaningful transitions are calculated at the optimal lag time which can be derived from implied timescales taken at several lag points to determine the relaxation timescales of the processes (**suppl. Figure S6**). After determining the ideal lag, which in the case of MYC88 is 18 ns, it is necessary to assign the clustered microstates to the three metastable macrostates, as predicted by the TICA free energy landscape. The PCCA++ method is used to extract a coarse representation of the MSM and the representative structures for each macrostate. PCCA++, or Perron-Cluster Cluster Analysis, calculates the membership distribution of the clustered structures within the metastable, or long-lived, states (**suppl. Figure S7**). The PCCA++ assignments neatly match the TICA state space, enabling the determination and inspection of the representative structures for each macrostate. **Figure 20** represents the re-weighted TICA landscape, using the stationary distribution, indicating the PCCA++ representative structures attributed to each metastable macrostate.

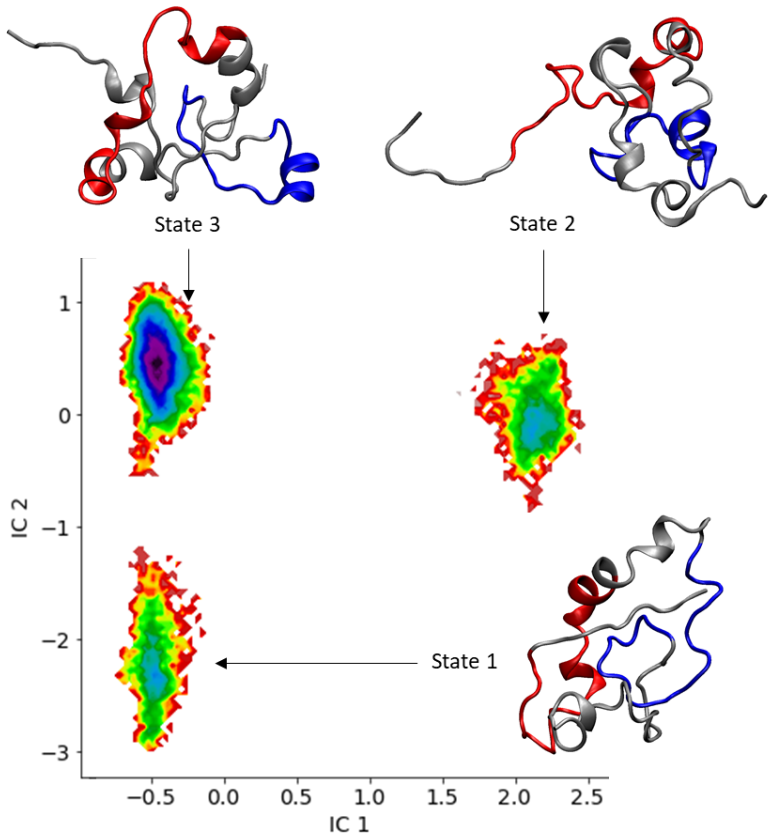

Figure 20 – Representative structures of each of the TICA-predicted macrostates. The structure figures highlight the location of the MB0 (in red) and MBI (in blue).

The main long-lived basin, corresponding to the most abundantly visited pool of conformations, is State 3. This is a key finding, as this pool of conformations affords a window of opportunity for drug discovery – it is a frequently visited protein state and at the heart of the two slowest transition processes. Hence, the representative structures from State 3 formed the structural basis for the 'druggability' studies presented in Chapter III. Exploring along the IC1, the main slowest transition process can be easily determined between State 3 and State 2. The graphical representations of the structures are coloured to identify MB0 in red and MBI in blue, so it is interesting to note that the structural dynamics from State 3 to State 2 entails the extension of MYC88's N-terminal, which includes the MB0 region. The second slowest process along the IC2 axis identifies the transition between State 3 and State 1, which moves MYC88 to a more compacted configuration. Thus, the TICA landscape identifies a protein with a very abundant pool of conformations (State 3), averaging 12.96 Å in Rg. The State 3 displays

the slowest transition to State 2, which mainly consists of the N-terminal extension – with an average Rg of 13.3 Å. The transition from State 3 to State 1 is the second slowest process in which the molecule acquires a slightly more compact structure, with an average Rg of 12.7 Å.

Despite the unsuccessful PCA calculation, it is still important to assess the highest amplitude motions because they usually correspond to rarer but important peak protein configurations. The strategy presented here goes back to basic geometrics measures, in this case the radius of gyration, and assesses the minima and maxima peaks over time (**Figure 21**).

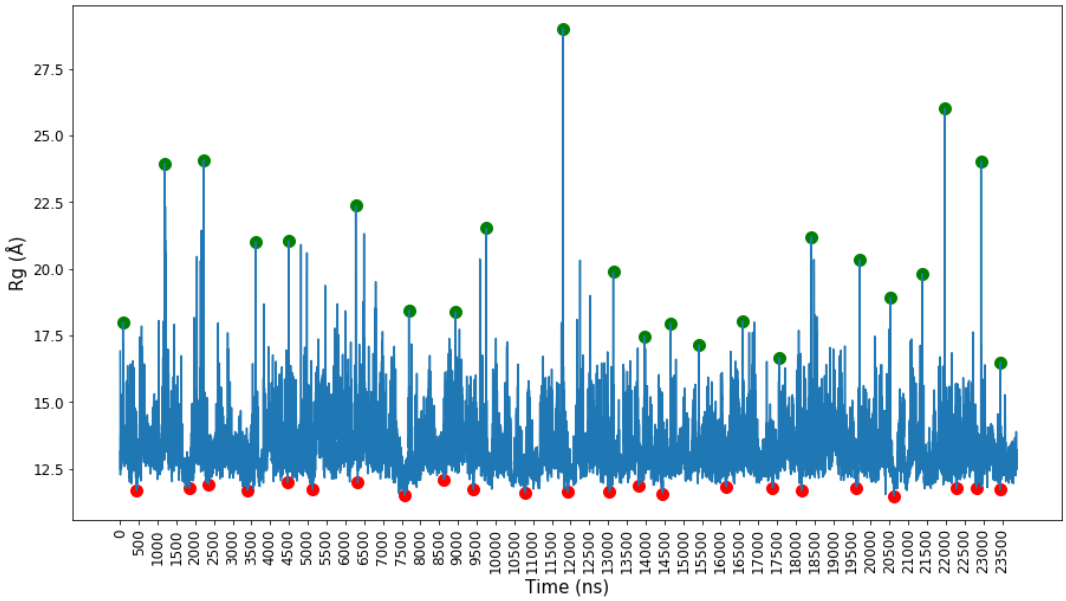

Figure 21 – The radius of gyration values over the course of the trajectory with the minimum and maxima point (coloured red and green, respectively) have been identified.

The noise reduction approach here consists of an algorithm that analyses the trajectory data over time without altering or averaging the peaks, but finds both the maximum and minimum value points, within a rolling window, whilst discarding smaller neighbouring peak values. This allows for the detection of the true peak events within the linear data while eliminating local noise. Considering at the configurations corresponding to the minimum (min) and maximum (max) peaks detected (**Figure 22**), it is obvious that the max values correspond to the most extended and the min values to the most compacted structures.

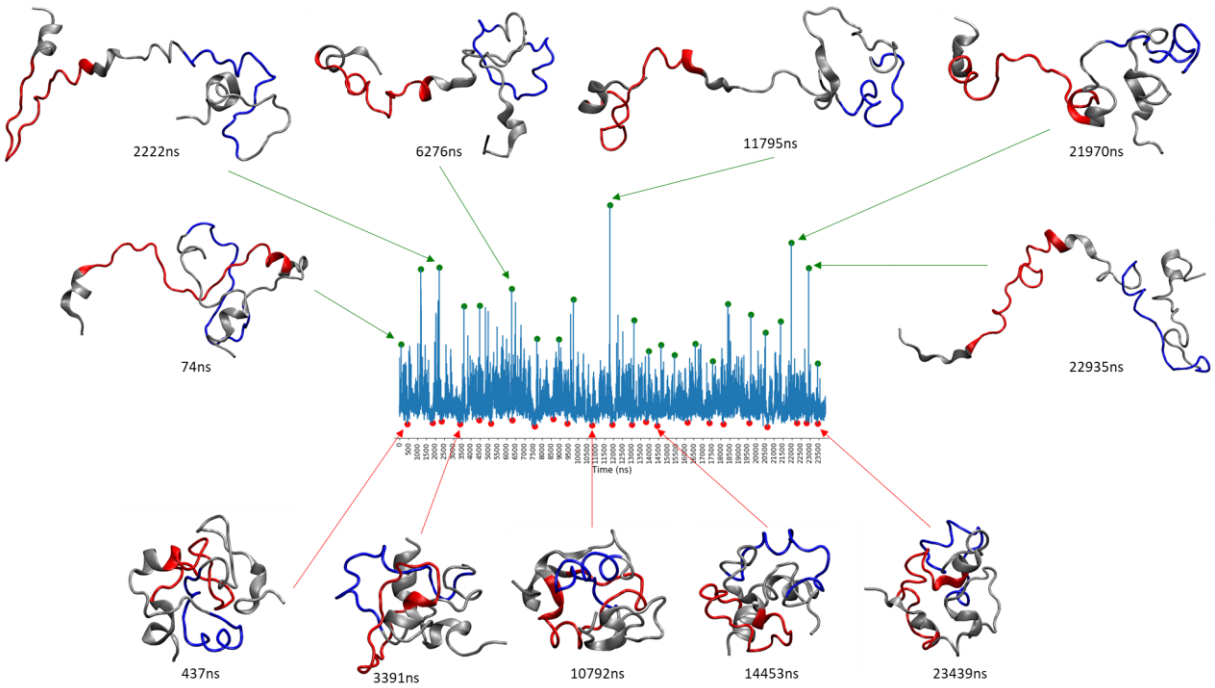

Figure 22 – Some of the minima and maxima configurations over time. The red corresponds to the location of MB0 and blue to the location of MBI.

The min peaks correspond to very compacted structures, constituting rare events within the State 1 predicted by the TICA landscape. The very extended max peak structures are extreme configurations, part of State 2 predicted by the TICA space. The structure of the max configurations reiterates that MYC88's extension involves its N-terminal and clearly includes MB0. These results echo the findings obtained from the TICA landscape, thus it is perhaps helpful to assess the protein's structural sequence, from most compacted to most extended, to fully appreciate the structural range predicted by the MD simulations (**Figure 23**).

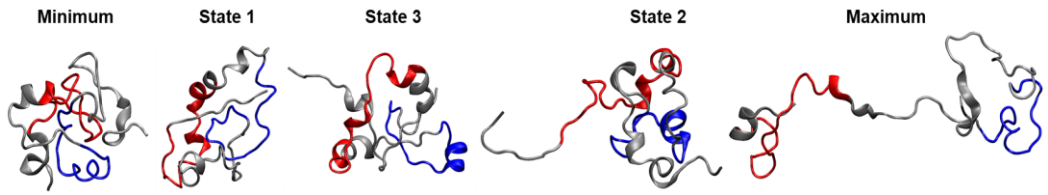

Figure 23 – The representative structures that summarise the structural dynamics of MYC88's MD simulations. The minimum and maximum were derived from the peaks detected from the Rg timeline and the 3 states correspond to the structures obtained from the three TICA metastable states.

The sequence illustrates the range of protein conformations available to MYC88. The minimum and maximum structures indicate rarer peak events, in which the protein acquires a

very compact and very extended configurations, respectively. The three TICA states correspond to averaged and well-sampled pool of conformations, especially State 3, the most abundant. State 3 is the most visited metastable state because it corresponds to an intermediate configuration that can easily go either way: (1) become more compact – therefore less likely to interact with molecular partners; or (2) project the N-terminal outwards in an extended configuration, which promotes the binding with key molecular partners. The N-terminal flexibility is clear when comparing the Rg and the RMSD of the first MYC88 24 amino acids with the remainder MYC88's 63 amino acids (**Figure 24**).

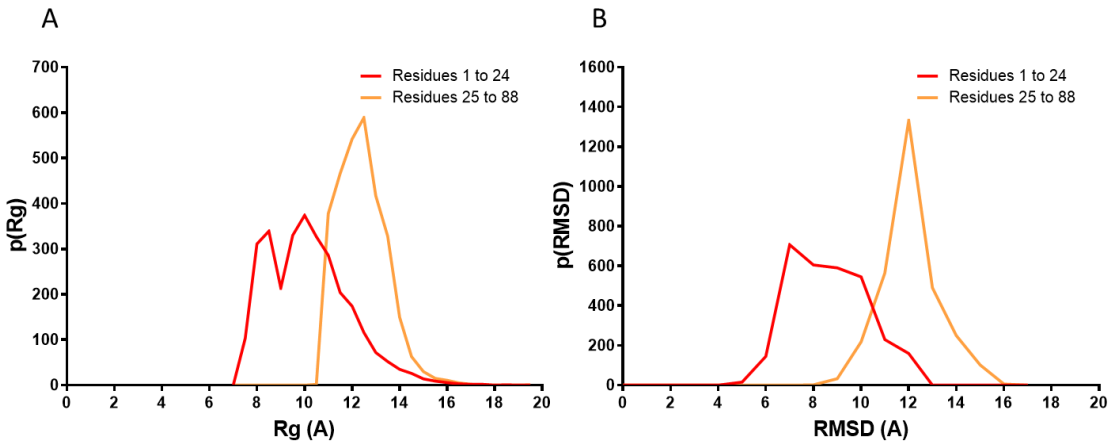

Figure 24 – Plot **A** shows the radius of gyration frequency of MB0's first 24 residues versus the rest of the protein and Plot **B** depicts the same for the RMSD values.

The frequency plots shows that MYC88's first 24 residues alone, a region which includes most of the MB0's residues, displays a wider range of Rg and RMSD values when compared to the remainder 63 amino acids. It is the most active protein region and greatly oscillates between extension and compaction. MYC88's N-terminal RMSD and Rg variability suggests it is the region mainly responsible for most of MYC88's conformational flexibility. MB0's structural dynamics, within a flexible N-terminal, are especially intriguing because until recently this region was not deemed a crucial part of the transactivation domain. Many research papers did not even mention MB0, describing only the activity of MBI and MBII. Only recently did Zhang *et al*. (2017) suggest that MB0 corresponds to a separate and independent transactivation

domain. The results in this Chapter support this view, as MB0's structural dynamics were further investigated to consider the contact formation rate per residue (**Figure 25**).

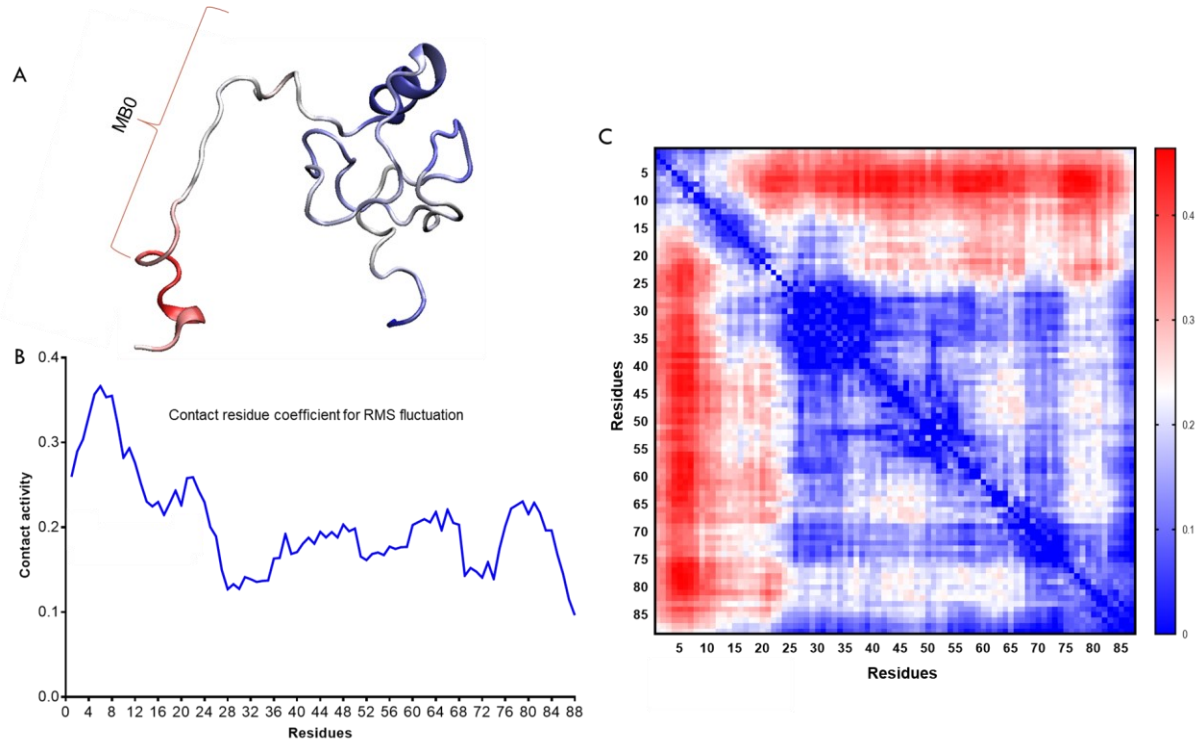

Figure 25 – Plot **B** is a linear graph containing the contact residue coefficient for RMS fluctuation per residue whilst plot **C** offers the data as a heat map to ease the visualisation of active vs inactive protein regions. (**A**) shows the contact activity data superimposed on a representative structure.

This metric allows for the identification of active and inactive protein regions in terms of their contact activity rate. **Figure 25** contains a graph and a heatmap (**plots B and C**) of the same data, depicting the contact residue coefficient per residue. The plots identify the MYC88 regions with the highest (in red) and lowest (in blue) contact formation and breaking activity. The regions most involved in the forming and breaking contacts are inevitably the most active and structurally more dynamic. In **Figure 25 - A** shows the data mapped onto a representative structure clearly reiterating that the region with the highest contact activity is the MYC88's N-terminal residues 1 to 24. The rest of the protein remains comparatively stable. Additionally, the pivot angle formation coefficient calculation allows for the identification of pivot residues that act as hinges and facilitators of the protein's contact activity. **Figure 26** presents the calculations for the pivot angle formation propensity for each residue.

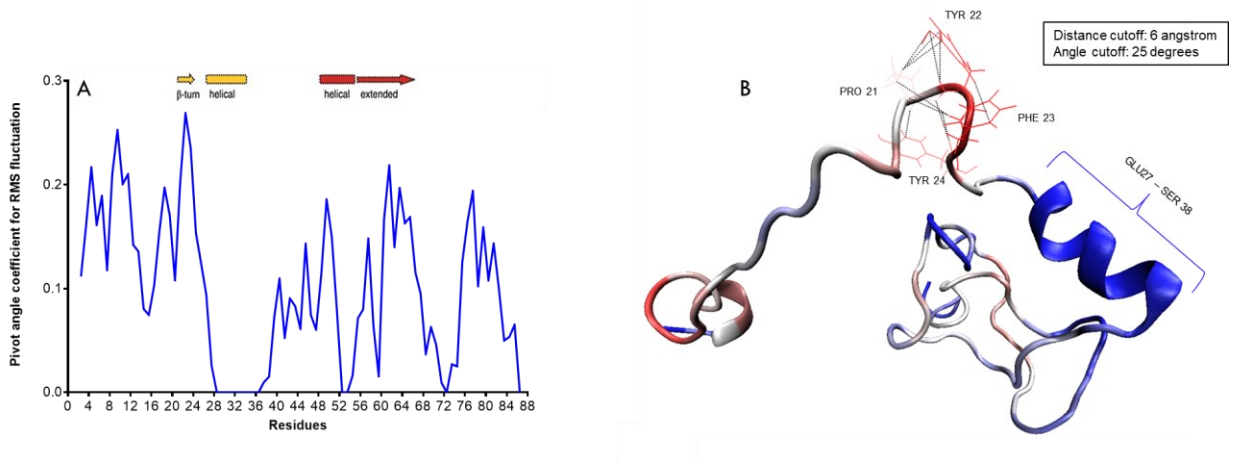

Figure 26 – The plot **A** is the linear data for the pivot angle coefficient, which identifies the residues involved in the formation of pivot angles. The same data is mapped onto a representative structure (**B**). The results show that the most prominent pivot angle is the β-turn formed by amino acids 20 to 24, which corresponds to a highly conserved region identified by NMR experiment (Andresen et al., 2012).

The pivot angle coefficient (**Figure 26 - B**) identifies the residues 20 to 24 as MYC88's main hinge which is responsible for most of MYC88's N-terminal and MB0's conformational extension. The hinge is formed by the same residues predicted by NMR data (**Figure 26 - A**) to form a β-turn. Other notable pivot angles, such as residue 8 to 12 allow the N-terminal to further extend away from the rest of the protein. An NMR predicted stable helix, spanning residues 27 to 38, corresponds in our MD simulations to an equally stable helical region. Up until now no function had been ascribed to this transiently ordered region but our findings show it acts as a divider, allowing both MB0 and MBI to act autonomously in terms of their structural dynamics. The dynamics of these highly conserved regions reiterate the idea that MB0 is a distinct transactivation domain and the ordered region creates an anchor which allows MB0 to perform its alternating extension and compaction motions, fly-catching its molecular partners, without disrupting MBI's stability.

3. Conclusion

Protein MD trajectories are noisy and complex datasets and IDPs MD trajectories even more so. The analysis of IDP trajectories need the careful deployment of methods able to address the trajectory noise. Many commonly used analytical methods fail to adequately

resolve the data complexity. In the case of MYC88's MD simulations, the clustering based on simple geometrics did not yield the expected definition of discrete protein states, neither did the deployment of the PCA. For MYC88, the identification of its metastable states entailed discovering its slowest transitions with TICA analysis. The TICA analysis identified three main metastable states: a slightly compacted State 1, a very abundant intermediate State 3, and a slightly extended State 2. MYC88's slowest transition entailed the movement from the abundant State 3 to State 2 and its second slowest transition from State 3 to State1. The State 3 to State 2 transition sees the protein project outwards its N-terminal, whilst the State 3 to State 1 sees the protein increase in compaction.

MYC88's maximum amplitude motions were analysed using the Rg linear data conjunctly with an algorithm that detects the data's highest min and max peaks. These min and max peaks correspond to rarer structural events: the minimum correspond to a very compacted configuration unlikely to establish any intermolecular interactions; the maximum corresponds to a full extension of the N-terminal and MB0, which restates the idea that MB0 is involved in a fly-catching motion by optimising its binding surface, to attract molecular partners crucial for c-MYC's activation and degradation. The flexible nature of MB0 was further echoed by the study of MYC88's contact activity. It identified MYC88's first N-terminal 24 residues as the region accountable for most of the protein's flexibility. It also identified the hinges which facilitate MB0's extension spanning residues 20 to 24 - an NMR-predicted β-turn, and residues 8 to 12. It also highlighted that the NMR-predicted helical region – covering residues 27 to 38 acts as a stabilising anchor allowing MB0 to act independently whilst preserving MBI's stability. The identification of the protein dynamics, its most abundant configurations and its slowest transitions allows for a window of opportunity in terms of drug targeting. Chapter III presents the results of the drug discovery process applied to MYC88.

Chapter III – Targeting MYC88

1. Introduction

Attempts have been made to target c-MYC at every step of its life - from modulating its transcription levels; affect its mRNA stability; amplify its degradation; or inhibit its interaction with molecular partners. Some drug discovery attempts have targeted it indirectly by disrupting the binding to its vast interactome, while others relied on a more direct approach by interfering with its production and/or degradation.

1.1 Indirect approaches

One of the main ways to indirectly target c-MYC has been to block its transcription and modulate its expression. Current research into bromo- and extra-terminal domain (BET) inhibitors, such as JQ1, found that this small molecule inhibitor can interfere with bromodomain chromatin regulators and downregulate c-MYC's transcription (Delmore et al., 2011). Specifically, JQ1 is thought to prevent the activity of the chromatin remodeller, and c-MYC coactivator, bromodomain-4 (BRD4) (Carabet, Rennie & Cherkasov, 2018), and induce cell senescence, cell-cycle arrest and reduce tumour activity in animal models of multiple myeloma, Burkitt lymphoma and acute myeloid leukaemia (Delmore et al., 2011; Jennifer A. Mertz et al., 2011). Recent research has disputed the idea that JQ1 inhibits c-MYC levels as, in some well-studied cases, c-MYC's levels remain quite high despite JQ1's activity (Ambrosini et al., 2015; Andrieu, Belkina & Denis, 2016; Bid et al., 2016; Garcia et al., 2016; Hogg et al., 2016; Yao et al., 2015). Furthermore, there is evidence suggesting that c-MYC can easily develop protective mutations, rendering it immune to the downregulatory effect induced by BET inhibitors (Fong et al., 2015; Rathert et al., 2015; Shi, X. et al., 2016; Shimamura et al., 2013). Lowering c-MYC's transcription levels was also attempted by targeting the cyclin-

dependent protein kinase 7 (CDK7), a catalytic subunit of the CDK-activating kinase (CAK). The CAK creates a complex with the human transcription factor II (TFIIH) and facilitates, via serine-5 phosphorylation of the RNA polymerase II (RNA Pol II), the subsequent elongation of the target transcripts (Sun, B. et al., 2020). The CDK7 inhibitor THZ1 has shown promise in a variety of cancers including neuroblastoma (Chipumuro et al., 2014; Kwiatkowski et al., 2014); lung cancer (Christensen et al., 2015) and triple-negative breast cancer (Wang, Y. et al., 2015). Nevertheless, although this inhibitor was found to lower c-MYC levels in neuroblastomas, it is still not fully established that it does so by blocking c-MYC's transcription (Whitfield, Beaulieu & Soucek, 2017).

Another approach to inhibit c-MYC's overexpression has been to block its translation. The mTOR translation regulatory mechanism has been particularly aimed at by several small molecular inhibitors. Many of these inhibitors, already approved for clinical use, affect mTOR directly or indirectly by regulating the activity of its many pathway partners (Polivka & Janku, 2014; Roohi & Hojjat-Farsangi, 2017). The eukaryotic initiation factor-4A (eIF4A) has been consistently targeted, notably by the inhibitor silvestrol, and was found to reduce c-MYC translation rates and impair tumour development (Wiegering et al., 2015). The Src kinase has also been targeted using the inhibitor saracatinib and found to downregulate eIF4A-mediated c-MYC translation (Jain et al., 2015).

Additionally, c-MYC has been the focus of synthetic lethality studies. In these studies, many proteins, unrelated to c-MYC, have been identified as potential targets since they display lethality when c-MYC is overexpressed (Cermelli et al., 2014; Xin Li et al., 2015). Namely, the artemisinin by-product compound dihydroartemisinin was found to indirectly destabilise c-MYC by activating the glycogen synthase kinase GSK3-β, promote THR58 phosphorylation, and subsequent c-MYC ubiquitination (Wei et al., 2018). Other lethality screen identified targets include Aurora kinases (Whitfield, Beaulieu & Soucek, 2017). Studies with Aurora-A kinase and n-MYC have suggested that the formation of Aurora A and n-MYC complex rescues n-MYC from degradation. Thus, it has been proposed that disrupting this interaction,

via small inhibitors, would be a viable tactic to promote normal n-MYC, and potentially c-MYC, degradation (Brockmann et al., 2016).

Attempts to find immunotherapy treatments for c-MYC-driven tumours have also gained traction. Notably, PCI-32765 (Ibrutinib) - an inhibitor of Bruton's tyrosine kinase (BTK) currently used in cancer treatments (Whitfield, Beaulieu & Soucek, 2017), was found to possess inhibitory activity against the formation of pancreatic islet tumours, a c-MYC-driven neoplasia (Soucek et al., 2011).

Many approaches have aimed to disrupt c-MYC's stability. Namely, the E3 ligases responsible for c-MYC ubiquitination, FBW7 and Skp2, have been induced by compounds such as oridonin, a diterpenoid, to promote c-MYC degradation (Huang, Hui-Lin et al., 2012). Targeting strategies have also aimed to compromise the deubiquitinating proteins, including USP28, USP38, and USP36, to enhance c-MYC's destruction (Sun, X., Sears & Dai, 2015).

Despite showing promise, and many compounds progressing to clinical trials, the success of the indirect approaches has been dampened by the lack of demonstrated specificity to c-MYC. Although several inhibitors are undergoing clinical trials, many researchers still question the idea that the observed clinical efficacy is caused by interference with the c-MYC's transcription or is even related to c-MYC at all (Whitfield, Beaulieu & Soucek, 2017).

1.2 Direct approaches

The direct targeting of c-MYC has proven equally difficult, to the extent that the protein has been deemed undruggable (Dang, Chi V. et al., 2017). This has mainly to do with c-MYC's disordered nature and the absence of identifiable stable druggable pockets and cavities (Whitfield, Beaulieu & Soucek, 2017). To complicate matters further, c-MYC is a nuclear protein with no enzymatic function and a history of poor interaction with small molecules (Posternak & Cole, 2016).

The most noteworthy attempt to directly target c-MYC involved blocking the MYC/MAX dimerization and downregulate c-MYC's transcriptional activity. One of the first small molecules to be tested as a MYC/MAX dimerization inhibitor was the peptide mimetic molecule IIA6B17 (Thorsten Berg et al., 2002). Unfortunately, this small compound lacked specificity and displayed heavy cross reactivity with other proteins, specifically with c-Jun, limiting its prospects as MYC/MAX inhibitor (Berg, 2008). Shortly after, another compound was developed, the 10058-F4, to bind c-MYC's bHLH-LZ domain in a more specific way (Yin et al., 2003). 10058-F4 performed excellently *in vitro*, and it is still commonly used in assays as a c-MYC inhibitor. However, it failed in animal testing due to its largely inadequate pharmacokinetic and pharmacodynamic qualities (Horiuchi, Anderton & Goga, 2014). Its fast metabolization, poor tissue bioavailability and inadequate tumour penetration made 10058-F4 an unacceptable candidate for further clinical trials (Fletcher & Prochownik, 2015; Guo, J. et al., 2009). Several other small molecule inhibitors have been described in literature, including 10074-G5 (Yin et al., 2003) and JY-3-094 (Wang, H. et al., 2013; Yap et al., 2013) which, likewise, suffered from low bioavailability and poor cell penetration.

c-MYC's direct drug discovery approaches have focused exclusively on the C-terminal, particularly the bHLH-LZ region, for a good reason: the bHLH-LZ is c-MYC's most ordered region with an established crystal structure. Conversely, the lack of structural knowledge regarding c-MYC's N-terminus makes it difficult to find suitable pockets within this region. However, as discussed in the previous Chapter, the MD trajectory analysis identified MYC88's slow transitions and abundant metastable states. Armed with this structural dynamics' knowledge, a rational approach to the assess the 'druggability' of this region is now feasible and is the main topic of this Chapter.

2. Results and Discussion

3.1 Pocket discovery

After determining MYC88's most abundant metastable state using TICA analysis, its 'druggability' largely depends on finding a suitable druggable pocket. The ideal cavity is assessed in terms of its geometry, favourable electrostatics, ideal polar and hydrophobic composition, and stability over time (**Figure 27**). This has typically precluded drug discovery research into c-MYC's N-terminal due to its intrinsically disordered nature. However, as previously discussed in Chapter II, MYC88 possesses transiently ordered regions that can be targeted. Additionally, the structure put forth to drug discovery analysis belongs to a very abundant metastable state which offers a good window of opportunity for pocket stability.

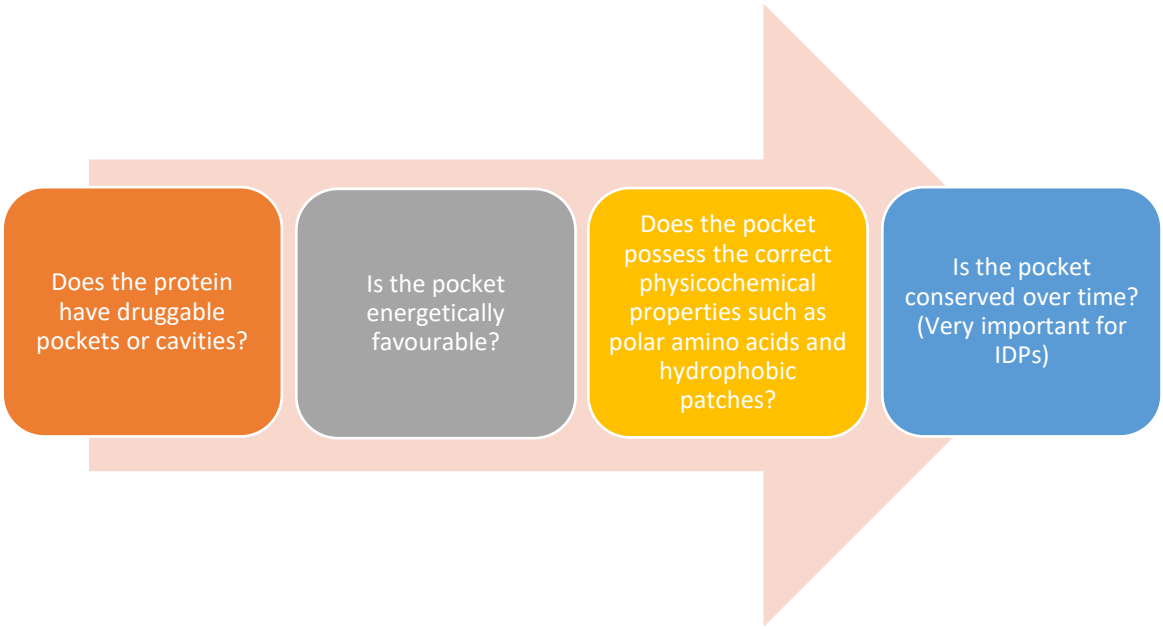

Figure 27– Diagram summarising the rationale informing the search for a druggable pocket.

The representative structure for TICA State 3 was used to look for a suitable pocket that would satisfy the step-by-step conditions defined in **Figure 27**. To do so, various *in silico* drug discovery methods were used in tandem to assess MYC88's surface for suitable pockets and cavities. A binding site consensus across different methods would robustly indicate a suitable druggable region.

The first tool used, CASTp, calculates the existence of suitable cavities based on geometry algorithms including the molecular surface volume of the pocket, the solvent-accessible surface volume, and the molecular surface area of the cavity's mouth. CASTp defines pockets as empty concavities on a protein surface with a mouth opening connecting the protein's interior with the solvent (with a probe sphere of 1.4 angstroms). **Figure 28** shows CASTp pocket prediction results mapped onto MYC88's representative structure.

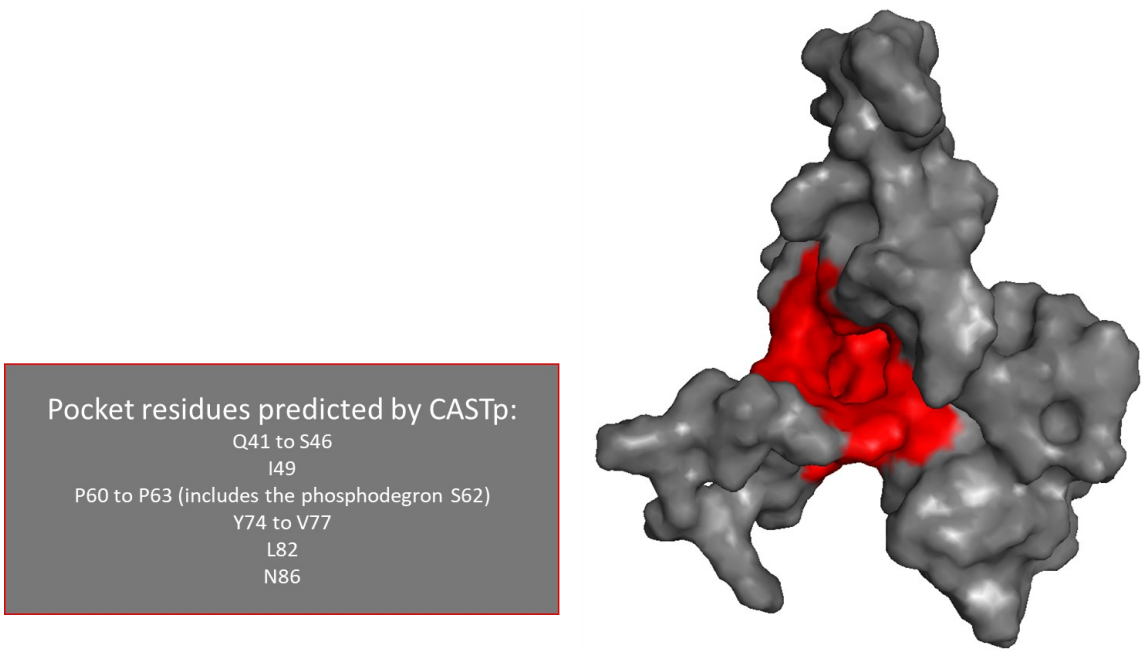

Figure 28 – CASTp pocket prediction results using a geometry-based approach. The ideal pocket region is highlighted in red and spans residues 41 to 46, 49, 60 to 63, 74 to 77, 82 and 86.

CASTp's prediction identifies a suitable pocket with correct geometric features. Interestingly, the residues involved in forming the pocket (**Figure 28** - in red) are mostly highly conserved amino acids within the MBI region, including SER62. A ligand binding this pocket is likely to interfere with MBI's interaction with molecular partners and interfere with the MBI mediated intramolecular contact network. However, a successful binding site must also display a high-level binding affinity to ligands. To assess this, FTMap was used as a tool to identify docking regions, or hotspots, possessing high affinity binding to organic probes (**Figure 29**). The probes consist of 16 small organic molecules of varying sizes, shapes and

polarities which help identify binding hotspots on the protein's surface. The identification of a consensus site for many probes indicates a highly druggable region.

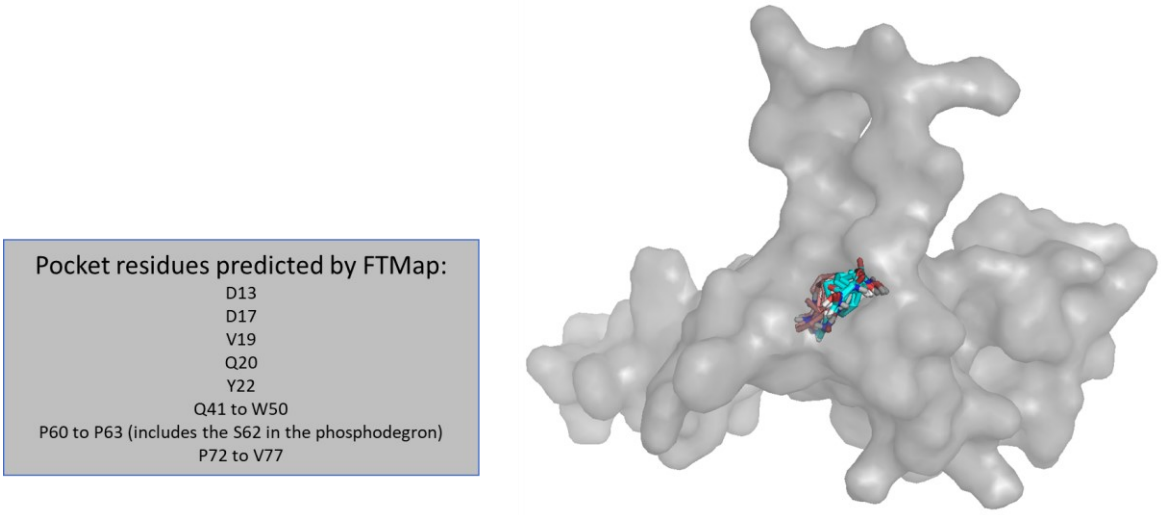

Figure 29 – FTMap results showing the consensus site identified by the 16 small molecule organic probes and the residues involved in creating the binding hotspot.

Extraordinarily, all 16 probes converged to the exact same spot, indicating a very strong druggable region. Furthermore, the predicted FTMap pocket is formed by most of the residues identified by CASTp, including the MBI amino acids and the all-crucial phosphodegron SER62 residue. Additionally, FTMap predicted a few highly conserved residues in the MB0 region as also participating in the formation of the binding hotspot.

The encouraging results obtained with FTMap and CASTp were replicated using another tool, PockDrug, which assesses the pocket in terms of its geometry, hydrophobicity, and polarity. PockDrug is an all-rounder application that assesses pockets in terms of their geometric aspects such as its shape, volume, and solvent-accessible area; and their biochemical aspects, including its residue composition, hydrophobicity, polarity and aromaticity to detect the presence of regions with high potential to form favourable intermolecular interactions with ligands (**Figure 30**).

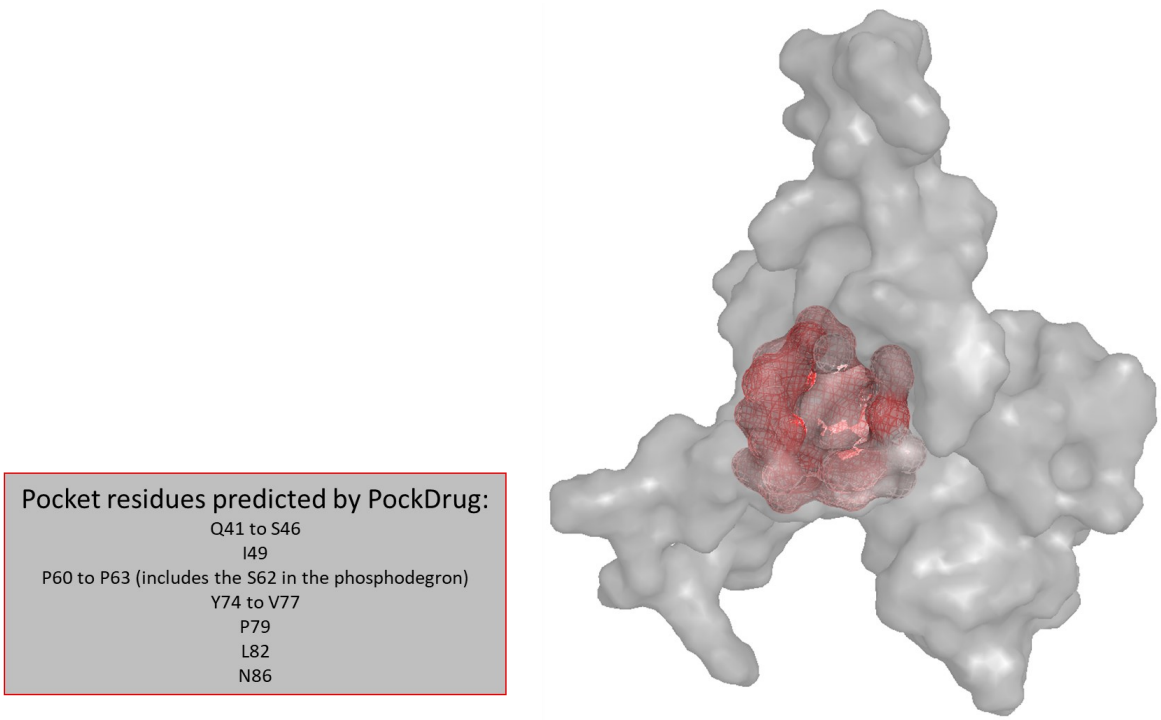

Figure 30 – PockDrug predicted pocket residues (in red) mapped onto MYC88's structure.

Yet again, the PockDrug analysis results reflect nearly exactly CASTp's findings. Both applications identify the same MBI residues, amino acids 41 to 46, 49; the phosphodegron 60 to 63 containing the all-important SER62; as well as some C-terminal residues 74 to 77, 79, 82 and 86, as the predicted pocket. Furthermore, PockDrug estimates that the pocket is composed of 33% polar residues and 31% hydrophobic residues based on Kyle-Doolitle hydrophobicity score, offering a high 'druggability' probability score of 84%.

Despite such encouraging results, it is imperative to consider the pocket stability over time and this is even more important for an IDP, such as MYC88. To assess the existence of a robust and well-maintained pocket another key analysis tool was deployed – MDPocket. This tool assesses the pocket stability by analysing its evolution over the course of the simulation. To do so, 10 snapshots were extracted from a 1000 ns simulation sampled at every 100 ns, allowing MDPocket to detect the presence of a conserved pocket throughout the trajectory. **Figure 31** shows the MDPocket results mapped onto the representative structure.

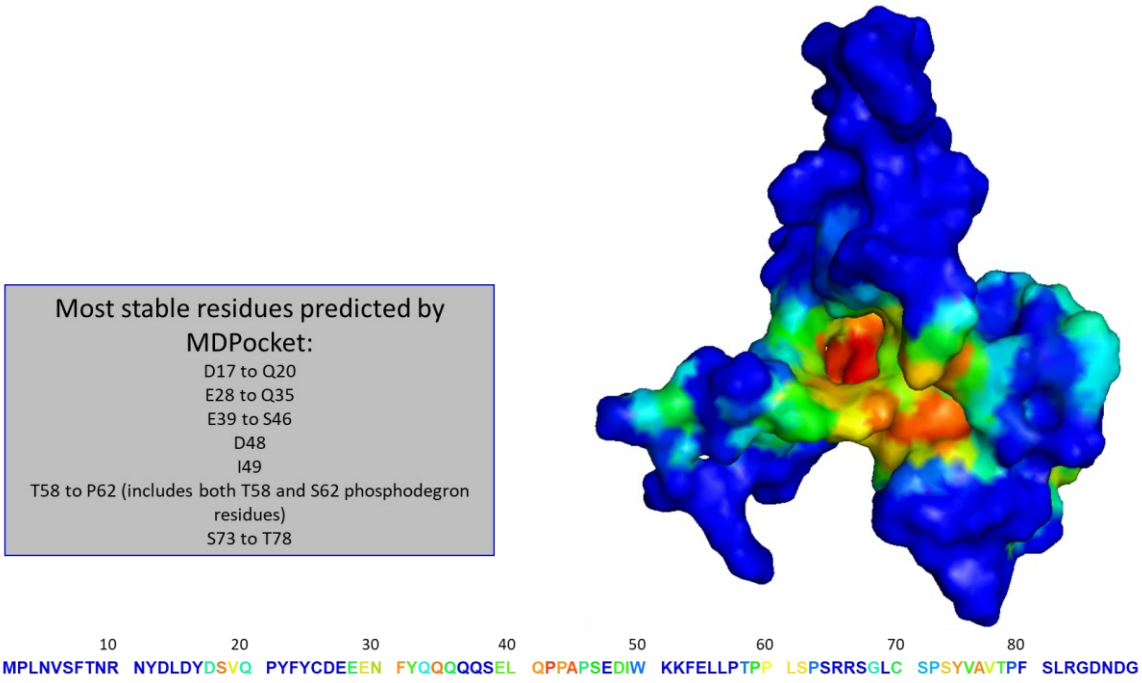

Figure 31 – MDPocket results identify which residues are the most stable over time the course of the trajectory (in red) and the least stable (in blue). The sequence below is coloured in the same colour scheme as the figure and identifies that the most stable regions correspond to the predicted pocket.

The MDPocket identifies the pocket residues as being maintained over time. In other words, MDPocket's findings show that the pocket created by the residues previously identified by CASTp, FTMap and PockDrug is very stable over the course of the MD trajectory. This is consistent with the results obtained from Chapter II, since the residues involved in forming the pocket are mostly MBI residues predicted to be the most stable part of the protein.

Therefore, based on the comparison of the results across the different methods a highly druggable pocket was identified. The consensus across the different methodologies identifies the residues 41 to 50, 60 to 63 and 74 to 77 as the main pocket. Since this pocket is mainly comprised of MBI residues, a region very specific to the MYC family of proteins, it is unlikely to suffer from the promiscuity, and cross-reactivity problems which plagued other c-MYC drug compounds. With a suitable pocket identified and its location defined as the docking site for potential ligands, the drug discovery process can proceed unimpeded.

3.2. Drug discovery

With MYC88's druggable pocket defined, tools were deployed to search for compound candidates which could target the predicted pocket with favourable free energy binding. This drug discovery process was undertaken using Autodock Vina (Trott & Olson, 2010) and iDock (Hongjian Li, Kwong-Sak Leung & Man-Hon Wong, May 2012) to identify the best drug candidates. Out of a total of 23,221,614 compounds screened, the 10,000 best scoring ligands were selected. To narrow this number down to a more manageable figure, only ligands that satisfied the Lipinski rule (rule of five) were considered, as this ensures that ligands with poor pharmacokinetics are discarded.

The Lipinski rule, derived from the work of (Lipinski et al., 1997), aims to provide sound guidelines for the selection of suitable compounds in drug discovery. According to the Lipinski rule, compounds with good pharmacodynamics, pharmacokinetics and satisfactory bioavailability properties should not violate more than one Lipinski condition. Ligands with two or more violations are expected to display inadequate absorption and permeability (Lipinski et al., 1997; Petit et al., 2012) and were not selected for testing. The Lipinski rule states that small ligands should have:

- No more than 5 hydrogen bond donors and no more than 10 hydrogen bond acceptors.
- A molecular mass less than 500 Daltons.
- An octanol-water partition coefficient log P not greater than 5.
- Fewer than 10 rotatable bonds.

Additionally, it is also possible to search for smaller lead-like compounds with more stringent parameters:

- An octanol-water partition coefficient log P not greater than 3.
- A molecular mass lower than 300 Daltons.
- 3 hydrogen bond donors or less; 3 hydrogen bond acceptors or less; and 3 rotatable bonds or less.

To choose the ideal drug candidates, with adequate physicochemical characteristics and solubility, the 10,000 compounds were triaged in order to identify the ligands that satisfied all Lipinski constraints and displayed a free binding energy of < -10.2 (kcal/mol); and the ligands that satisfied the lead-like parameters with a free binding energy of < -8.5 (kcal/mol) (**Figure 32**).

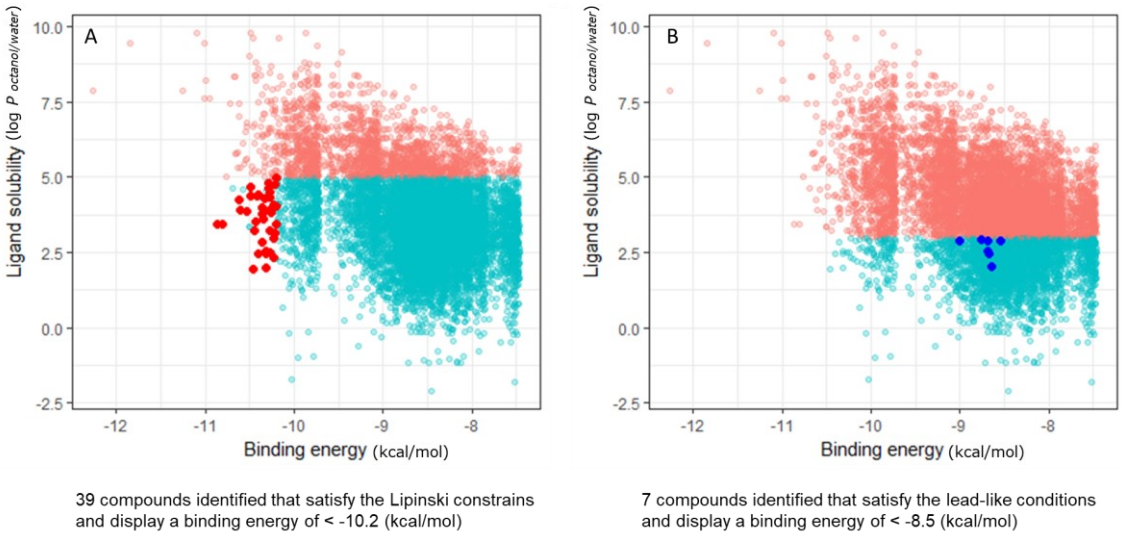

Figure 32 – Identification of compounds out of the top scoring 10.000 candidates that (A) fulfil the Lipinski rule and have a binding energy of < -10.2 (kcal/mol); (B) fulfil the lead-like constraints and display a binding energy of < -8.5 (kcal/mol).

A total of 46 compounds were identified: of these 39 satisfied the Lipinski rule and bound MYC88 with a free energy less than -10.2 (kcal/mol); whilst further 7 complied with lead-like constraints and bound the protein with free energy less than -8.5 (kcal/mol). These compounds were all tested using MD simulations to assess their interaction with MYC88. Of the 46 compounds, only 14 demonstrated stable binding throughout the MD trajectory and of these, the best performing 6 were chosen due to their noteworthy binding stability and energies (**Figure 33**). The chemical structures of these 6 ligands can be found in **suppl. Figure S5**.

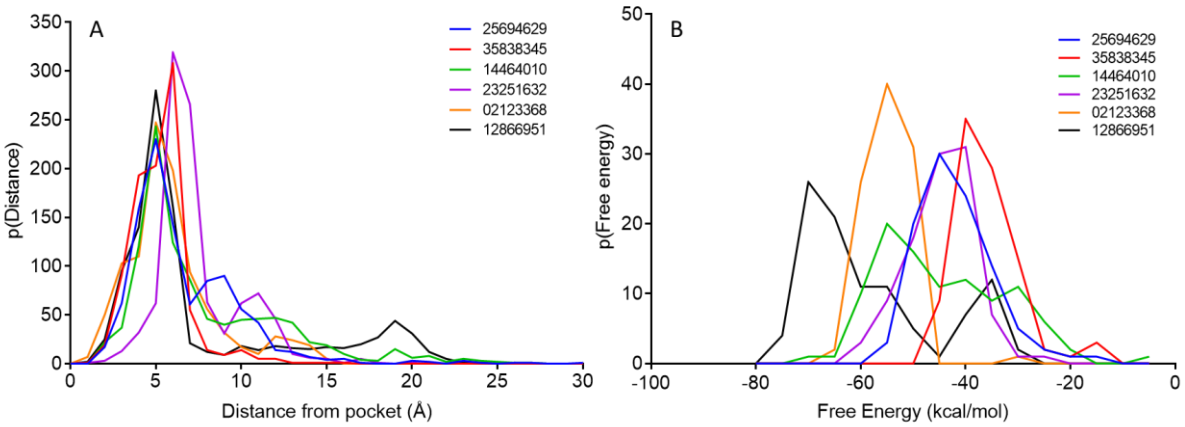

Figure 33 – Frequency distribution plots showing each ligand's distance from the pocket throughout the simulation (A) and the ligand's binding energy to MYC88's pocket throughout the simulation.

Binding stability is defined by the ligand's sustained binding and its closeness to the predicted pocket throughout the course of the MD trajectory. On average, all ligands maintained a very close distance of 5 Å to the predicted pocket (**Figure 33 - A**). In tandem, the ligands were also found to favourably bind the pocket, as seen on **Figure 33** – **B**. Each tested ligand bound MYC88's pocket very favourably creating complexes with binding energies lower than -20 kcal/mol.

The tested compounds' high binding affinity and a strong binding stability maintained throughout the trajectory demonstrates the robustness of their binding kinetics. Such strong interactions are likely to have consequences to MYC88's conformational landscape, which can now be evaluated by deploying analytical methods previously unsuccessful due to MYC88 intrinsic disorder. In other words, PCA can now be used to analyse the MYC88-ligand complex conformational landscape (**Figure 34**).

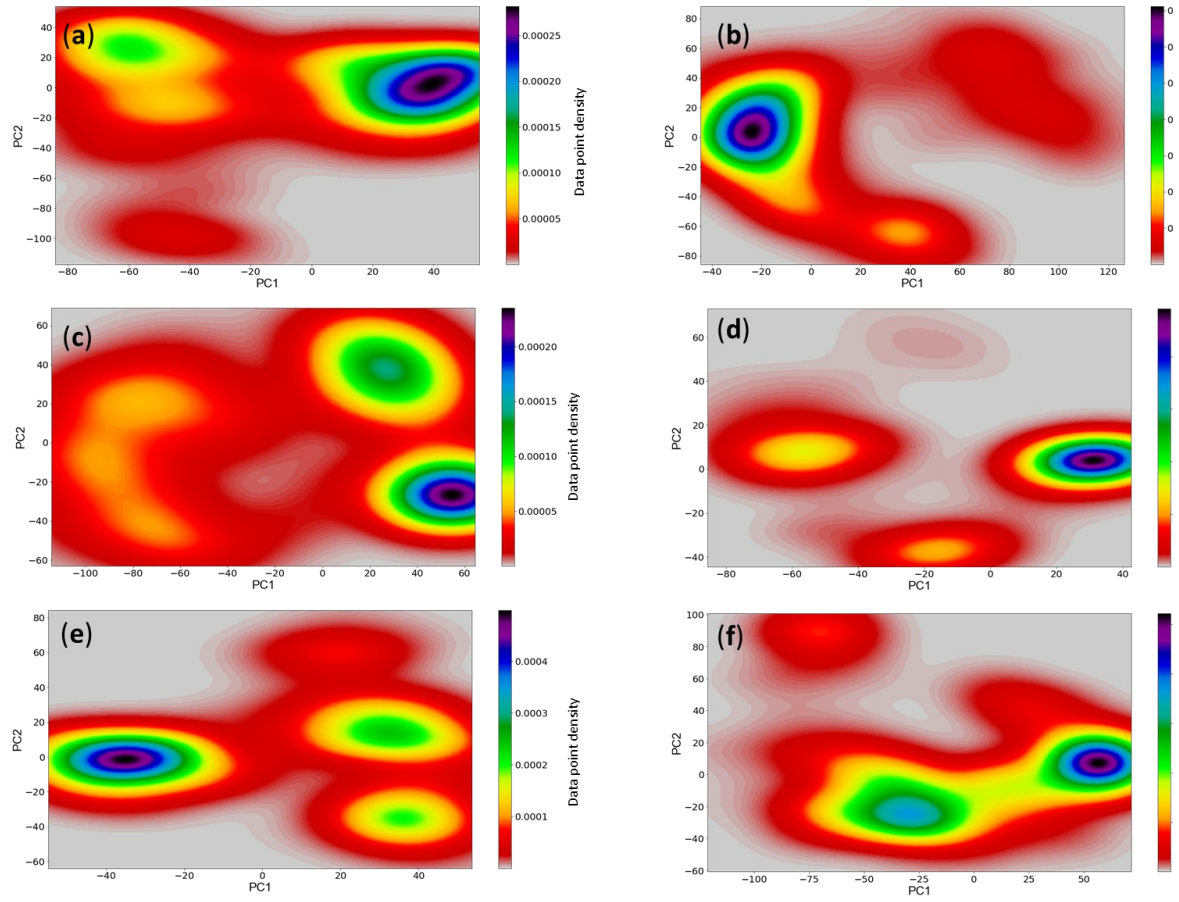

Figure 34 – PCA landscapes for each of the ligands: (**a**) 02123368, (**b**) 12866951, (**c**) 14464010, (**d**) 23251632, (**e**) 35838345 and (**f**) 25694629.

Previously inadequate to resolve MYC88 conformational states, the PCA plots now reveals that MYC88-ligand trajectories display a significant loss of protein disorder, now revealing 'clusterable' landscapes. This implicates the ligands in MYC88's loss of natural disorder and suggests their role in forcing MYC88 to adopt discrete states, which can now be directly discerned from a simple PCA noise reduction analysis.

Ligands, such as 23251632, compel MYC88 to inhabit well-defined PCA clusters with PC's that explain close to 60% of the data variance. In the case of ligand 23251632 there are three distinct clusters, which can then be used to deploy the K-means clustering algorithm (**Figure 35**).

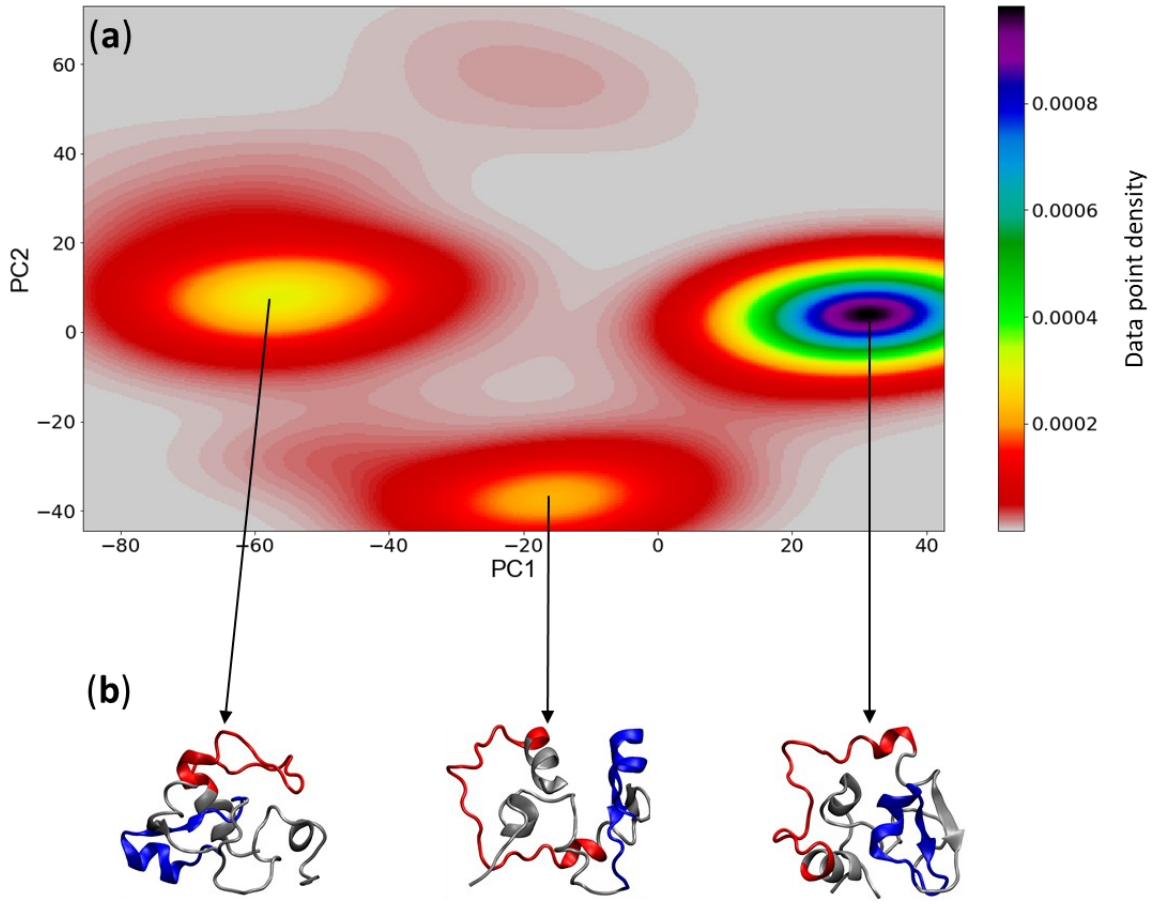

Figure 35 – This figure shows the data analysis for the MYC88 targeted with ligand 23251632 (a) the PCA landscape showing the allocation of the 3 predicted clusters and (b) the representative centroid structures for each of the clusters.

The representative structure obtained from each cluster (**Figure 35**) shows that MYC88, when drugged with ligand 23251632, appears structurally folded, ordered, and compacted in all its conformational clusters. The visual inspection of the structures indicates that the loss of disorder is linked to MYC88 N-terminal's, which includes the MB0's region (coloured in red in **Figure 35**), loss of its dynamic 'fly-casting' movement. This abrogated N-terminal extension is the main cause of the protein compaction. This can be further discerned by considering the overall frequency distribution of the Rg for the unbound MYC88 simulation versus the MYC88 in interaction with ligand 23251632 (**Figure 36**).

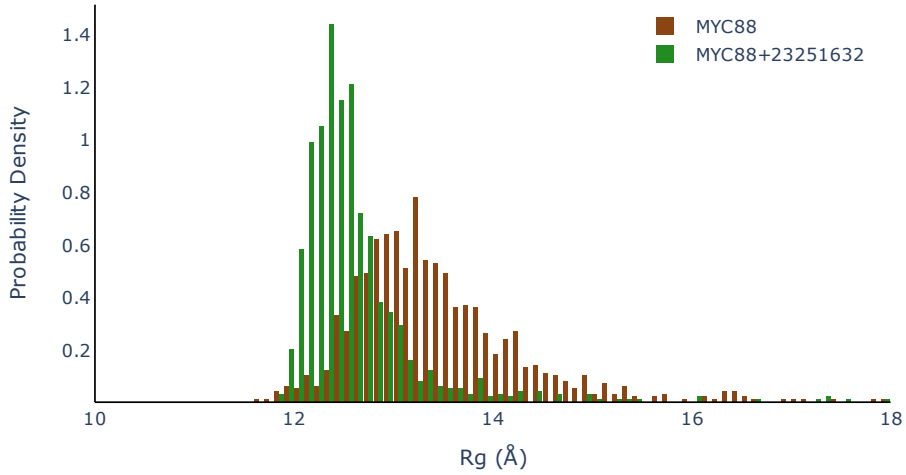

Figure 36 – Comparative probability density Rg histogram for the MYC88 simulation (brown) and MYC88 targeted with the ligand 23251632 (green).

The ligand interaction with MYC88 dramatically shifts the protein's Rg towards a more compacted and less conformationally diverse space. Also, the comparative analysis of the secondary structure propensities shows that the ligand enhances MYC88's ordered content when compared to unbound MYC88 (**Figure 37**).

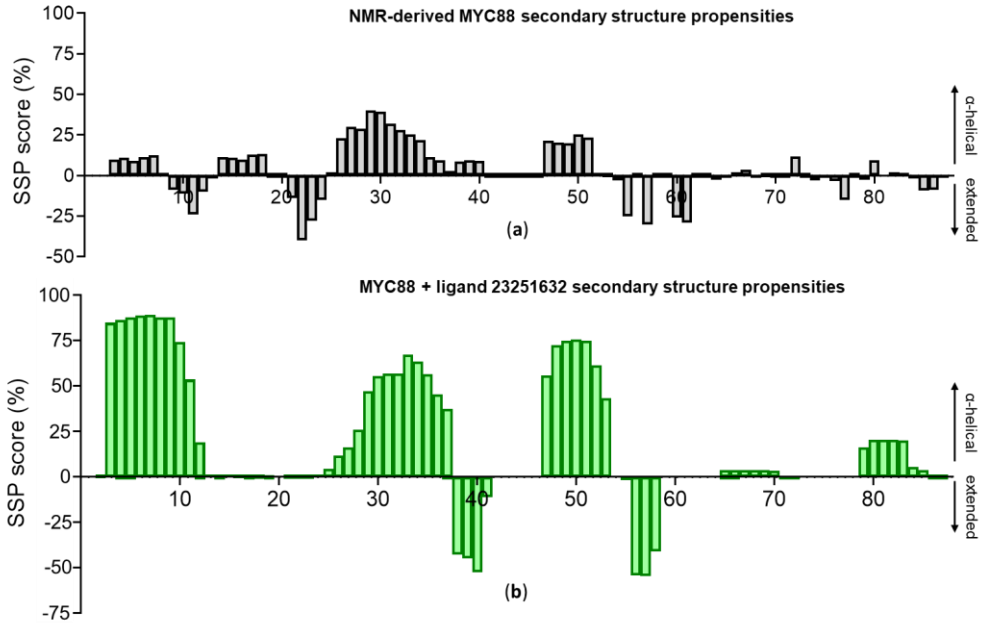

Figure 37 – The secondary structure propensities (SSP) scores for the MYC88 simulation (**a**) and MYC88 targeted with the ligand 23251632 (**b**).

MYC88 when targeted with the ligand 23251632 displays a remarkable 50% increase in helical formations of residues 3 to 12, which serves to stabilise the entire MB0 region. It also displays an abrogation of the β-turn formation between residues 22 and 25 which, as previously seen, acts as a crucial pivot angle and a hinge facilitating MYC88's N-terminal extension and retraction.

The ligand's stabilizing effect on MYC88's conformational potential is more evident when considering the MYC88's peak minimum and maximum configurations extracted directly from the Rg over time linear plot (**Figure 38**).

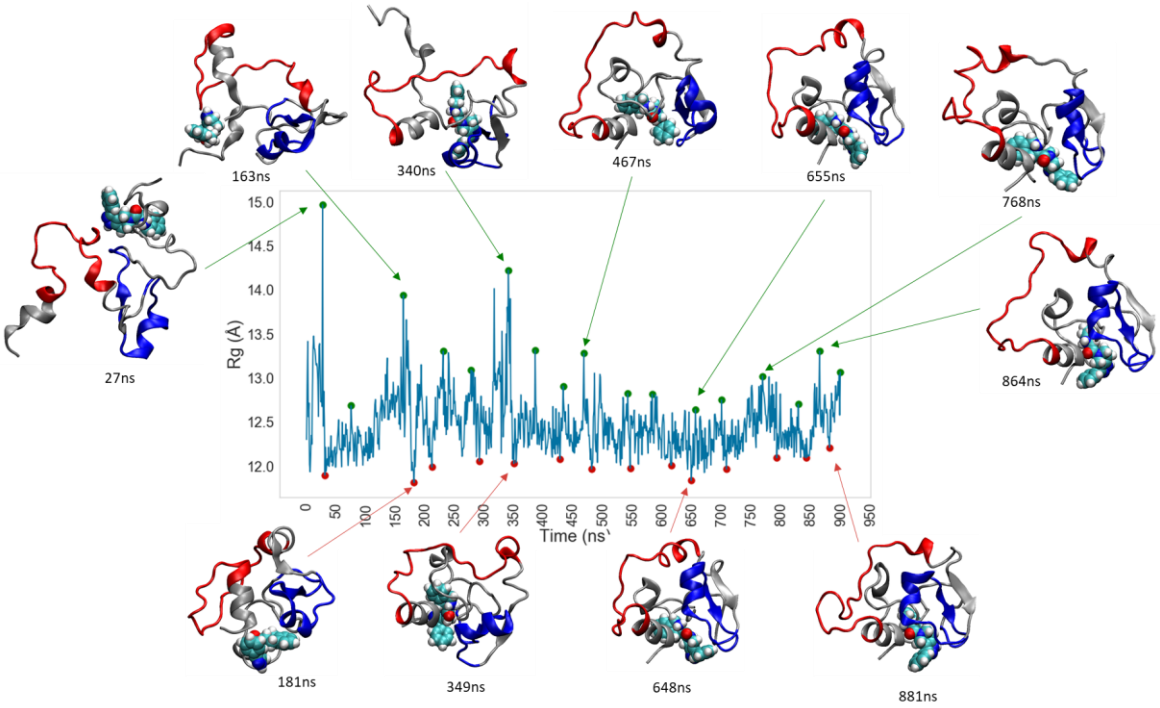

Figure 38 – Timeline evolution of MYC88's Rg showcasing the peak minimum and maximum structures. The figures display the ligand and the location of both MYC boxes: MB0 in red and MBI in blue.

Figure 38, which shows the MYC88 configurations timeline evolution, also displays the location of the ligand 23251632 for each represented structure and allows for the visual assessment of the ligand's interaction with MYC88 over the course of the trajectory. It is curious to note that at the start of the simulation MYC88's N-terminal initially attempts an extension at 27 ns. However, the subsequent maximum peaks highlight how MYC88's N-

terminal becomes so trapped in its interaction with the ligand that is unable to extend out from the protein. So much so that, after the first 500 ns, the variance in Rg between the maximum (~12.5 Å) and minimum peaks (~12 Å) is only 0.5 angstroms, suggesting extraordinary protein stabilisation and narrowing of the conformational range.

By investigating the MYC88/ligand contact map it was possible to identify which exact protein residues the ligand is targeting throughout the trajectory (**Figure 39**).

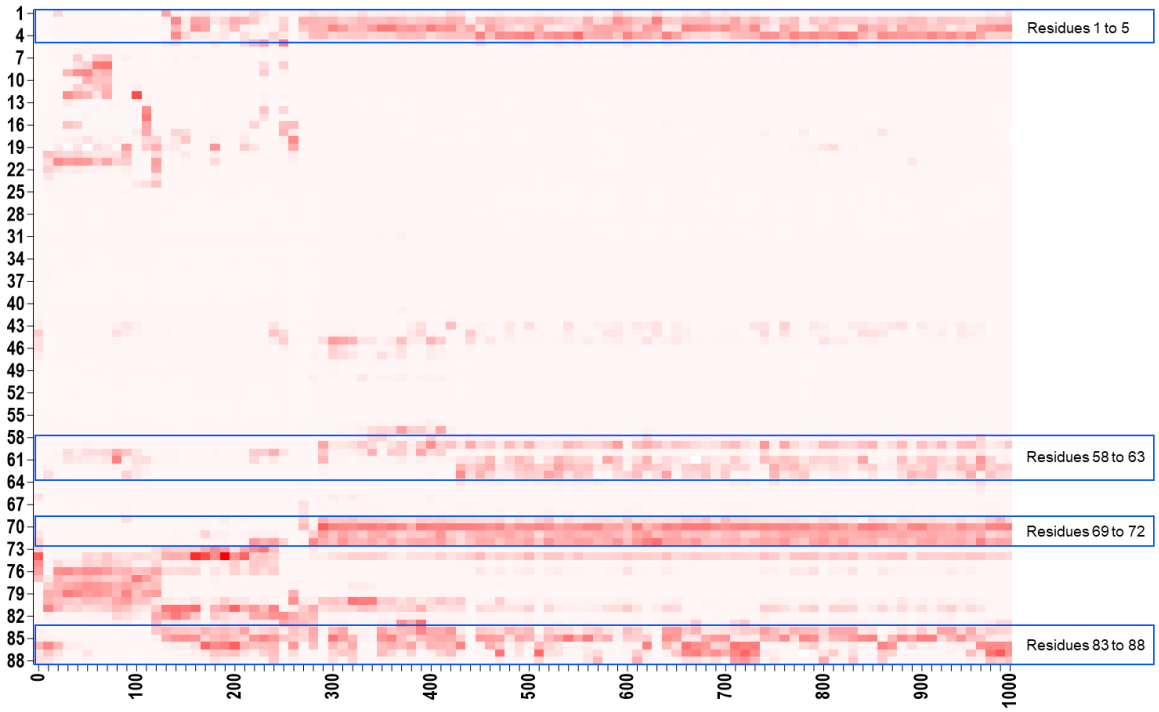

Figure 39 – The MYC88 contact map for ligand 23251632 which shows that the compound binds consistently throughout the simulation to four main regions highlighted in blue frames: N-terminal region - residues 1 to 5; phosphodegron region – residues 58 to 63 and the C-teminal regions – residues 69 to 72 and residues 83 to 88.

Figure 39 results show that the ligand stably and consistently binds four MYC88 regions: residues 1 to 5; residues 58 to 63; residues 69 to 72; and residues 83 to 88. This sees the ligand binding together N-terminal residues to C-terminal residues, involving the phosphodegron as well, which keeps the protein in a folded configuration. The ligand occupancy of the phosphodegron region is likely to negatively affect any interactions that could lead to c-MYC activation. The protein activation is further negatively impacted by the

abrogation of the N-terminal extension which interferes with MB0's recruitment of PIN1, a molecular interaction crucial for c-MYC transcriptional activation and this likely causes c-MYC inactivation - a very desirable outcome in cancer.

The 'druggability' proof-of-concept achieved with ligand 23251632 is in no way unique but reflects the findings obtained with the other 6 ligands. **Figure 40** offers the K-means clustering landscape for the ligand 358383345 along with the representative centroid structures.

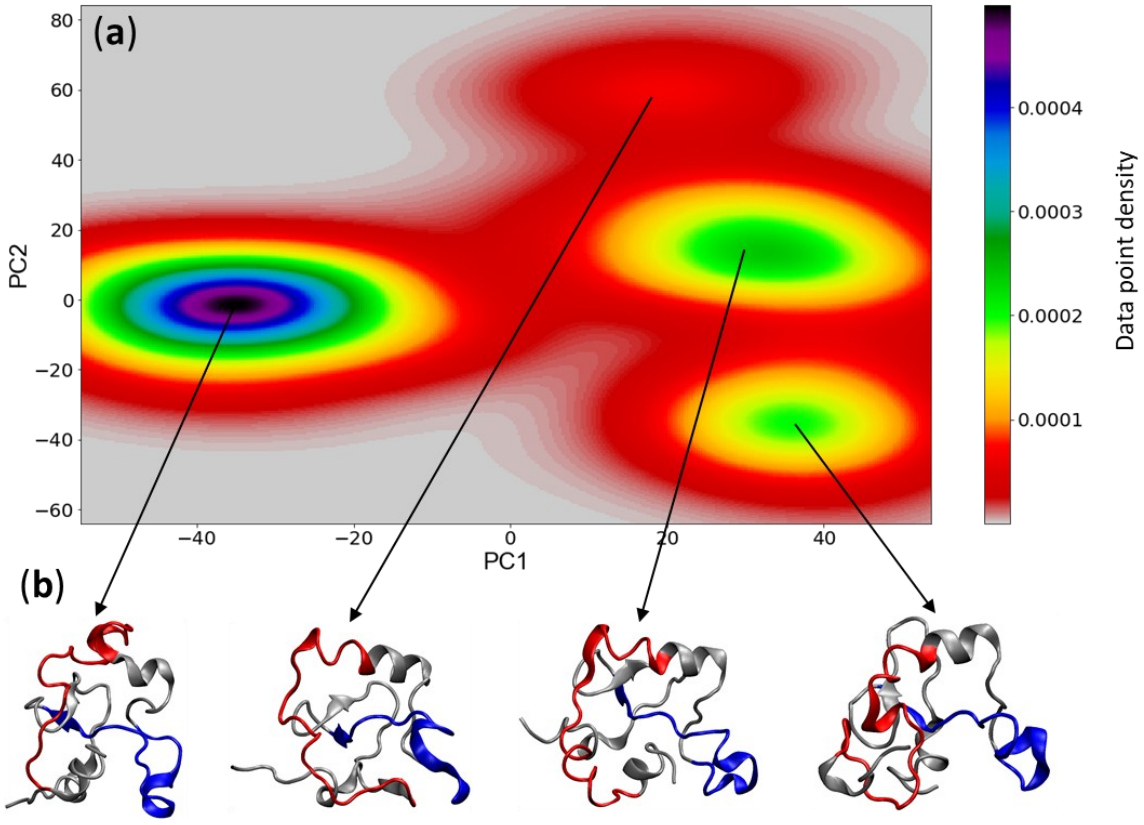

Figure 40 – PCA landscape obtained from the simulation of MYC88 targeted by ligand 358383345. It shows the allocation of the 4 predicted clusters using the K-means clustering algorithm alongside the representative structures for each of the cluster centroids.

It is immediately evident upon inspection that the representative structures obtained for the ligand 358383345 closely replicate the findings obtained for ligand 23251632: the centroid structures are too folded, and the N-terminal does not appear extended in any of the clusters. Further investigation with the minimum and maximum Rg peaks reveals that only two

maximum Rg of gyration peaks, occurring at 366 ns and 391 ns, involve the extension of the N-terminal (**Figure 41**). For the remainder of the trajectory, the N-terminal remains firmly compacted.

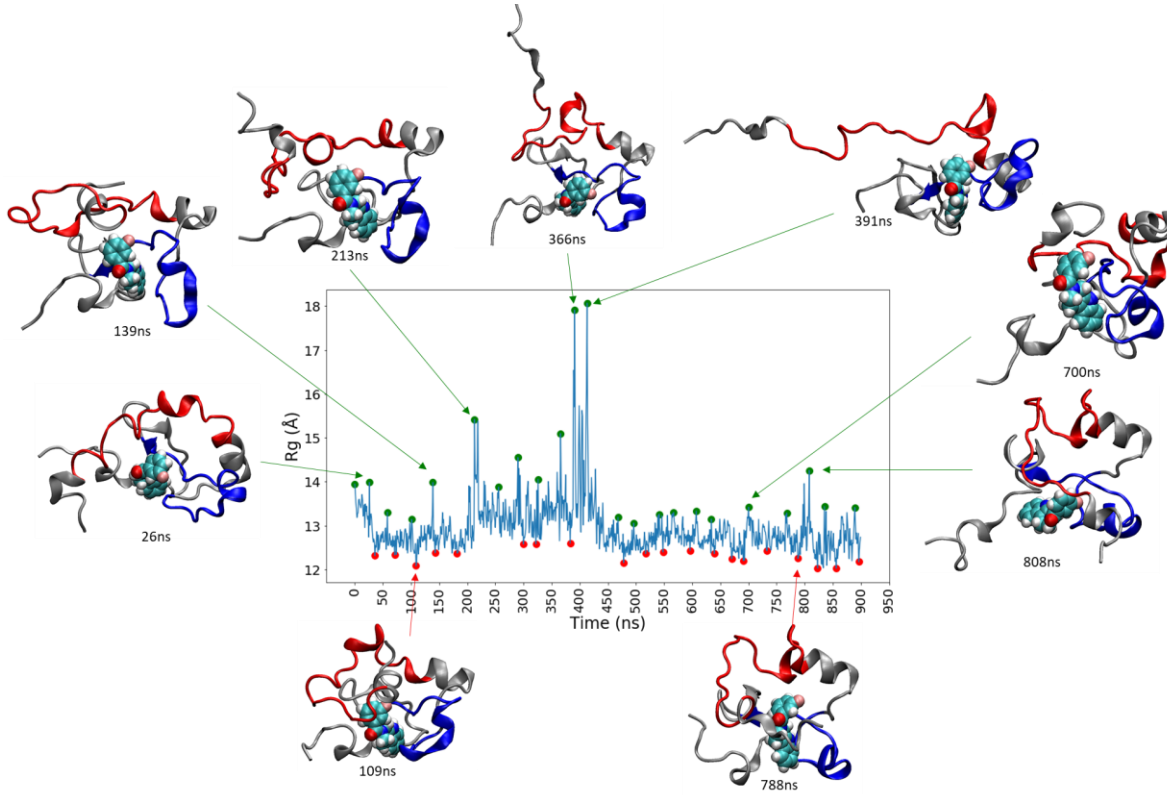

Figure 41– Timeline evolution of MYC88's Rg showcasing the peak minimum and maximum structures. The figures display the ligand and the location of both MYC boxes: MB0 in red and MBI in blue.

The reason for the two mid-trajectory extension events is obvious when looking at the contact map between MYC88 and ligand 358383345 - the max N-terminal extension events coincide with the period during which the ligand transiently unbinds the N-terminal' residues 2 to 13 (**Figure 42**).

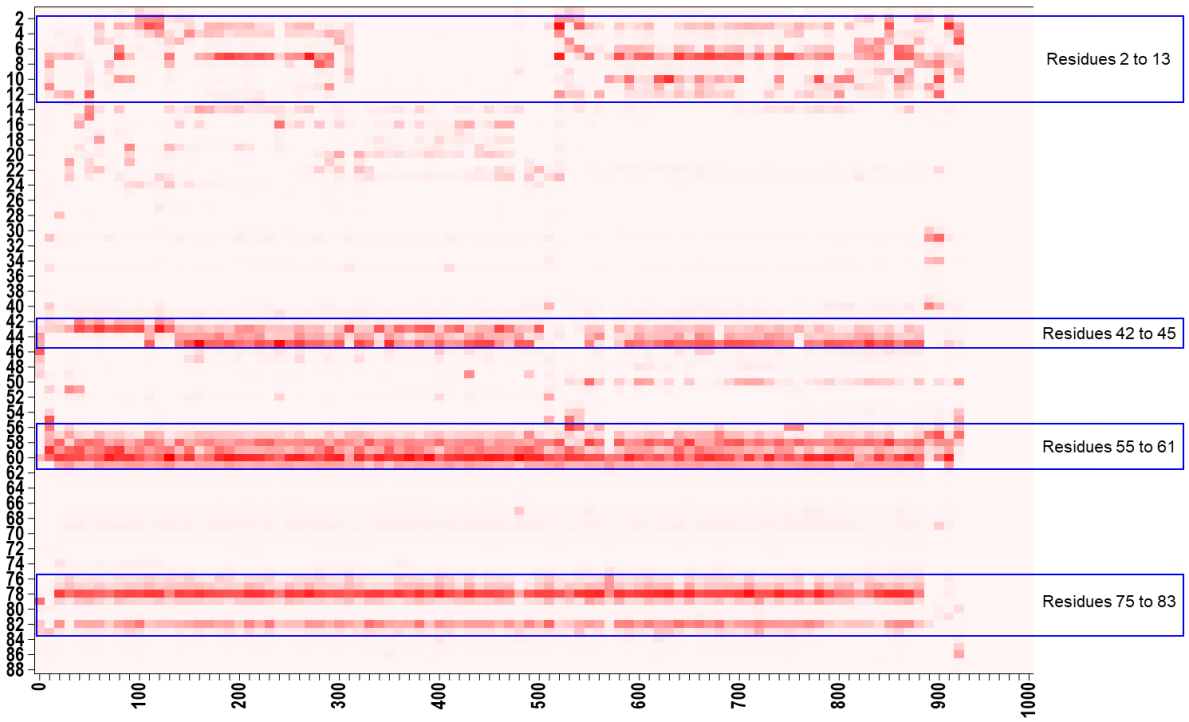

Figure 42 – Timeline evolution of MYC88's interaction with ligand 358383345.

The ligand transient unbinding occurring between 310 and 500 ns leaves MYC88's N-terminal momentarily free and able to extend, causing the maximum peak of extension observed at 366 and 391 ns. Upon reassociation of the ligand to the N-terminal, occurring at the 500 ns mark, MYC88's N-terminal 'fly-casting' activity is once more abolished and the protein becomes once again trapped in a compacted configuration. This establishes that the loss of the structural mobility of the N-terminal is a direct result of the ligand's interaction. Furthermore, **Figure 42** also confirms that ligand 358383345 strongly binds to the phosphodegron region and is also likely to interfere with c-MYC activation.

These findings reveal a very exciting pattern of interaction between the ligands and c-MYC which leads to an extremely desired outcome – to tackle c-MYC overexpression by preventing, and interfering with, its activation in cancer.

3. Conclusion

This Chapter started with a representative structure from MYC88's most abundant and well-sampled metastable state and the intention to challenge the idea that c-MYC is an undruggable 'black box'. The first step was to search for a druggable pocket within the MYC88 against which to build the drug discovery process. Several methodologies were employed to find the elusive pocket. The results, originating from a variety of distinct tools, recognised the same region as a suitable pocket formed by a cavity with the correct geometry, favourable residue composition and stability over time. This region created mainly by MBI residues, includes the phosphodegron and some C-terminal residues, became the target site for drug screening. The extensive screening process identified suitable compounds that bound MYC88 with excellent binding affinity and displayed exceptional binding stability. These compounds also respect the Lipinski rules and are, thus, expected to display optimal pharmacokinetics and pharmacodynamics, making them strong candidates for additional medicinal chemistry research and drug compound optimisation.

Analysing the MD trajectories of several MYC88-ligand complexes revealed a wealth of information, including the mode by which these ligands interfere with MYC88's structural dynamics. These compounds turn MYC88's noisy and disordered PCA landscape into a 'clusterable' conformational landscape. The ligands cause MYC88 to lose its intrinsic disorder, forcing it to adopt ordered and compact conformational states. This loss of disorder correlates with the abrogation of the N-terminal extension activity and occurs because of direct ligand binding to key N-terminal residues. The ligand traps the N-terminal into a folded, compacted configuration and is likely to have a significant negative impact to MYC88's function. The N-terminal's loss of structural dynamics interferes with the 'fly-casting' recruitment of molecular partners and makes MB0 docking site unavailable. Molecular partners, such as PIN1, become unable to activate c-MYC for transcriptional activity. With impaired activation it is probable that c-MYC-driven carcinogenesis is negatively impacted since the ligands also stably bind to the phosphodegron region, further limiting access to c-MYC's activation switch – SER62.

Ultimately, this Chapter provides strong evidence that c-MYC's TAD domain can and should be successfully considered for drug discovery. It offers robust proof-of-concept that it is viable, and even advantageous, to target c-MYC regions other than its DNA binding domain. These conserved c-MYC regions are less likely to trigger compound cross-reactivity since they are highly specific to MYC family of proteins. The findings in this Chapter can help jumpstart further enquiry into TAD domain drug discovery and highlight how the deployment of robust Molecular dynamics simulations and *in silico* drug discovery tools can tackle previously intractable IDPs, confirming that c-MYC should no longer be deemed an undruggable 'black box'.

Chapter IV – c-MYC TAD domain

1. Introduction

The focus of this Chapter is on the study of c-MYC's first 150 amino acids (MYC150), which contain the entirety of its TAD domain. This provides validation for the MYC88 findings and, most importantly, offers further insight into c-MYC's epicentre of interaction with its molecular partners – its TAD domain. These interactions, and how they are modulated, the molecular mechanisms behind c-MYC's interaction decisions remain poorly understood, which strictly frustrates the search for solutions to c-MYC's overexpression.

MYC150 contains the first three highly conserved MYC boxes. This includes MB0 (residues 10 to 32), MBI (residues 45 to 63) and MBII (residues 128 to 143). As described, MB0 has been recently identified as the target docking site for a prolyl isomerase enzyme – PIN1 (Helander et al., 2015) and as an independent transactivation domain (Zhang, Q. et al., 2017). Prior to that, MB0 was not even described in literature and the TAD domain was solely defined in terms of its two other MYC boxes – MBI and MBII. Our research into MYC88 finds that MB0 acts as an independent, autonomous flexible region, engaged in periodic extensions and reassociations which serve as a fly-casting mechanism. In our MYC88 studies, MB0 structural dynamics account for MYC88's highest amplitude conformational changes. The most notable of MB0's interactors - PIN1 has been identified as a molecular switch for a variety of substrates. It is known that Pin1-mediated *cis-trans* isomerization of prolines neighbouring the phosphodegron residues, THR58 and SER62, regulates c-MYC's activation and degradation and, therefore, dictates c-MYC's fate. c-MYC's proliferative activity is activated by phosphorylation of SER62, whilst the subsequent phosphorylation of THR58 causes c-MYC to be flagged for proteasomal degradation (Hann, 2006). PIN1 is thought to mediate the phosphodegron phosphorylation and dephosphorylation events and regulate c-MYC's interaction with different kinases and phosphatases (Helander et al., 2015). This regulatory

PIN1 function is further evidenced in cancer, as cancer-induced phosphodegron mutations result in c-MYC aberrantly maintaining its transcriptional activated state (Bahram et al., 2000; Wang, X. et al., 2011). PIN1 in c-MYC-driven cancers often functions only as a co-activator and loses its function as a c-MYC degradation promoter (Farrell & Sears, 2014). This highlights how important it is to understand these regulatory molecular interactions. How the PIN1 control is achieved and how c-MYC recognises and coordinates its TAD domain dynamics in response to its many co-factors is still very inadequately understood. This final Chapter aims to investigate the structural dynamics of the entire TAD domain to reveal more about key residues involved in the MYC150's structural dynamics and how these relate to mechanisms that control and modulate c-MYC's activity.

2. Results and Discussion

3.1 Exploring the MYC150 structural dynamics

This sub-Chapter seeks to validate the results obtained for MYC88 in terms of its clustering and noise reduction data analysis and applies the same analytical methods. Firstly, it looks at the MYC150 landscape created by simple geometrics including radius of gyration (Rg) and the distance end-to-end (from the first to the last residue) for insight into the protein's compactness; solvent-accessible surface area (SASA); and the molecule's hydrogen bonds, to explore the protein's stability. The results in **Figure 43** indicate that MYC150 mimics the results obtained for MYC88 and fails to create a 'clusterable' landscape with any of its simple metrics. Of course, this is unsurprising given that MYC150 is larger and more complex than MYC88, and more residues means more degrees of freedom, and ultimately noise.

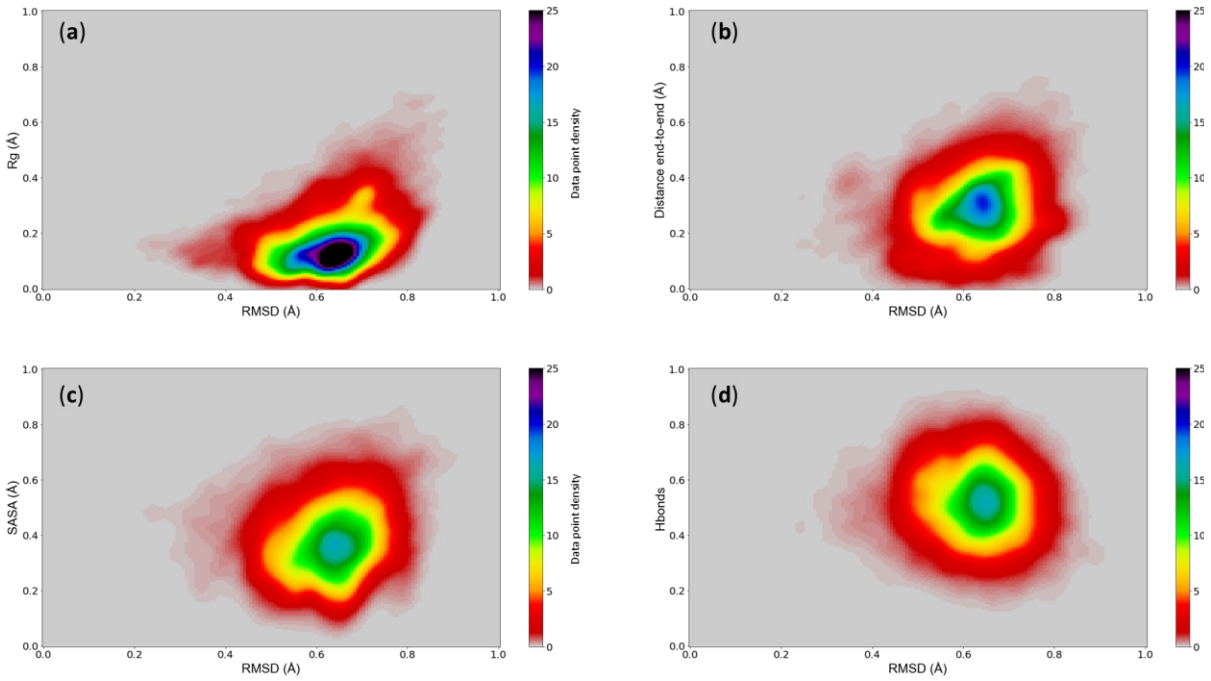

Figure 43 – Normalised MYC150 landscapes obtained by plotting RMSD values against different simple MD simulation metrics: (**a**) radius of gyration (Rg), (**b**) the molecule's distance from N-terminal to the C-terminal (Distance end-to-end), (**c**) solvent-accessible surface area (SASA) and (**d**) the number of hydrogen bonds (Hbonds).

Again, TICA analysis is deployed to give insight into the major average metastable states the protein might be inhabiting. TICA analysis for MYC150 (**Figure 44 - B**) reveals a much flatter landscape when compared to MYC88's landscape (**Figure 44 - A**).

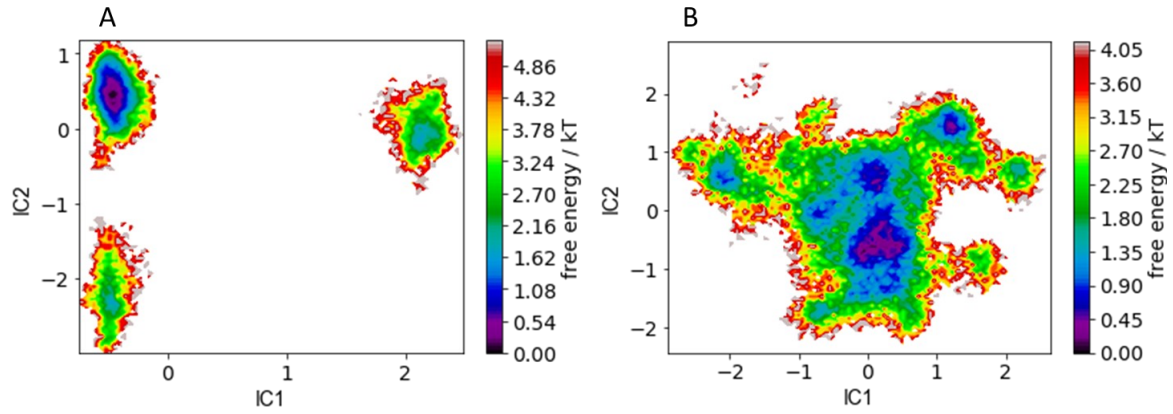

Figure 44 – Free energy plots revealing the conformational basins created by the first two ICs after TICA analysis for both the MYC88 (**A**) and the MYC150 (**B**).

Unlike MYC88, the MYC150 conformational space does not inhabit well-defined, and clearly

separated, energy basins and contains a very large conformational basin with several local minima. MYC150 rapidly moves between a substantial ensemble of conformations separated by very small energy barriers. An attempt at clustering revealed the predicted structures for each of the states including the large basin (**Figure 45**).

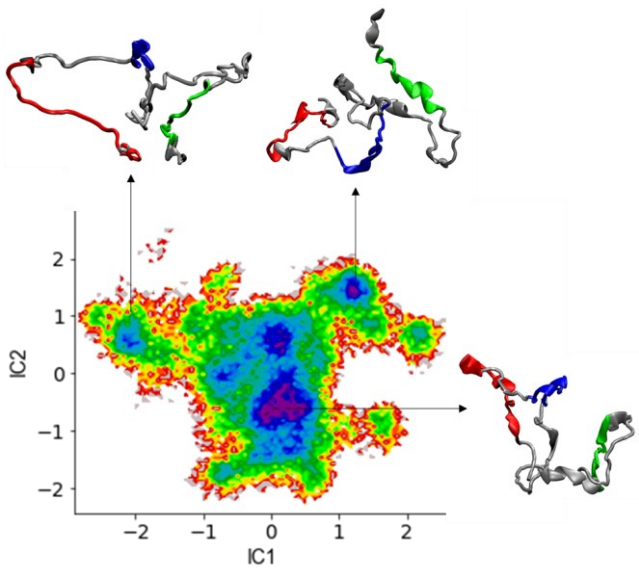

Figure 45 – Conformational averages for each basin mapped onto the TICA free energy landscape. The representative structures show MB0 in red, MBI in blue and MBII in green.

Since this is such a flat landscape, several representative structures were extracted for each basin, and then averaged to reveal a clearer description its conformational range. Hence, the average structure should not be taken a true conformation but rather as an indication of the representative structures' mean. The results show that the largest basin corresponds to structures with the highest compaction, which includes the folding of each of the three MYC boxes. The upper left cluster (IC1 of -2) corresponds to structures exhibiting mostly N-terminal and MB0 extension - consistent with the findings obtained for MYC88. The upper right cluster (IC1 of 1.5) reveals a new pool of conformations characterised mostly by flexibility and extension associated with the C-terminal and the MBII region.

Given the evident limitations of the TICA analysis, these results were validated by the analysis of the linear Rg over time to find the conformational maximum and minimum peaks, affording a reliable way to gain insight into MYC150's highest amplitude transitions (**Figure 46**).

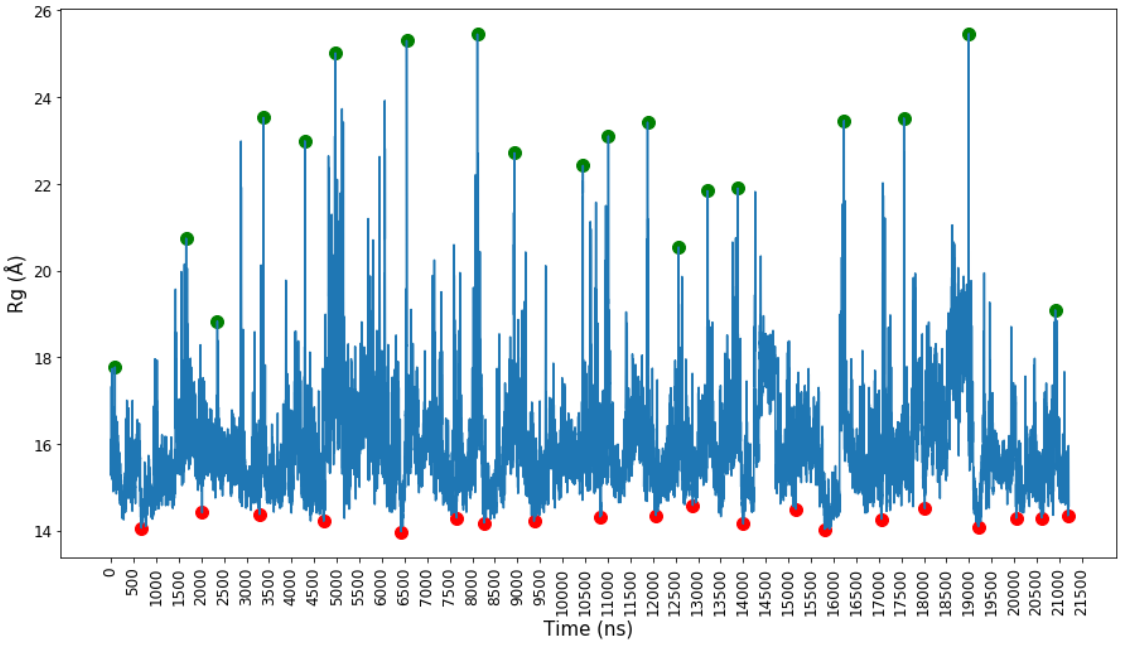

Figure 46 – MYC150 Rg evolution over time showing the identified minimum and maximum peaks.

As with MYC88, MYC150's Rg evolution over time also reveals that the protein undergoes frequent and periodic sampling of maximum and minimum peak conformations. **Figure 47** maps the max and min peak conformations onto the Rg linear plot to identify their structural features.

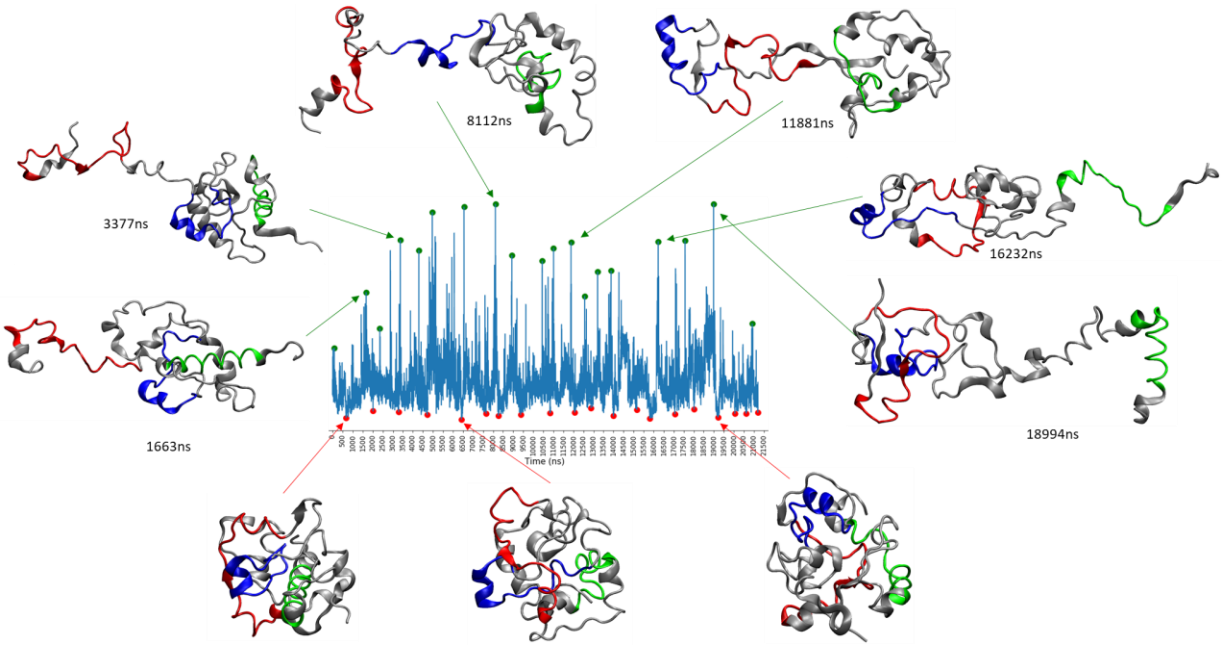

Figure 47 – Representative structures of some maximum and minimum peaks sampled by MYC150 throughout the MD simulations. MB0 is depicted in red, MBI in blue and MBII in green.

The max peak structures, colour-coded with the MB0 in red, MBI in blue and MBII in green, show that the maximum extension and flexibility occurs due to three main structural motions: (1) when the N-terminal and the MB0 extends out from the protein which can be seen at 1663 ns and 3377 ns; (2) when MBI displays an extended configuration at 8112 ns and (3) when the C-terminal and MBII extends out which can be seen occurring at 16232 ns and 18994 ns.

As previously observed for MYC88, MYC150 also undertakes periodic peak configurations involving transitions from compacted to extended, followed by a reassociation. In MYC150, however, the extension events involve each of the MYC boxes separately. This suggests a dynamic mechanism which aims to maximise interaction with molecular partners and to promote the formation of molecular complexes with these conserved regions. Since each MYC box seems to be extending independently, MYC150's dynamics serve the purpose of selecting molecular partners by modulating access to the binding site – making it accessible during extension and inaccessible during compaction.

MYC150 conformational selectivity can be assessed by looking at the long-range intramolecular interactions involving the MB regions. **Figure 48** maps the MYC150 long-distance internal connectivity with a 10-angstrom cut off, considering only very close contacts between residues.

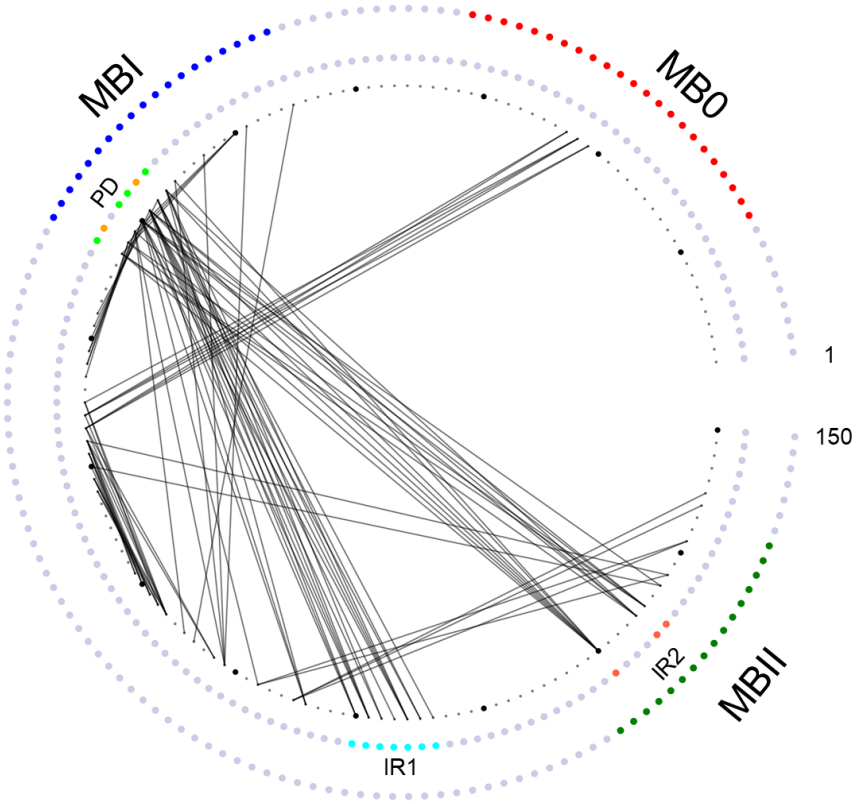

Figure 48 – MYC150 long-range interaction contact maps. MB0 is depicted in red, MBI in blue and MBII in green. The phosphodegron (PD)'s prolines are in marked in light green, the T58 and S62 in orange. The conserved interaction region 1 (IR1) is coloured in cyan and the interaction region 2 (IR2) in salmon.

The contact map shows that the phosphodegron region mediates MYC150's long-range intramolecular interactions. The phosphodegron is in direct contact with two main interaction regions: interaction region 1 (IR1) which corresponds to a conserved region outside of the MYC boxes and interaction region 2 (IR2) a highly conserved region inside MBII. The network centrality of the phosphodegron, previously identified for MYC88, suggests that the region is more than a phosphorylation switch but mediates the structural dynamics of the unphosphorylated protein as well. This can be further assessed with network analysis to determine the most important residues in terms of centrality (**Figure 49**).

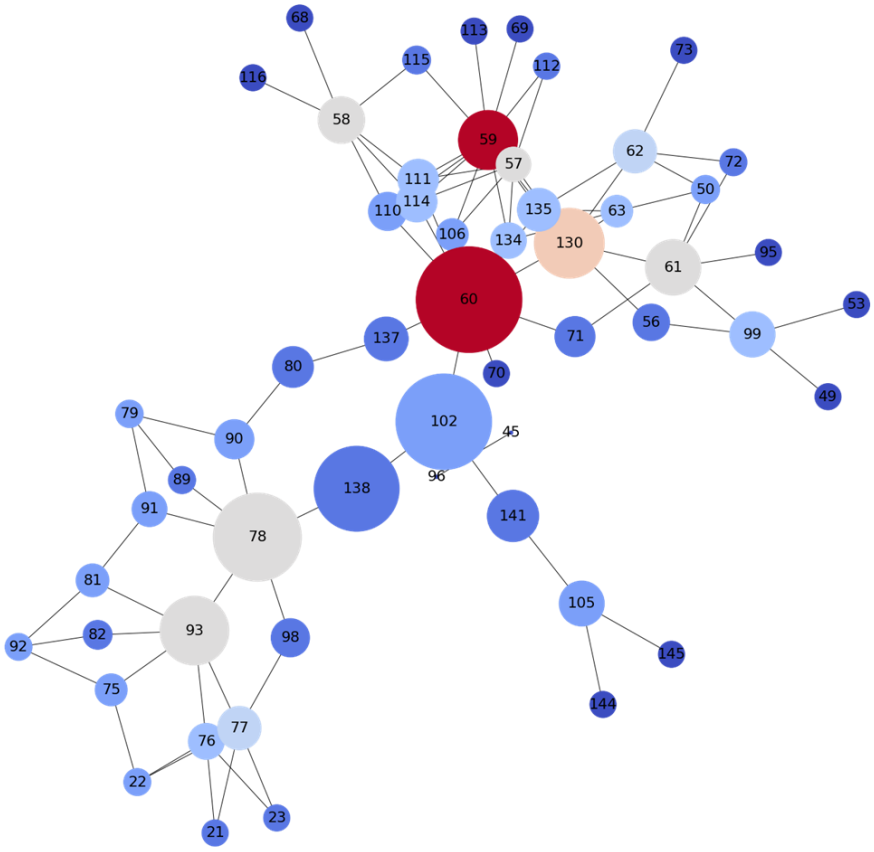

Figure 49 – MYC150 network analysis. The nodes are coloured by their degree centrality and sized according to their betweenness centrality.

The main hub in the MYC150 network is PRO60, at the centre of the phosphodegron, displaying both the highest degree and the highest betweenness centrality. This makes it the most connected residue and the main orchestrator between the different regions of the network. PRO60 is closely followed in importance by another phosphodegron proline – PRO59 with an equally high degree centrality, the same 11 connections as PRO60 (**Table 8**).

Table 8 – MYC150 network centrality measures for the highest-scoring residues.

Residue	Degree centrality (edges)	Betweenness centrality (score)
PRO60	11	0.54
PRO59	11	0.145
ILE130	7	0.22
LEU61	6	0.123
THR58	6	0.07

Out of the 5-top scoring residues in terms of degree centrality - four are phosphodegron residues, including THR58, indicating that the phosphodegron is key for the long-range interconnectedness of MYC150. The main recipient of the phosphodegron connections is the MBII residue ILE130, highlighting that the connectivity between MBI and MBII is of central importance to establishing MYC150 pattern of intramolecular interactions.

Considering the protein's structural dynamics as information flow systems, the central residues identified by the network analysis constitute main hubs in the information transmission between different parts of the system. These key network residues, involved in network information modulation, are often crucial for protein folding regulation (Atilgan, Akan & Baysal, 2004; Nikolay V. Dokholyan et al., 2002; Vendruscolo et al., 2002), found in binding sites controlling interactions with other proteins (del Sol & O'Meara, 2005; del Sol, Fujihashi & O'Meara, 2005) and associated with the active site in several enzymes (Amitai et al., 2004). The identification of MYC150's functionally important residues is, again, pointing towards the importance of the phosphodegron, making it is crucial to further assess the role of the phosphorylation for MYC150's structural dynamics.

3.2 MYC150 phosphorylation and mutagenesis

The first step in assessing the effects of the phosphorylation is to study the Rg linear graph showing the maximum and minimum Rg peaks for the three conditions: MYC150; MYC150 phosphorylated at SER62 (pSER62); and MYC150 with phosphorylated THR58 (pTHR58) (**Figure 50**).

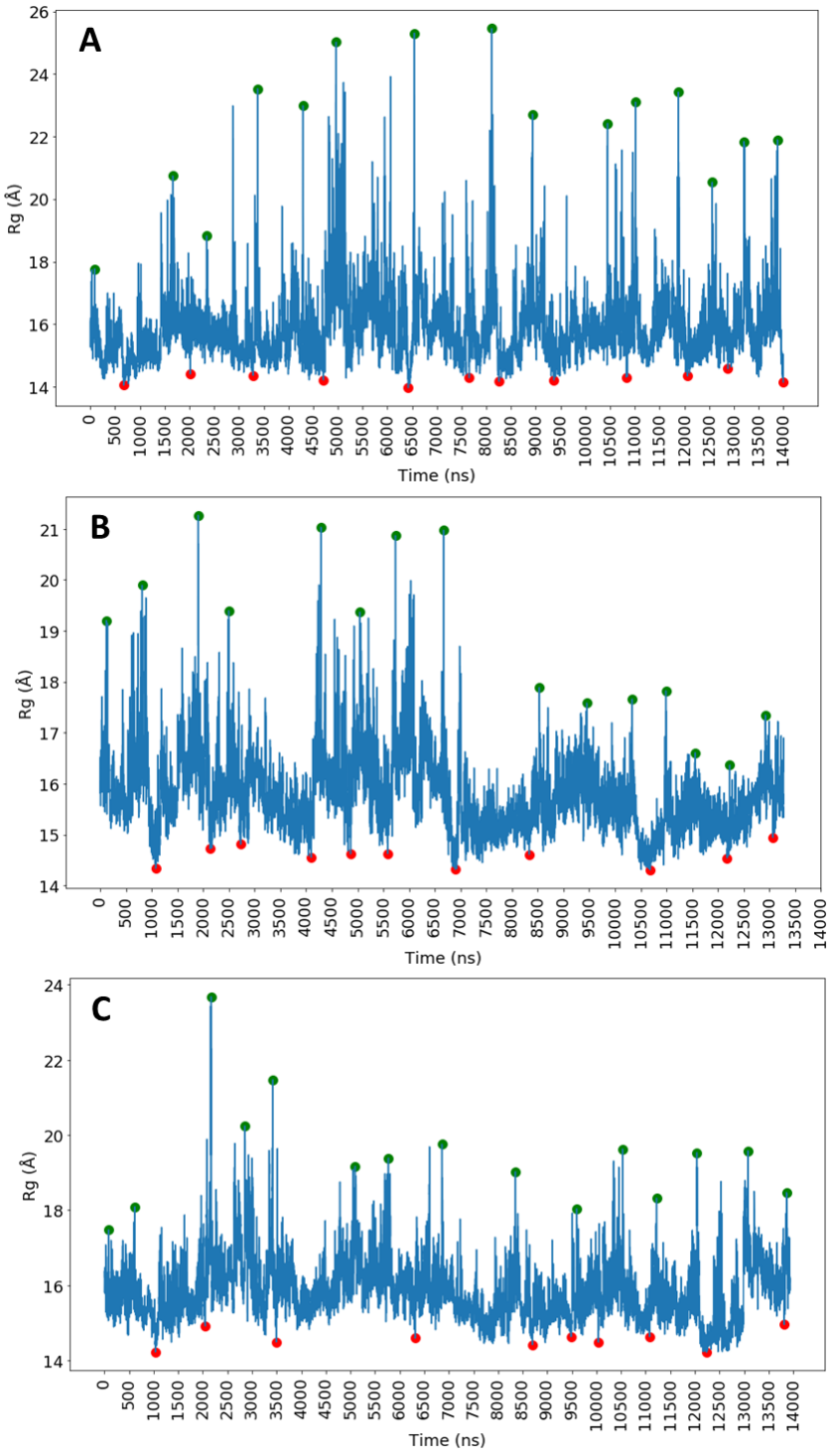

Figure 50 – Rg evolution over time for (A) MYC150, (B) pSER62 and (C) pTHR58. The graph identified the minimum peaks in red and maximum in green.

Observing the three plots it is immediately obvious that the simulations containing the phosphorylated residues, both pTHR58 (**Figure 50 - C**) and pSER62 (**Figure 50 - B**)

simulations, display a decreased number of maximum peak configurations when compared to the unphosphorylated MYC150 (**Figure 50 - A**). The count of maximum Rg peaks events, defined as Rg values above 20 Å, reveals that MYC150 displays a higher frequency of maximum peak events – 137 (**Table 9**). This when compared to pSER62, which only has 5 maximum peaks, and pTHR58, displaying 14 maximum peaks, suggests that phosphorylation of either residue interferes with MYC150's conformational flexibility and its ability to achieve its periodic states of maximum extension.

Table 9 – Descriptive statistics for the Rg over time data concerning the MYC150 simulations and the phosphorylated simulations: pSER62 and pTHR58.

Simulation	Number of Rg peaks (>20 Å)	Highest Rg (Å)	Rg mean (Å)
MYC150	137	25.5	15.9
pTHR58	14	23.7	15.8
pSER62	5	21.3	15.7

Considering the range of Y-axis values for each of the **Figure 50**'s plots, makes it evident that even when the pTHR58 and pSER62 trajectories sample the extended peaks their extension range does not compare to the range sampled by MYC150. The maximum extension Rg value for MYC150 is 25.5 Å; whilst pSER62 only achieves a maximum Rg conformation of 21.3 Å; and pTHR58 a maximum of 23.7 Å (**Table 9**).

It is interesting to note that for pSER62 the maximum Rg peaks are caused by a modest extension of MB0, whilst the other two MYC boxes remain largely folded and compacted. Whilst for pTHR58 the maximum peak of extension does not involve the extension of any MYC boxes, which remain all quite compacted. To visually highlight this, **Figure 51** presents the most unfolded structures for each simulation phosphorylation condition.

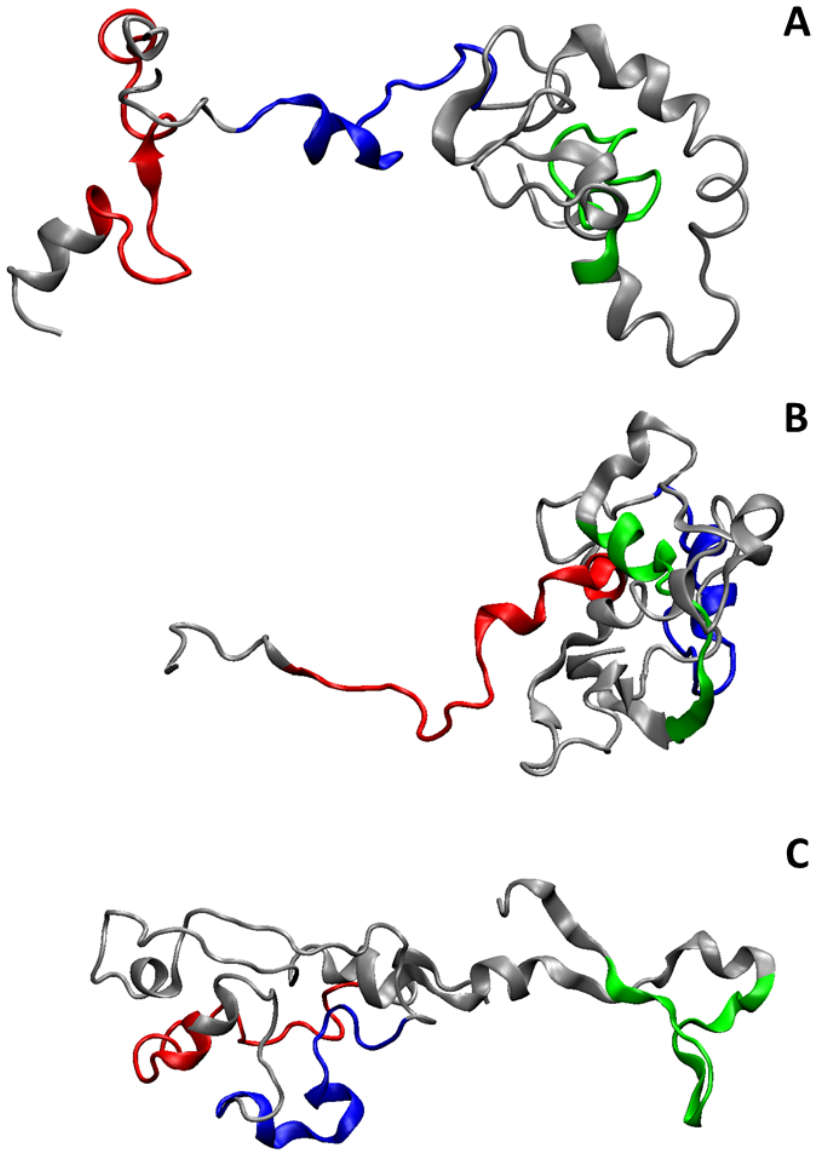

Figure 51 – Maximum Rg peak configurations for (A) MYC150, (B) pSER62 and (C) pTHR58. The figures highlight the MB0 in red, MBI in blue and MBII in green.

SER62 phosphorylation drives a dramatical loss, in the number and amplitude, of unfolded peak events - these become exclusively caused by a modest MB0 extension motion whilst the other MYC boxes remain stably folded. It is important to note that although pSER62 displays a reduction in maximum Rg peak events and becomes stabilised, the mean Rg does not seem to be affected, the pSER62's Rg mean of 15.7 Å is very close to MYC150's Rg mean of 15.9 Å (**Table 9**).These findings are in line with current research by (Helander et al., 2015) which suggests that upon SER62 phosphorylation, the binding patterns of MB0 to PIN1 are maintained and this is the main reason why, in pSER62 simulations, the MB0 extension motion

continues to be observed and accounts for the maximum Rg events. PIN1's interaction continues to be crucial even after the phosphorylation of SER62 because PIN1 also mediates the pSER62 dephosphorylation, which together with the phosphorylation of THR58 flags the protein for destruction.

Interestingly, the pTHR58's dynamics also show a remarkable decrease in the number of maximum Rg peak that does not seem to interfere with the overall intrinsic disorder - the mean Rg of pTHR58 is 15.8 Å and comparable to MYC150's Rg mean of 15.9 Å (**Table 9**). Considering pTHR58's highest Rg value structure in **Figure 51 – C,** it is curious to note that even its most unfolded conformation displays highly compacted MYC boxes. This suggests that upon THR58 phosphorylation, since the protein is marked for degradation, the decrease in the MYC boxes' binding surface due to folding and compaction makes them unavailable for further intermolecular interaction. This raises the idea that specific phosphorylation patterns change the overall MYC150's structural dynamics involved in modulating local accessibility to the MBs and to phosphodegron. The accessibility to the phosphodegron switch residues, THR58 and SER62, is of utmost importance and can be measured in the terms of each residue's solvent accessible surface area or SASA.

The first phosphorylation event occurs at SER62, which activates MYC150 for transcription. Subsequently, a second phosphorylation event at THR58 starts the process of flagging MYC150 for destruction. The process is finalised when PIN1 mediates the dephosphorylation of SER62 at which point only THR58 remains phosphorylated. When this occurs, the protein is finally flagged for destruction via proteasomal degradation. To assess the changes each phosphorylation event causes to the SASA values of both THR58 and SER62, the boxplots in **Figure 52** present the solvent accessibility shifts for both residues.

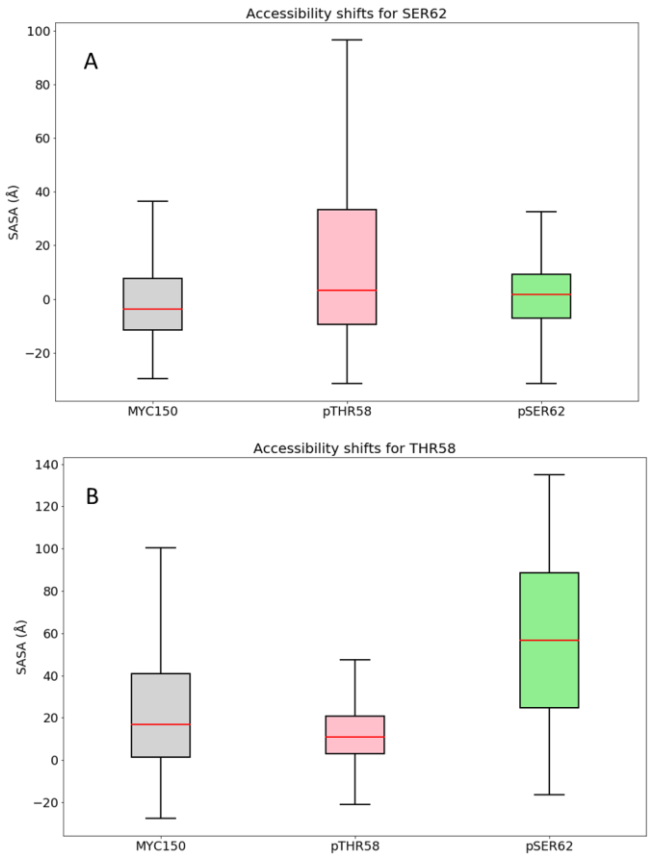

Figure 52 – Boxplots presenting the SASA values for the SER62 (top) and THR58 (bottom) across the different simulation conditions (MYC150 refers to the unphosphorylated protein in grey; pTHR58 refers to the simulations with a phosphorylated THR58 in red; and pSER62 to the simulations with a phosphorylated SER62 in green).

Figure 52 depicts an interesting accessibility modulation pattern involving the phosphodegron residues and caused by the phosphorylation events. **Figure 52 - A** indicates that when SER62 is phosphorylated (green boxplot) its solvent accessible area is very similar to the unphosphorylated SER62 (grey boxplot), however SER62's accessibility dramatically increases when THR58 gets phosphorylated (red boxplot). This is likely caused by the need to dephosphorylate SER62 after THR58 becomes phosphorylated, increasing its accessibility. **Figure 52 - B** shows the opposite effect occurring in terms of THR58 accessibility: when SER62 gets phosphorylated (green boxplot) THR58's accessibility dramatically increases which helps to promote its phosphorylation. On the other hand, when THR58 itself is phosphorylated (red boxplot) its accessibility noticeably decreases - undoubtedly to prevent undue dephosphorylation that would interfere with the protein's degradation. This dynamic switch, facilitated by the phosphorylation events, serves to modulate access to c-MYC's TAD

domain ‘activation’ and ‘destruction’ buttons, SER62 and THR58, respectively. Thus, it is crucial to assess MYC150’s network of intramolecular contacts upon phosphorylation. **Figure 53** presents the contact (**A**) and network (**B**) maps for pSER62.

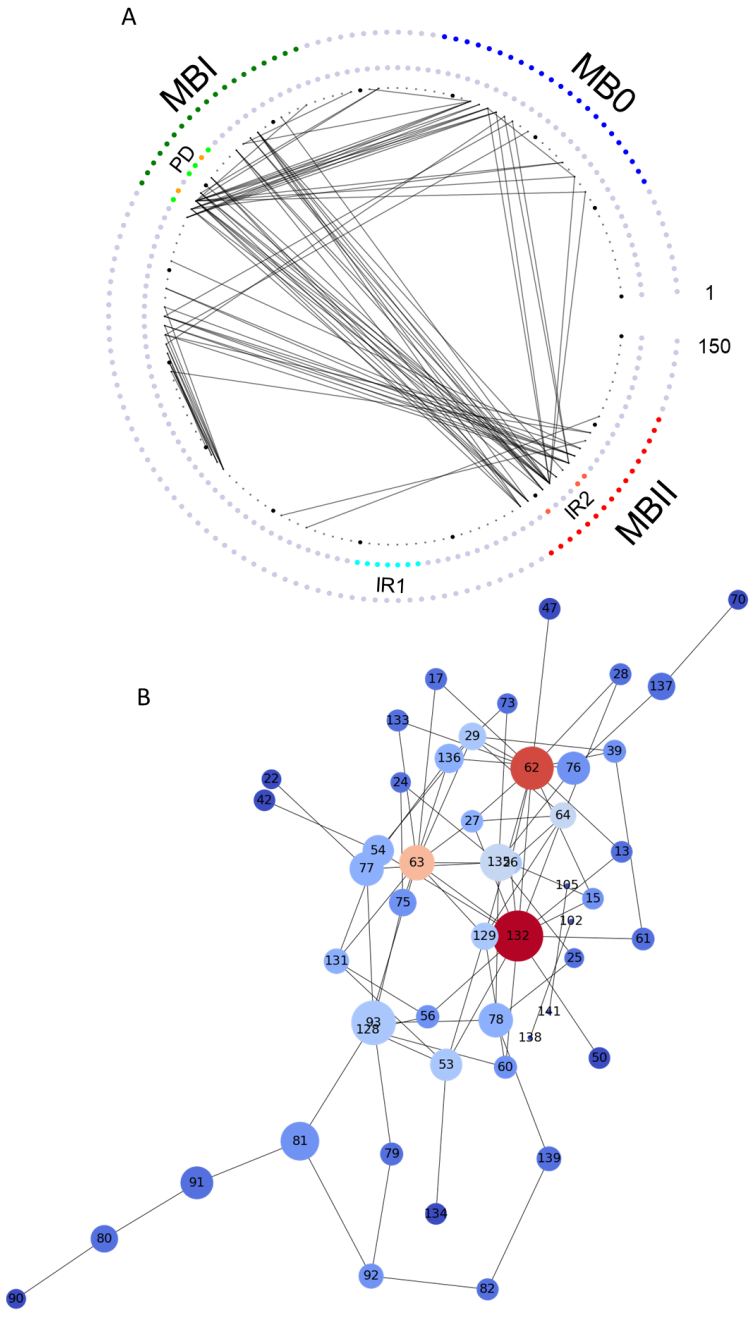

Figure 53 – Long-range internal connectivity (**A**) and network analysis (**B**) for pSER62.

It is evident from the contact map that the phosphodegron remains central to most of the long-range interactions. There is a high prevalence of contacts between the phosphodegron

residues and MBII, which accounts for the higher stability displayed by MYC150 after SER62 becomes phosphorylated. The phosphodegron residues are also in close interaction with MB0 residues, which explains why the MB0 extension frequency is decreased when compared to the unphosphorylated MYC150. Also, when compared to MYC150, pSER62 phosphodegron region displays increased intermolecular contact to the MYC150 IR2 region and abrogates contacts with the IR1 region. Most interestingly, the contact network map for pSER62 (**Figure 53 - B**) identifies SER62 as the most important functional residue alongside MBII residue 132. These two residues become central hubs of connectivity and mediate the contacts between different regions of the network. These residues are at the core of structural dynamics which strengthen the interaction between MB0, MBI and MBI, lead to protein stabilisation, decreased frequency of MB0 fly-catching motion and increased accessibility to THR58.

Similarly, the contact and network assessment for pTHR58 is presented in **Figure 54**. Upon THR58 phosphorylation, it is striking to note the loss of long-range intramolecular contacts displayed by the protein. However, despite a dramatic difference in connectivity pattern, the contact map (**Figure 54** - **A**) shows how, yet again, the phosphodegron is at the heart of the protein's intramolecular interactions, establishing contacts with MB0 and, sparsely, with the MYC150's C-terminal, including some IR2 MBII residues. Most remarkably, the protein's interconnectivity becomes predominantly dominated by short-range local interactions (**Figure 54 – A red box**) and involves low-conserved regions outside any of the MYC boxes. The network map (**Figure 54** - **B**) reveals that upon being phosphorylated, THR58 becomes the most prominent residue in the network displaying an incredible score in degree centrality with a total of 18 connections. The second-best scoring residue in terms of degree centrality (residue 93) displays only half of THR68's connection – a total of 9 network edges. This highlights how critical THR58's role is as a modulator of the pTHR58 contact network.

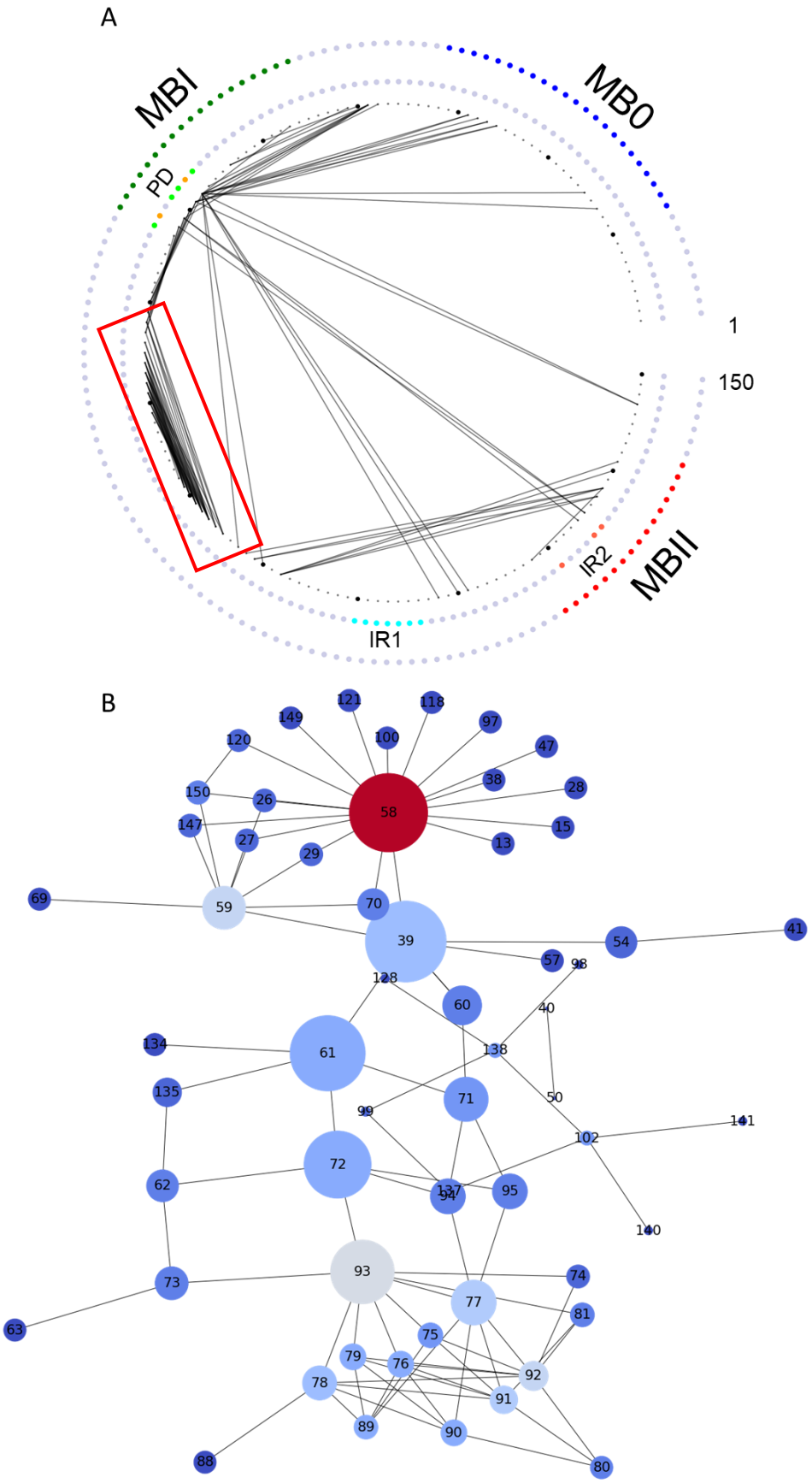

Figure 54 – Long-range internal connectivity (**A**) and network analysis (**B**) for pTHR58.

This finding suggests that when THR58 becomes phosphorylated it also becomes the most central residue in the network of contacts, modulating the structural dynamics to make the MYC boxes unavailable for binding.

The structural changes modulated by phosphorylation at THR58 and SER62 are crucially vital to the protein's function and interaction with other molecules, which would explain why mutations at or near these residues are so deleterious. To further assess the effect of mutations involving either SER62 or THR58, and how these alter MYC150's structural dynamics *in silico* mutagenesis was attempted. In both cases, SER62 and THR58 were substituted by the phosphomimetic residue glutamic acid. This is a best-case scenario, given that the mutation of a phosphorylated residue by a phosphomimetic residue has the potential to replicate the phosphorylated residues' functions. However, despite being a best-case scenario the mutagenesis uncovered the adverse effects underlying altering the phosphodegron's dynamics. **Figure 55** presents the accessibility values for both SER62 and when serine is mutated to glutamic acid (S62E).

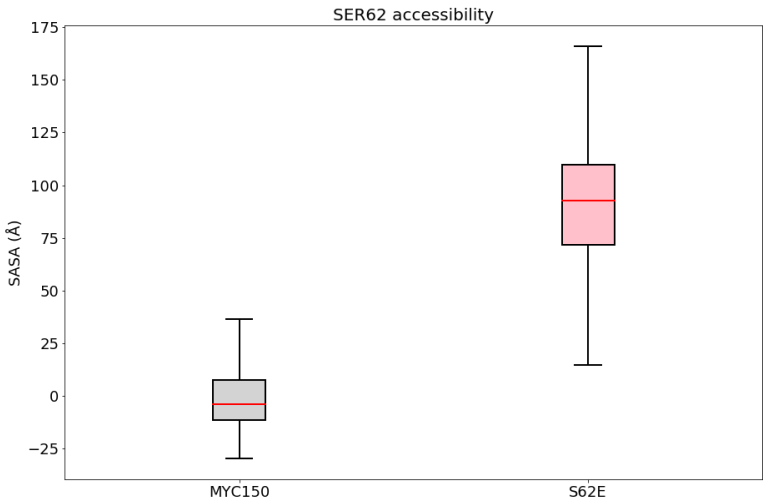

Figure 55 – Boxplots presenting SER62's solvent accessible area range for the unphosphorylated protein (in grey) and when SER62 is mutated to a GLU62 (red).

The S62E (red boxplot) mutation leads to a dramatic increase in the residue's accessibility when compared to the unphosphorylated protein (in grey). This, of course, has grave implications as an increased access to the activation residue is likely to promote

phosphorylation and, consequentially, undue c-MYC transcriptional activation. This explains why phosphodegron mutations cause c-MYC to aberrantly maintain its transcriptional activated state and increase its stability (Chakraborty et al., 2015). They do so by increasing accessibility to the activation switch: SER62. This increase in SER62's accessibility is accompanied and likely caused by the spectacular changes in its contacts and network control (**Figure 56**).

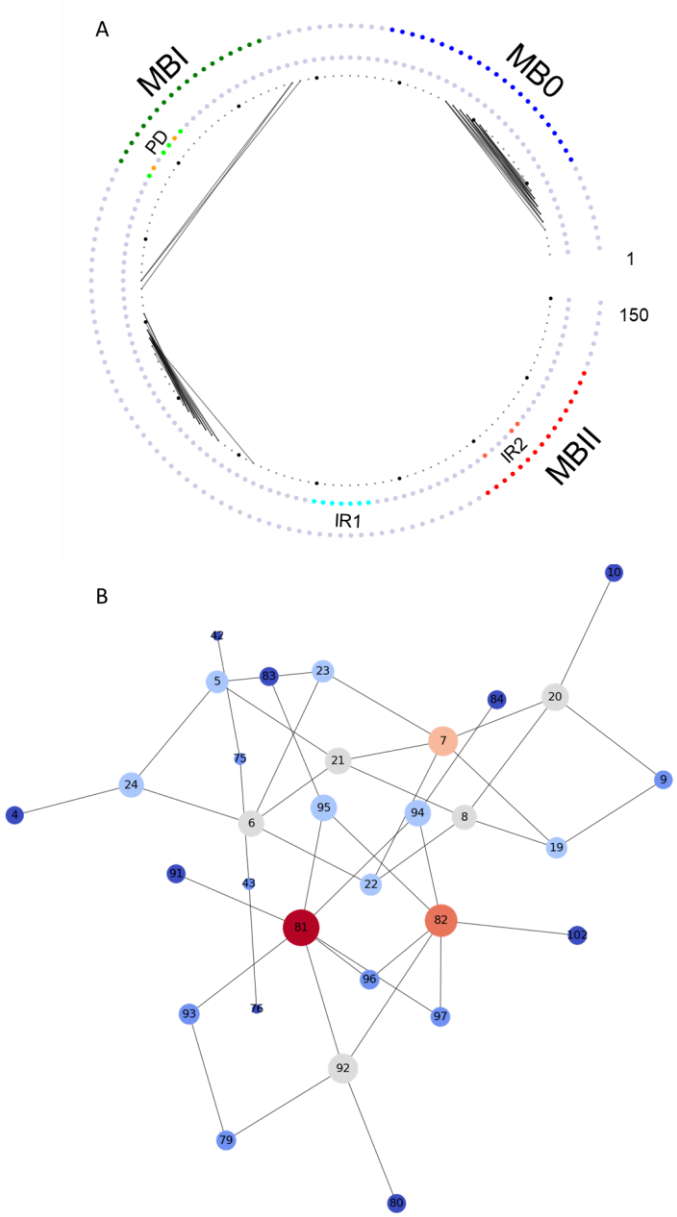

Figure 56 – Long-range internal connectivity (**A**) and network analysis (**B**) for S62E.

The phosphodegron, which in normal circumstances is the focal point in the protein's long-range intramolecular connectivity, loses all control of the protein's contact activity. In fact, the network shows a substantial decrease in long-range contacts and becomes modulated by short-range interactions modulated by unimportant residues, outside any of the MYC boxes. The mutation of T58E delivers similar results (**Figure 57**) creating a completely altered network of contacts and intramolecular connectivity pattern.

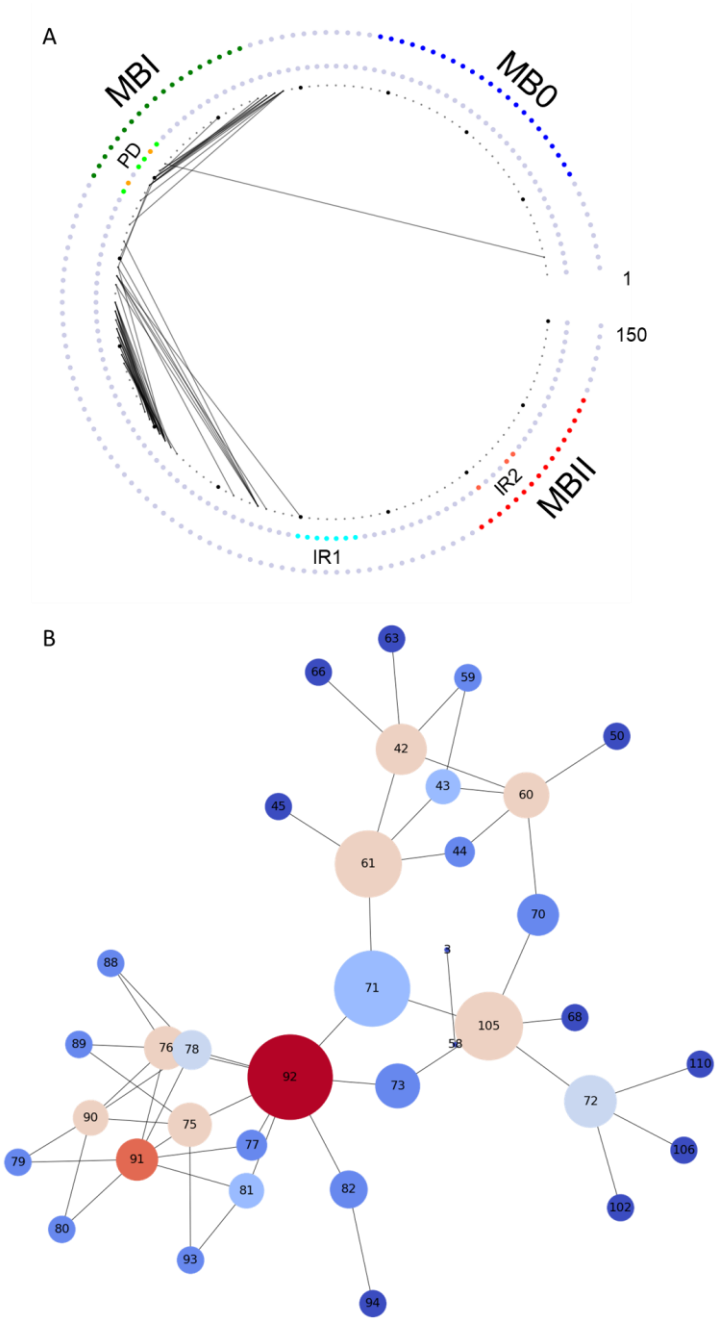

Figure 57 – Long-range internal connectivity (**A**) and network analysis (**B**) for T58E.

In the case of T58E, although the phosphodegron is still involved in some of the protein's contacts, these are mainly local contacts within MBI itself. The phosphodegron loses its long-range connectivity and none of the phosphodegron residues displays high centrality, especially when compared to the unphosphorylated MYC150's PRO59 and PRO60.

Upon investigation into why PRO59 and PRO 60 are so prominent in the MYC150's network, it was found that PRO59 displays a peculiar state of isomerisation. Proline isomerisation has been identified as a molecular timer, or a switch, crucial for the regulation of the c-MYC's biological functions (Helander et al., 2015).

Prolines preferentially adopt a *trans* configuration, although they can often sample the *cis* configuration (Nicholson et al., 2007). This offers a mode of protein regulation, since certain interactors only recognise specific states of prolyl isomerisation (Hamelberg & McCammon, 2009). **Figure 58** presents the PRO59 isomerisation state results, in the form of a psi-omega Ramachandran plot, for the unphosphorylated protein and how it changes when SER62 and THR58 become phosphorylated. The unphosphorylated MYC150 (**Figure 58 - grey plots**) displays a PRO59 isomerisation pattern in which the residue equally inhabits the *trans* and *cis* configurations. The Ramachandran analysis reveals that the THR58 β-region is the most populated, with some sampling of both left-handed and right-handed α-helical conformations. However, this is altered when SER62 becomes phosphorylated (**Figure 58 – green plots**). Not only is PRO59 switched exclusively to the *cis* configuration but also the THR58-PRO59 sees its α-helical content dramatically reduced in favour of exclusive β-region sampling. When THR58 is phosphorylated (**Figure 58 – red plots**), PRO59 remains in *cis* configuration, but the THR58 Ramachandran plot reveals an increase in right-handed α-helical content.

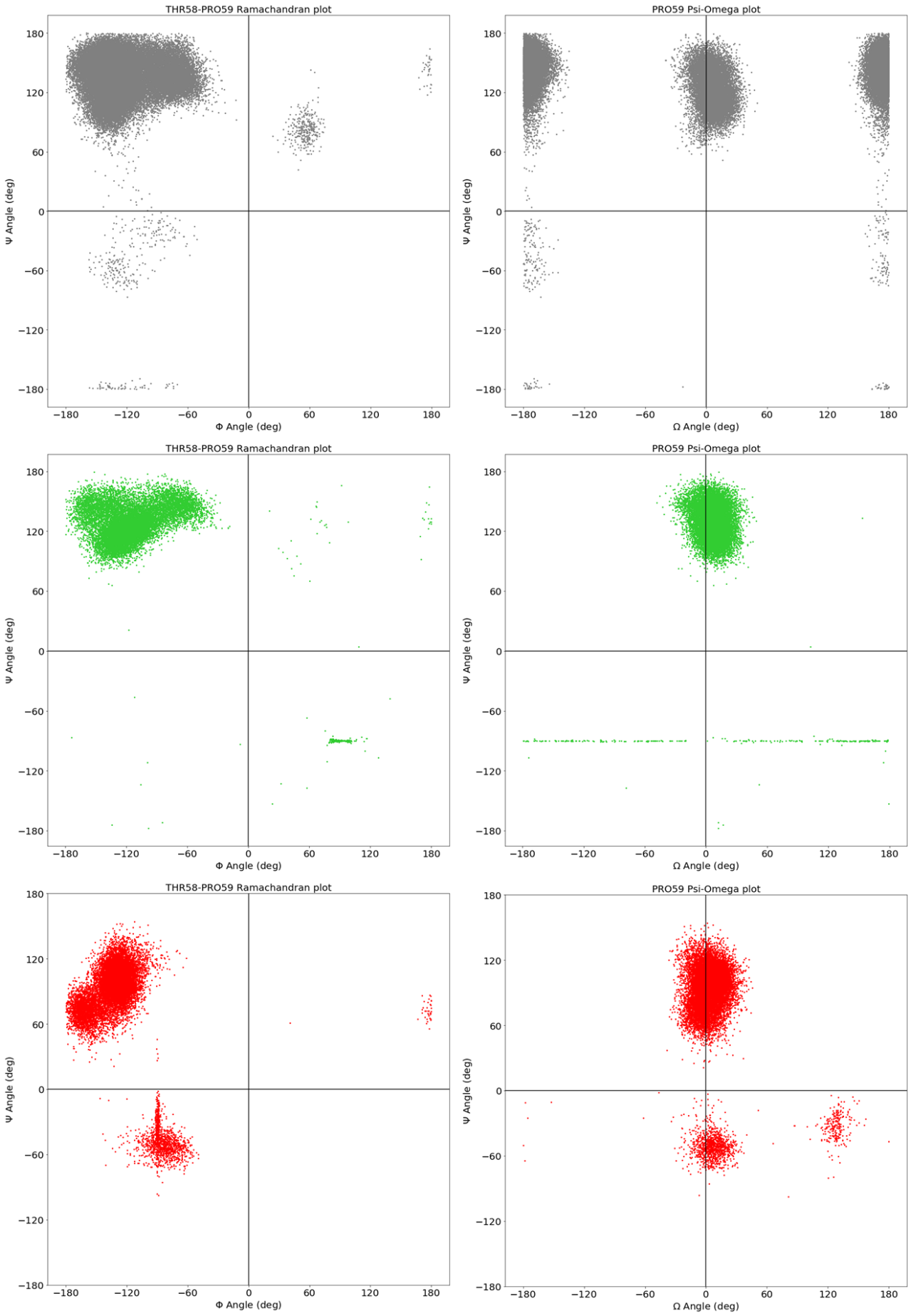

Figure 58 –The Ramachandran and Psi-omega for MYC150 (grey), pSER62 (green) and pTHR58 (red).

These structural dynamic patterns alongside proline isomerisation transitions are very crucial in regulating the processes uncovered in this Chapter. This becomes more evident when considering the dramatic impact of the mutations on the isomerisation arrangement of PRO59 and the Ramachandran plot of THR58 (**Figure 59**).

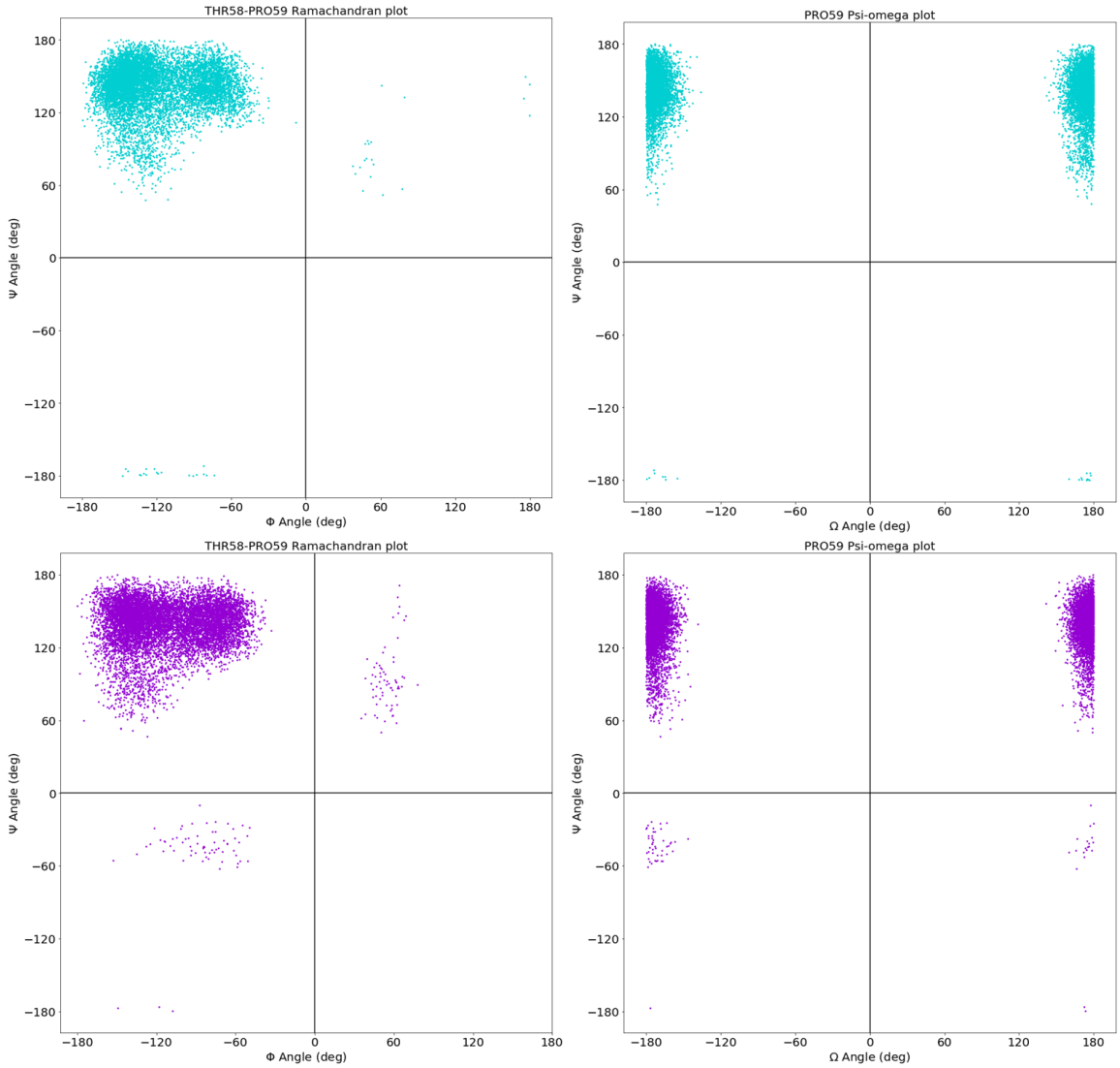

Figure 59 –The Ramachandran and Psi-omega S62E (cyan) and T58E (violet).

Both mutations cause the same exact effect - they both turn PRO59's isomerisation state to *trans* only. This abolishes PRO59's role as a molecular switch to control and mediate the phosphodegron interactions. Since some interactors, including kinases, prefer to bind proteins in the *trans* state, it is not unreasonable to think that having PRO59 switched to the *trans* only configuration, alongside increased accessibility to SER62, leads to amplified interaction with

those specific molecular partners becoming an additional causal factor in the undue, excessive protein activation and/or compromised protein degradation.

3. Conclusion

c-MYC's TAD domain has been, for the most part, unexplored due to its highly disordered nature. Its interaction decisions and the coordination between the protein's activation and degradation pathways were always poorly understood. In this Chapter, c-MYC's TAD domain (MYC150) was found to display frequently and periodic conformational extensions which involved each of the three MYC boxes independently. The MB0 was found frequently extended in a manner consistent with the findings for MYC88 in Chapter II. This was identified as a motion destined to optimise interaction with molecular interactors, notably PIN1, whose binding site is found within MB0. PIN1 is known to bind MB0 and allosterically modulate proline isomerisation events in the phosphodegron prolines in MBI. Curiously, the long-range contact analysis for MYC150 identified phosphodegron prolines 59 and 60 as the most functionally important residues in the network for MYC150's intramolecular interactions. It was also found that specifically PRO59 modulates the phosphodegron activity by acting a switch moving from *trans-cis* to exclusive *cis* conformations upon phosphorylation of either THR58 or SER62. The control of c-MYC activity was found to take shape in terms of accessibility to the all-important phosphoresidues THR58 and SER62. When SER62 gets phosphorylated, the accessibility to THR58 increases which improves its availability for phosphorylation. Vice-versa when THR58 gets phosphorylated the accessibility to SER62 increases since increased accessibility is necessary for the dephosphorylation event to take place, as a final step to protein degradation. These dynamics are crucial and highly depend on the precise residue composition. Mutations of either THR58 or SER62, even by phosphomimetic alternatives, lead to a complete alteration of MYC150's delicate balance of intramolecular interactions and structural dynamics. These alterations are likely responsible for the impairment of c-MYC's degradation, its aberrant activation and increased stability.

Summary and final thoughts

This work uncovers a journey intended to tackle c-MYC, a protein deemed important as it is elusive. c-MYC's intrinsically disordered nature has severely hindered its structural study using conventional experimental methods. This has naturally spurred interest into alternative *in-silico* methods. Molecular dynamics simulations, capable of studying systems with atomistic detail, showed promise in overcoming the experimental difficulties.

To be a true alternative, MD simulations must reproduce the complex workings of a disordered protein system and many methods for simulation setup, which proved their worth in simulating ordered systems, consistently failed to replicate the richness of IDP conformational diversity. Thus, the first step became finding novel ways for MD simulation optimisation. Different molecular dynamics force fields and solvation methods were tested and compared to experimental data. The combination of ff14sconly force field with the Generalised Born 8 implicit solvation model produced highly accurate simulations, consistent with different experimental models. Implicit solvation is often overlooked for fear of compromising biological realism, however, the GB8 model demonstrably outperformed any other parameterisation solution. Furthermore, when compared to a well-sampled and unencumbered by temporal progression Markov-Chain Monte Carlo simulation, the GB8 simulations demonstrated great overlap in sampling range and a wide conformational space. This coupled with its computational accuracy, and unmatched efficiency, made the GB8 solvation method the prime parameterisation protocol used for c-MYC simulations. Certainly, an optimised protocol that can be useful to study other similar IDPs.

Having troubleshooted the simulation parameterisation, assessed it for accuracy and equilibrium, the path was prepared for the analysis of the simulation results. Of the many available methods commonly used to make sense of the noisy and complex datasets, tools such as K-means clustering and PCA did not perform adequately in terms of IDP trajectory analysis. These methods failed to characterise the important IDP features and the deployment

of alternative methods, including TICA and the Rg linear analysis over time, proved more helpful in uncovering the structural dynamics of MYC88. MYC88, which corresponds to the first 88 amino acids of the c-MYC protein, contains two highly conserved MYC boxes: MB0 and MBI. The TICA analysis identified three metastable MYC88 states: a compact state, a well-sampled intermediate state, and a more unfolded state created by the N-terminal and the MB0 intermittent outward extension. The linear analysis of the radius of gyration, which ascertained the maximum and minimum peaks in terms of conformation compactness, uncovers MYC88's highest amplitude motion clearly dominated by the extension of the N-terminal and MB0. This motion created by the MB0 full extension and subsequent reassociation is proposed to be involved in 'fly-casting' and optimising c-MYC's binding to key MB0 molecular partners. Such molecular partners, as in the case of PIN1, are crucial to c-MYC's life cycle – including modulating the protein's transcriptional activation and degradation.

With MYC88's metastable states defined, its slowest and most abundant state offered a window of opportunity for drug discovery, targeting a region never attempted before – its TAD domain. The search for a druggable pocket revealed the presence of a suitable cavity endowed with the correct geometry, electrostatics, residue composition and preservation over time. This region mainly comprised of MBI residues, included the phosphodegron residues, became the target site for drug screening. The drug screening process identified a series of suitable compounds, with six of them demonstrating excellent binding kinetics. These compounds selected for their pharmacodynamic characteristics, respect the Lipinski and lead-like compound rules and are exceptional candidates for further work in medicinal chemistry optimisation. The analysis of MYC88's interaction with the ligands uncovered an overwhelming consensus regarding the ligand influence on MYC88's conformational space: when bound to a ligand MYC88 loses its disorder, becomes stabilised, ordered, and more folded. The cause for this lies mainly with the loss of N-terminal extension, exposing the compound's two-fold action – on one hand, the ligands trap the N-terminal in a compacted

configuration, impeding its interaction with molecular partners; and, on the other, the ligands occupy the all-important phosphodegron region, obstructing by competition the access to the MBI phosphorylation sites. This loss of function is necessary and essential when battling c-MYC overexpression in c-MYC-driven cancers. These findings provide a proof-of-concept that c-MYC should not be regarded 'undruggable' when regions, other than its DNA-binding domain, can demonstrably be successfully tackled.

Finally, the c-MYC research extended to the entirety of its TAD domain including its first 3 MYC boxes: MB0, MBI and MBII. This spans c-MYC's first 150 amino acids (MYC150). The investigation revealed that MYC150 validated much of MYC88's structural activity, particularly with MB0 displaying the same pattern of periodic and frequent extensions. In tandem, MYC150 peak analysis discovered additional rare extensions which involved the other two MYC boxes. The contact and network analysis reiterated the centrality of the phosphodegron residues as main hubs modulating the protein's long-range intramolecular connectivity. For the unphosphorylated MYC150, the network orchestration relies on the contact patterns of two prolines – PRO59 and PRO60. When SER62 gets phosphorylated, SER62 assumes a position of high centrality, likewise when THR58 gets phosphorylated, THR58 itself becomes the most important functional residues in the network displaying the highest degree and betweenness centrality. The network of contacts regulation creates structural dynamics which modulate accessibility to the phosphodegron residues: when SER62 gets phosphorylated its own accessibility decreases, reducing the chances for a potential undue dephosphorylation; however, the accessibility of THR58 increases since it is appropriate and necessary to phosphorylate THR58 in order to start the process of degradation. Likewise, when THR58 becomes phosphorylated, its own accessibility is reduced to avoid undue dephosphorylation which would put the c-MYC back into active mode; while SER62's accessibility is increased since its dephosphorylation is now necessary to complete the flagging process crucial for its degradation. All these processes are highly dependent on the precise residue composition in the phosphodegron and mutations here are known to cause

extremely deleterious effects. Substituting either SER62 or THR58, with a phosphomimetic residue produces catastrophic results, including a completely altered intramolecular connectivity network which abrogates the phosphodegron control. Additionally, it was found that PRO59 possesses a peculiar pattern of *cis-trans* configuration, inhabiting the *cis* and *trans* states equally in the unphosphorylated protein. PRO59 then moves to a *cis* exclusive mode when MYC150 becomes phosphorylated. Mutations of either SER62 and THR58 abrogate this isomerisation switch and move PRO59 to a *trans* only configuration, making c-MYC more likely to interact with molecular partners, for example kinases, which tend to prefer a *trans* substate. This likely explains for why mutations involving, or nearby, any of the phosphoresidues lead to undue c-MYC activation.

Overall, the explorative research pursued in this work offers proof-of-concept which highlights the robustness of well-parameterised *in-silico* simulation methods, and machine learning analysis algorithms, to study previously deemed intractable proteins. The intention is not to replace the experimental work, as the experimental validation is, of course, crucial for the advancement of our knowledge of c-MYC. Rather, it aims to offer additional and/or alternative methods to delve deeper where experimental work cannot, due to its limitations; and provide extra guidance where the experimental work is insufficient to establish the full picture. Owing to our ever-expanding computational capabilities, the refinement of our algorithms and the advancements in data science, the use of *in-silico* simulation methods and bioinformatic resources to make sense of molecular biological systems is more than practical, it is absolutely necessary.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX.* 1-2 (C), 19-25. Available from: http://dx.doi.org/10.1016/j.softx.2015.06.001. Available from: doi: 10.1016/j.softx.2015.06.001.

Alder, B. J. & Wainwright, T. E. (1957) Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics.* 27 (5), 1208-1209. Available from: https://www.osti.gov/biblio/4322875. Available from: doi: 10.1063/1.1743957.

Ambrosini, G., Sawle, A. D., Musi, E. & Schwartz, G. K. (2015) BRD4-targeted therapy induces Myc-independent cytotoxicity in Gnaq/11-mutatant uveal melanoma cells. *Oncotarget.* 6 (32), 33397-33409. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26397223. Available from: doi: 10.18632/oncotarget.5179.

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. (2004) Network Analysis of Protein Structures Identifies Functional Residues. *Journal of Molecular Biology.* 344 (4), 1135-1146. Available from: http://dx.doi.org/10.1016/j.jmb.2004.10.055. Available from: doi: 10.1016/j.jmb.2004.10.055.

Amy S. Farrell, Carl Pelz, Xiaoyan Wang, Colin J. Daniel, Zhiping Wang, Yulong Su, Mahnaz Janghorban, Xiaoli Zhang, Charlie Morgan, Soren Impey & Rosalie C. Sears. (2013) Pin1 Regulates the Dynamics of c-Myc DNA Binding To Facilitate Target Gene Regulation and Oncogenesis. *Molecular and Cellular Biology.* 33 (15), 2930-2949. Available from: http://mcb.asm.org/content/33/15/2930.abstract. Available from: doi: 10.1128/MCB.01455-12.

Andresen, C., Helander, S., Lemak, A., Farès, C., Csizmok, V., Carlsson, J., Penn, L. Z., Forman-Kay, J. D., Arrowsmith, C. H., Lundström, P. & Sunnerhagen, M. (2012) Transient structure and dynamics in the disordered c-Myc transactivation domain affect Bin1 binding. *Nucleic Acids Research.* 40 (13), 6353-6366. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22457068. Available from: doi: 10.1093/nar/gks263.

Andrieu, G., Belkina, A. C. & Denis, G. V. (2016) Clinical trials for BET inhibitors run ahead of the science. *Drug Discovery Today: Technologies.* 19 45-50. Available from: http://dx.doi.org/10.1016/j.ddtec.2016.06.004. Available from: doi: 10.1016/j.ddtec.2016.06.004.

Antje Menssen & Heiko Hermeking. (2002) Characterization of the c-MYC-Regulated Transcriptome by SAGE: Identification and Analysis of c-MYC Target Genes. *Proceedings of the National Academy of Sciences of the United States of America.* 99 (9), 6274-6279. Available from: https://www.jstor.org/stable/3058662. Available from: doi: 10.1073/pnas.082005599.

Arnold, H. K. & Sears, R. C. (2006) Protein phosphatase 2A regulatory subunit B56alpha associates with c-myc and negatively regulates c-myc accumulation. *Molecular and Cellular Biology.* 26 (7), 2832. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16537924.

Askew, D. S., Ashmun, R. A., Simmons, B. C. & Cleveland, J. L. (1991) Constitutive c-myc expression in an IL-3-dependent myeloid cell line suppresses cell cycle arrest and accelerates apoptosis. *Oncogene.* 6 (10), 1915-1922.

Atilgan, A. R., Akan, P. & Baysal, C. (2004) Small-World Communication of Residues and Significance for Protein Dynamics. *Biophysical Journal.* 86 (1), 85-91. Available from: http://dx.doi.org/10.1016/S0006-3495(04)74086-2. Available from: doi: 10.1016/S0006-3495(04)74086-2.

Austen, M., Vervoorts, J., Lüscher-Firzlaff, J. M., Rottmann, S., Lilischkis, R., Lüscher, B., Dohmann, K. & Walsemann, G. (2003) Stimulation of c-MYC transcriptional activity and acetylation by recruitment of the cofactor CBP. *EMBO Reports.* 4 (5), 484-490. Available from: http://dx.doi.org/10.1038/sj.embor.embor821. Available from: doi: 10.1038/sj.embor.embor821.

B Vennstrom, D Sheiness, J Zabielski & J M Bishop. (1982) Isolation and characterization of c-myc, a cellular homolog of the oncogene (v-myc) of avian myelocytomatosis virus strain 29. *Journal of Virology.* 42 (3), 773-779. Available from: http://jvi.asm.org/content/42/3/773.abstract.

Babu, M. M. (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society Transactions.* 44 (5), 1185-1200. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27911701. Available from: doi: 10.1042/BST20160172.

Bahram, F., von der Lehr, N., Cetinkaya, C. & Larsson, L. G. (2000) c-Myc hot spot mutations in lymphomas result in inefficient ubiquitination and decreased proteasome-mediated turnover. *Blood.* 95 (6), 2104-2110. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10706881. Available from: doi: 10.1182/blood.V95.6.2104.

Baudino, T. A., McKay, C., Pendeville-Samain, H., Nilsson, J. A., Maclean, K. H., White, E. L., Davis, A. C., Ihle, J. N. & Cleveland, J. L. (2002) c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes & Development.* 16 (19), 2530-2543. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12368264. Available from: doi: 10.1101/gad.1024602.

Beauchamp, K. A., Lin, Y., Das, R. & Pande, V. S. (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation.* 8 (4), 1409-1414. Available from: http://dx.doi.org/10.1021/ct2007814. Available from: doi: 10.1021/ct2007814.

Beer, S., Zetterberg, A., Ihrie, R. A., McTaggart, R. A., Yang, Q., Bradon, N., Arvanitis, C., Attardi, L. D., Feng, S., Ruebner, B., Cardiff, R. D. & Felsher, D. W. (2004) Developmental Context Determines Latency of MYC-Induced Tumorigenesis. *PLoS Biology.* 2 (11), e332. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15455033. Available from: doi: 10.1371/journal.pbio.0020332.

Berg, T. (2008) Inhibition of transcription factors with small organic molecules. *Current Opinion in Chemical Biology.* 12 (4), 464-471. Available from: http://dx.doi.org/10.1016/j.cbpa.2008.07.023. Available from: doi: 10.1016/j.cbpa.2008.07.023.

Bertness, V., Minna, J. D., Brooks, B. J., Gazdar, A. F., Hollis, G. F., Sausville, E., Kirsch, I. R., McBride, O. W., Nau, M. M. & Battey, J. (1985) L- myc , a new myc -related gene amplified and expressed in human small cell lung cancer. *Nature.* 318 (6041), 69-73. Available from: http://dx.doi.org/10.1038/318069a0. Available from: doi: 10.1038/318069a0.

Best, R. B. (2017) Computational and theoretical advances in studies of intrinsically disordered proteins. *Current Opinion in Structural Biology.* 42 147-154. Available from: http://dx.doi.org/10.1016/j.sbi.2017.01.006. Available from: doi: 10.1016/j.sbi.2017.01.006.

Best, R. B., Zheng, W. & Mittal, J. (2014) Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation.* 10 (11), 5113-5124. Available from: http://dx.doi.org/10.1021/ct500569b. Available from: doi: 10.1021/ct500569b.

Bid, H. K., Phelps, D. A., Xaio, L., Guttridge, D. C., Lin, J., London, C., Baker, L. H., Mo, X. & Houghton, P. J. (2016) The Bromodomain BET Inhibitor JQ1 Suppresses Tumor Angiogenesis in Models of Childhood Sarcoma. *Molecular Cancer Therapeutics.* 15 (5), 1018-1028. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26908627. Available from: doi: 10.1158/1535-7163.MCT-15-0567.

Biljana Culjkovic, Ivan Topisirovic, Lucy Skrabanek, Melisa Ruiz-Gutierrez & Katherine L. B. Borden. (2006) eIF4E Is a Central Node of an RNA Regulon That Governs Cellular Proliferation. *The Journal of Cell Biology.* 175 (3), 415-426. Available from: https://www.jstor.org/stable/4152221. Available from: doi: 10.1083/jcb.200607020.

Boomsma, W., Frellsen, J., Harder, T., Bottaro, S., Johansson, K. E., Tian, P., Stovgaard, K., Andreetta, C., Olsson, S., Valentin, J. B., Antonov, L. D., Christensen, A. S., Borg, M., Jensen, J. H., Lindorff-Larsen, K., Ferkinghoff-Borg, J. & Hamelryck, T. (2013) PHAISTOS: A framework for Markov chain Monte Carlo simulation and inference of protein structure. *Journal of Computational Chemistry.* 34 (19), 1697-1705. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23292. Available from: doi: 10.1002/jcc.23292.

Bretones, G., Delgado, M. D. & León, J. (2015) Myc and cell cycle control. *BBA - Gene Regulatory Mechanisms.* 1849 (5), 506-516. Available from: https://www.sciencedirect.com/science/article/pii/S187493991400073X. Available from: doi: 10.1016/j.bbagrm.2014.03.013.

Brockmann, M., Poon, E., Berry, T., Carstensen, A., Deubzer, H. E., Rycak, L., Jamin, Y., Thway, K., Robinson, S. P., Roels, F., Witt, O., Fischer, M., Chesler, L. & Eilers, M. (2016) Small Molecule Inhibitors of Aurora-A Induce Proteasomal Degradation of N-Myc in Childhood Neuroblastoma. *Cancer Cell.* 30 (2), 357-358. Available from: http://dx.doi.org/10.1016/j.ccell.2016.07.002. Available from: doi: 10.1016/j.ccell.2016.07.002.

Carabet, L. A., Rennie, P. S. & Cherkasov, A. (2018) Therapeutic Inhibition of Myc in Cancer. Structural Bases and Computer-Aided Drug Discovery Approaches. *International Journal of Molecular Sciences.* 20 (1), 120. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30597997. Available from: doi: 10.3390/ijms20010120.

Case, D. A., Betz, R. M., Cerutti, D. S., Cheatham III, T. E., Darden, T. A., Duke, R. E., Giese, T. J., Gohlke, H., Goetz, A. W. & Homeyer, N. (2016) AMBER 2016 reference manual. *University of California: San Francisco, CA, USA.* 1-923.

Cermelli, S., Jang, I. S., Bernard, B. & Grandori, C. (2014) Synthetic Lethal Screens as a Means to Understand and Treat MYC-Driven Cancers. *Cold Spring Harbor Perspectives in Medicine.* 4 (3), a014209. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24591535. Available from: doi: 10.1101/cshperspect.a014209.

Chakraborty, A. A., Scuoppo, C., Dey, S., Thomas, L. R., Lorey, S. L., Lowe, S. W. & Tansey, W. P. (2015) A common functional consequence of tumor-derived mutations within c-MYC. *Oncogene.* 34 (18), 2406-2409. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24998853. Available from: doi: 10.1038/onc.2014.186.

Charron, J., Malynn, B. A., Fisher, P., Stewart, V., Jeannotte, L., Goff, S. P., Robertson, E. J. & Alt, F. W. (1992) Embryonic lethality in mice homozygous for a targeted disruption of the N-myc gene. *Genes & Development.* 6 (12A), 2248-2257. Available from: https://www.ncbi.nlm.nih.gov/pubmed/1459450. Available from: doi: 10.1101/gad.6.12a.2248.

Cheng, P. F., Eisenman, R. N., McMahon, S. B., Zhang, X., Knoepfler, P. S. & Gafken, P. R. (2006) Myc influences global chromatin structure. *The EMBO Journal.* 25 (12), 2723-2734. Available from: http://dx.doi.org/10.1038/sj.emboj.7601152. Available from: doi: 10.1038/sj.emboj.7601152.

Chi V. Dang, Anne Le & Ping Gao. (2009) MYC-Induced Cancer Cell Energy Metabolism and Therapeutic Opportunities. *Clinical Cancer Research.* 15 (21), 6479-6483. Available from: http://clincancerres.aacrjournals.org/content/15/21/6479.abstract. Available from: doi: 10.1158/1078-0432.CCR-09-0889.

Chipumuro, E., Marco, E., Christensen, C., Kwiatkowski, N., Zhang, T., Hatheway, C., Abraham, B., Sharma, B., Yeung, C., Altabef, A., Perez-Atayde, A., Wong, K., Yuan, G., Gray, N., Young, R. & George, R. (2014) CDK7 Inhibition Suppresses Super-Enhancer-Linked Oncogenic Transcription in MYCN-Driven Cancer. *Cell.* 159 (5), 1126-1139. Available from: http://dx.doi.org/10.1016/j.cell.2014.10.024. Available from: doi: 10.1016/j.cell.2014.10.024.

Chodera, J. D. & Noé, F. (2014) Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology.* 25 135-144. Available from: http://dx.doi.org/10.1016/j.sbi.2014.04.002. Available from: doi: 10.1016/j.sbi.2014.04.002.

Chong, S., Chatterjee, P. & Ham, S. (2017) Computer Simulations of Intrinsically Disordered Proteins. *Annual Review of Physical Chemistry.* 68 (1), 117-134. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28226222. Available from: doi: 10.1146/annurev-physchem-052516-050843.

Christensen, C., Kwiatkowski, N., Abraham, B., Carretero, J., Al-Shahrour, F., Zhang, T., Chipumuro, E., Herter-Sprie, G., Akbay, E., Altabef, A., Zhang, J., Shimamura, T., Capelletti, M., Reibel, J., Cavanaugh, J., Gao, P., Liu, Y., Michaelsen, S., Poulsen, H., Aref, A., Barbie, D., Bradner, J., George, R., Gray, N., Young, R. & Wong, K. (2015) Targeting Transcriptional Addictions in Small Cell Lung Cancer with a Covalent CDK7 Inhibitor. *Cancer Cell.* 27 (1), 149. Available from: http://dx.doi.org/10.1016/j.ccell.2014.12.007. Available from: doi: 10.1016/j.ccell.2014.12.007.

Conacci-Sorrell, M., Ngouenet, C. & Eisenman, R. N. (2010) Myc-Nick: A Cytoplasmic Cleavage Product of Myc that Promotes α-Tubulin Acetylation and Cell Differentiation. *Cell.* 142 (3), 480-493. Available from: http://dx.doi.org/10.1016/j.cell.2010.06.037. Available from: doi: 10.1016/j.cell.2010.06.037.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society.* 117 (19), 5179-5197. Available from: http://dx.doi.org/10.1021/ja00124a002. Available from: doi: 10.1021/ja00124a002.

Corvetta, D., Chayka, O., Gherardi, S., D'Acunto, C. W., Cantilena, S., Valli, E., Piotrowska, I., Perini, G. & Sala, A. (2013) Physical interaction between MYCN oncogene and polycomb repressive complex 2 (PRC2) in neuroblastoma: functional and therapeutic implications. *The Journal of Biological Chemistry.* 288 (12), 8332-8341. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23362253. Available from: doi: 10.1074/jbc.M113.454280.

Cragnell, C., Durand, D., Cabane, B. & Skepö, M. (2016) Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. *Proteins: Structure, Function, and Bioinformatics.* 84 (6), 777-791. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25025. Available from: doi: 10.1002/prot.25025.

D Sheiness, L Fanshier & J M Bishop. (1978) Identification of nucleotide sequences which may encode the oncogenic capacity of avian retrovirus MC29. *Journal of Virology.* 28 (2), 600-610. Available from: http://jvi.asm.org/content/28/2/600.abstract.

Dang, C. (2012) MYC on the Path to Cancer. *Cell.* 149 (1), 22-35. Available from: http://dx.doi.org/10.1016/j.cell.2012.03.003. Available from: doi: 10.1016/j.cell.2012.03.003.

Dang, C. V., Li, F. & Lee, L. A. (2005) Could MYC Induction of Mitochondrial Biogenesis be linked to ROS Production and Genomic Instability? *Cell Cycle.* 4 (11), 1465-1466. Available from: http://www.tandfonline.com/doi/abs/10.4161/cc.4.11.2121. Available from: doi: 10.4161/cc.4.11.2121.

Dang, C. V., Reddy, E. P., Shokat, K. M. & Soucek, L. (2017) Drugging the 'undruggable' cancer targets. *Nature Reviews. Cancer.* 17 (8), 502-508. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28643779. Available from: doi: 10.1038/nrc.2017.36.

David R. Wise, Ralph J. DeBerardinis, Anthony Mancuso, Nabil Sayed, Xiao-Yong Zhang, Harla K. Pfeiffer, Ilana Nissim, Evgueni Daikhin, Marc Yudkoff, Steven B. McMahon & Craig B. Thompson. (2008) Myc Regulates a Transcriptional Program That Stimulates Mitochondrial Glutaminolysis and Leads to Glutamine Addiction. *Proceedings of the National Academy of Sciences of the United States of America.* 105 (48), 18782-18787. Available from: https://www.jstor.org/stable/25465544. Available from: doi: 10.1073/pnas.0810199105.

David, C. C. & Jacobs, D. J. (2014) Principal component analysis: a method for determining the essential dynamics of proteins. *Methods in Molecular Biology (Clifton, N.J.).* 1084 193-226. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24061923. Available from: doi: 10.1007/978-1-62703-658-0_11.

Davis, A. C., Wims, M., Spotts, G. D., Hann, S. R. & Bradley, A. (1993) A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. *Genes & Development.* 7 (4), 671-682. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8458579. Available from: doi: 10.1101/gad.7.4.671.

De Paris, R., Quevedo, C. V., Ruiz, D. D., Norberto de Souza, O. & Barros, R. C. (2015) Clustering Molecular Dynamics Trajectories for Optimizing Docking Experiments.

Computational Intelligence and Neuroscience. 2015 916240-9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25873944. Available from: doi: 10.1155/2015/916240.

del Sol, A., Fujihashi, H. & O'Meara, P. (2005) Topology of small-world networks of protein-protein complex structures. *Bioinformatics (Oxford, England).* 21 (8), 1311-1315. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15659419. Available from: doi: 10.1093/bioinformatics/bti167.

del Sol, A. & O'Meara, P. (2005) Small-world network approach to identify key residues in protein–protein interaction. *Proteins: Structure, Function, and Bioinformatics.* 58 (3), 672-682. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20348. Available from: doi: 10.1002/prot.20348.

Delmore, J., Issa, G., Lemieux, M., Rahl, P., Shi, J., Jacobs, H., Kastritis, E., Gilpatrick, T., Paranal, R., Qi, J., Chesi, M., Schinzel, A., McKeown, M., Heffernan, T., Vakoc, C., Bergsagel, P. ., Ghobrial, I., Richardson, P., Young, R., Hahn, W., Anderson, K., Kung, A., Bradner, J. & Mitsiades, C. (2011) BET Bromodomain Inhibition as a Therapeutic Strategy to Target c-Myc. *Cell.* 146 (6), 904-917. Available from: http://dx.doi.org/10.1016/j.cell.2011.08.017. Available from: doi: 10.1016/j.cell.2011.08.017.

DePinho, R. A., Nisen, P. D., Kohl, N. E., Smith, R. K., Gee, C. E., Alt, F. W. & Legouy, E. (1986) Human N- myc is closely related in organization and nucleotide sequence to c- myc. *Nature.* 319 (6048), 73-77. Available from: http://dx.doi.org/10.1038/319073a0. Available from: doi: 10.1038/319073a0.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y. & Liang, J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research.* 34 (Web Server issue), W116-W118. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16844972. Available from: doi: 10.1093/nar/gkl282.

Dunn, S. & Cowling, V. H. (2015) Myc and mRNA capping. *BBA - Gene Regulatory Mechanisms.* 1849 (5), 501-505. Available from: https://www.sciencedirect.com/science/article/pii/S1874939914000674. Available from: doi: 10.1016/j.bbagrm.2014.03.007.

Eisenman, R. N., Gomez-Roman, N., White, R. J. & Grandori, C. (2003) Direct activation of RNA polymerase III transcription by c-Myc. *Nature.* 421 (6920), 290-294. Available from: http://dx.doi.org/10.1038/nature01327. Available from: doi: 10.1038/nature01327.

Evan, G. I., Wyllie, A. H., Gilbert, C. S., Littlewood, T. D., Land, H., Brooks, M., Waters, C. M., Penn, L. Z. & Hancock, D. C. (1992) Induction of apoptosis in fibroblasts by c-myc protein. *Cell.* 69 (1), 119-128. Available from: http://dx.doi.org/10.1016/0092-8674(92)90123-T. Available from: doi: 10.1016/0092-8674(92)90123-T.

Fanidi, A., Harrington, E. A. & Evan, G. I. (1992) Cooperative interaction between c-myc and bcl-2 proto-oncogenes. *Nature.* 359 (6395), 554-556.

Farrell, A. S. & Sears, R. C. (2014) MYC Degradation. *Cold Spring Harbor Perspectives in Medicine.* 4 (3), a014365. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24591536. Available from: doi: 10.1101/cshperspect.a014365.

Feig, M., Karanicolas, J. & Brooks, C. L. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling.* 22 (5), 377-395. Available from: http://dx.doi.org/10.1016/j.jmgm.2003.12.005. Available from: doi: 10.1016/j.jmgm.2003.12.005.

Fernandez, P. C., Frank, S. R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A. & Amati, B. (2003) Genomic targets of the human c-Myc protein. *Genes & Development.* 17 (9), 1115-1129. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12695333. Available from: doi: 10.1101/gad.1067003.

Fletcher, S. & Prochownik, E. V. (2015) Small-molecule inhibitors of the Myc oncoprotein. *BBA - Gene Regulatory Mechanisms.* 1849 (5), 525-543. Available from: http://dx.doi.org/10.1016/j.bbagrm.2014.03.005. Available from: doi: 10.1016/j.bbagrm.2014.03.005.

Fong, C. Y., Gilan, O., Lam, E. Y. N., Rubin, A. F., Ftouni, S., Tyler, D., Stanley, K., Sinha, D., Yeh, P., Morison, J., Giotopoulos, G., Lugo, D., Jeffrey, P., Lee, S. C., Carpenter, C., Gregory, R., Ramsay, R. G., Lane, S. W., Abdel-Wahab, O., Kouzarides, T., Johnstone, R. W., Dawson, S., Huntly, B. J. P., Prinjha, R. K., Papenfuss, A. T. & Dawson, M. A. (2015) BET inhibitor resistance emerges from leukaemia stem cells. *Nature.* 525 (7570), 538-542. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26367796. Available from: doi: 10.1038/nature14888.

Francesco Faiola, Xiaohui Liu, Szuying Lo, Songqin Pan, Kangling Zhang, Elena Lymar, Anthony Farina & Ernest Martinez. (2005) Dual Regulation of c-Myc by p300 via Acetylation-Dependent Control of Myc Protein Turnover and Coactivation of Myc-Induced Transcription. *Molecular and Cellular Biology.* 25 (23), 10220-10234. Available from: http://mcb.asm.org/content/25/23/10220.abstract. Available from: doi: 10.1128/MCB.25.23.10220-10234.2005.

Gabay, M., Li, Y. & Felsher, D. W. (2014) MYC Activation Is a Hallmark of Cancer Initiation and Maintenance. *Cold Spring Harbor Perspectives in Medicine.* 4 (6), a014241. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24890832. Available from: doi: 10.1101/cshperspect.a014241.

Galloway, D. A., Felton-Edkins, Z. A., Eisenman, R. N., White, R. J., Grandori, C., Ngouenet, C. & Gomez-Roman, N. (2005) c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature Cell Biology.* 7 (3), 311-318. Available from: http://dx.doi.org/10.1038/ncb1224. Available from: doi: 10.1038/ncb1224.

Gao, P., Tchernyshyov, I., Chang, T., Lee, Y., Kita, K., Ochi, T., Zeller, K. I., De Marzo, A. M., Van Eyk, J. E., Mendell, J. T. & Dang, C. V. (2009) c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature.* 458 (7239), 762-765. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19219026. Available from: doi: 10.1038/nature07823.

Garcia, P. L., Miller, A. L., Kreitzburg, K. M., Council, L. N., Gamblin, T. L., Christein, J. D., Heslin, M. J., Arnoletti, J. P., Richardson, J. H., Chen, D., Hanna, C. A., Cramer, S. L., Yang, E. S., Qi, J., Bradner, J. E. & Yoon, K. J. (2016) The BET bromodomain inhibitor JQ1 suppresses growth of pancreatic ductal adenocarcinoma in patient-derived xenograft models. *Oncogene.* 35 (7), 833-845. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25961927. Available from: doi: 10.1038/onc.2015.126.

García-Gutiérrez, L., Delgado, M. D. & León, J. (2019) MYC Oncogene Contributions to Release of Cell Cycle Brakes. *Genes.* 10 (3), 244. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30909496. Available from: doi: 10.3390/genes10030244.

Gelin, B. R., Karplus, M. & McCammon, J. A. (1977) Dynamics of folded proteins. *Nature.* 267 (5612), 585-590. Available from: http://dx.doi.org/10.1038/267585a0. Available from: doi: 10.1038/267585a0.

Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., Domanski, J., Dotson, D. L., Buchoux, S., Kenney, I. M. & Beckstein, O. (Sep 11, 2016) MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. United StatesAvailable from: https://www.osti.gov/servlets/purl/1565806 .

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England).* 22 (21), 2695-2696. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16940322. Available from: doi: 10.1093/bioinformatics/btl461.

GraphPad Software. (2020) *GraphPad Prism* (8.3.1) San Diego, California USA, .

Guo, J., Guo, J., Parise, R., Parise, R., Joseph, E., Joseph, E., Egorin, M., Egorin, M., Lazo, J., Lazo, J., Prochownik, E., Prochownik, E., Eiseman, J. & Eiseman, J. (2009) Efficacy, pharmacokinetics, tisssue distribution, and metabolism of the Myc–Max disruptor, 10058-F4 [Z,E]-5-[4-ethylbenzylidine]-2-thioxothiazolidin-4-one, in mice. *Cancer Chemotherapy and Pharmacology.* 63 (4), 615-625. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18509642. Available from: doi: 10.1007/s00280-008-0774-y.

Guo, Q. M., Malek, R. L., Kim, S., Chiao, C., He, M., Ruffy, M., Sanka, K., Lee, N. H., Dang, C. V. & Liu, E. T. (2000) Identification of c-Myc Responsive Genes Using Rat cDNA Microarray. *Cancer Research.* 60 (21), 5922. Available from: http://cancerres.aacrjournals.org/cgi/content/abstract/60/21/5922.

Hamelberg, D. & McCammon, J. A. (2009) Mechanistic Insight into the Role of Transition-State Stabilization in Cyclophilin A. *Journal of the American Chemical Society.* 131 (1), 147-152. Available from: http://dx.doi.org/10.1021/ja806146g. Available from: doi: 10.1021/ja806146g.

Hann, S. R. (2006) Role of post-translational modifications in regulating c-Myc proteolysis, transcriptional activity and biological function. *Seminars in Cancer Biology.* 16 (4), 288-302. Available from: http://dx.doi.org/10.1016/j.semcancer.2006.08.004. Available from: doi: 10.1016/j.semcancer.2006.08.004.

Helander, S., Montecchio, M., Pilstål, R., Su, Y., Kuruvilla, J., Elvén, M., Ziauddin, J. M. E., Anandapadamanaban, M., Cristobal, S., Lundström, P., Sears, R. C., Wallner, B. & Sunnerhagen, M. (2015) Pre-Anchoring of Pin1 to Unphosphorylated c-Myc in a Fuzzy Complex Regulates c-Myc Activity. *Structure.* 23 (12), 2267-2279. Available from: http://dx.doi.org/10.1016/j.str.2015.10.010. Available from: doi: 10.1016/j.str.2015.10.010.

Henriques, J., Cragnell, C. & Skepö, M. (2015) Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation.* 11 (7), 3420-3431.

Henriques, J. & Skepö, M. (2016) Molecular Dynamics Simulations of Intrinsically Disordered Proteins: On the Accuracy of the TIP4P-D Water Model and the Representativeness of Protein Disorder Models. *Journal of Chemical Theory and Computation.* 12 (7), 3407-3415. Available from: http://dx.doi.org/10.1021/acs.jctc.6b00429. Available from: doi: 10.1021/acs.jctc.6b00429.

Hilary A. Coller, Carla Grandori, Pablo Tamayo, Trent Colbert, Eric S. Lander, Robert N. Eisenman & Todd R. Golub. (2000) Expression Analysis with Oligonucleotide Microarrays Reveals That MYC Regulates Genes Involved in Growth, Cell Cycle, Signaling, and Adhesion. *Proceedings of the National Academy of Sciences of the United States of America.* 97 (7), 3260-3265. Available from: https://www.jstor.org/stable/121879. Available from: doi: 10.1073/pnas.97.7.3260.

Hogg, S. J., Newbold, A., Vervoort, S. J., Cluse, L. A., Martin, B. P., Gregory, G. P., Lefebure, M., Vidacs, E., Tothill, R. W., Bradner, J. E., Shortt, J. & Johnstone, R. W. (2016) BET Inhibition Induces Apoptosis in Aggressive B-Cell Lymphoma via Epigenetic Regulation of BCL-2 Family Members. *Molecular Cancer Therapeutics.* 15 (9), 2030-2041. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27406984. Available from: doi: 10.1158/1535-7163.MCT-15-0924.

Hollingsworth, S. A. & Dror, R. O. (2018) Molecular dynamics simulation for all. *Neuron.* 99 (6), 1129-1143.

Hongjian Li, Kwong-Sak Leung & Man-Hon Wong. (May 2012) idock: A multithreaded virtual screening tool for flexible ligand docking. *CIBCB.* , IEEE. pp.77-84Available from: https://ieeexplore.ieee.org/document/6217214 .

Horiuchi, D., Anderton, B. & Goga, A. (2014) Taking on Challenging Targets: Making MYC Druggable. *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting.* (34), e497-e502. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24857145. Available from: doi: 10.14694/EdBook_AM.2014.34.e497.

Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. (2015) Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry : AABC.* 8 37-47. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26604800. Available from: doi: 10.2147/AABC.S70333.

Hsieh, A. L., Walton, Z. E., Altman, B. J., Stine, Z. E. & Dang, C. V. (2015) MYC and metabolism on the path to cancer. *Seminars in Cell and Developmental Biology.* 43 11-21. Available from: https://www.sciencedirect.com/science/article/pii/S1084952115001470. Available from: doi: 10.1016/j.semcdb.2015.08.003.

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B. S., Mesquita, A., Liu, C., Yuan, C. L., Hu, Y., Hüttelmaier, S., Skibbe, J. R., Su, R., Deng, X., Dong, L., Sun, M., Li, C., Nachtergaele, S., Wang, Y., Hu, C., Ferchen, K., Greis, K. D., Jiang, X., Wei, M., Qu, L., Guan, J., He, C., Yang, J. & Chen, J. (2018) Recognition of RNA N 6 -methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature Cell Biology.* 20 (3), 285-295. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29476152. Available from: doi: 10.1038/s41556-018-0045-z.

Huang, H., Weng, H., Wang, L., Yu, C., Huang, Q., Zhao, P., Wen, J., Zhou, H. & Qu, L. (2012) Triggering Fbw7-Mediated Proteasomal Degradation of c-Myc by Oridonin Induces Cell Growth Inhibition and Apoptosis. *Molecular Cancer Therapeutics.* 11 (5), 1155-1165. Available

from: https://www.ncbi.nlm.nih.gov/pubmed/22389469. Available from: doi: 10.1158/1535-7163.MCT-12-0066.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H. & MacKerell, J., Alexander D. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods.* 14 (1), 71-73. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27819658. Available from: doi: 10.1038/nmeth.4067.

Humphrey, W., Dalke, A. & Schulten, K. (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics.* 14 (1), 33-38. Available from: http://dx.doi.org/10.1016/0263-7855(96)00018-5. Available from: doi: 10.1016/0263-7855(96)00018-5.

Hussein, H. A., Borrel, A., Geneix, C., Petitjean, M., Regad, L. & Camproux, A. (2015) PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Research.* 43 (W1), W436-W442. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25956651. Available from: doi: 10.1093/nar/gkv462.

Hyunsuk Shim, Christine Dolde, Brian C. Lewis, Chyi-Sun Wu, Gerard Dang, Richard A. Jungmann, Riccardo Dalla-Favera & Chi V. Dang. (1997) c-Myc Transactivation of LDH-A: Implications for Tumor Metabolism and Growth. *Proceedings of the National Academy of Sciences of the United States of America.* 94 (13), 6658-6663. Available from: https://www.jstor.org/stable/42208. Available from: doi: 10.1073/pnas.94.13.6658.

Hyunsuk Shim, Yoon S. Chun, Brian C. Lewis & Chi V. Dang. (1998) A Unique Glucose-Dependent Apoptotic Pathway Induced by c-Myc. *Proceedings of the National Academy of Sciences of the United States of America.* 95 (4), 1511-1516. Available from: https://www.jstor.org/stable/44314. Available from: doi: 10.1073/pnas.95.4.1511.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. (2012) ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling.* 52 (7), 1757-1768. Available from: http://dx.doi.org/10.1021/ci3001277. Available from: doi: 10.1021/ci3001277.

Ishida, N., Hatakeyama, S., Nakayama, K. I., Tsunematsu, R., Okumura, F., Kamura, T., Nakayama, K., Imaki, H., Yada, M. & Nishiyama, M. (2004) Phosphorylation-dependent degradation of c-Myc is mediated by the F-box protein Fbw7. *The EMBO Journal.* 23 (10), 2116-2125. Available from: http://dx.doi.org/10.1038/sj.emboj.7600217. Available from: doi: 10.1038/sj.emboj.7600217.

J. A. Hartigan & M. A. Wong. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics).* 28 (1), 100-108. Available from: https://www.jstor.org/stable/2346830. Available from: doi: 10.2307/2346830.

Jacob Sarid, Thanos D. Halazonetis, William Murphy & Philip Leder. (1987) Evolutionarily Conserved Regions of the Human c-myc Protein Can Be Uncoupled from Transforming Activity. *Proceedings of the National Academy of Sciences of the United States of America.* 84 (1), 170-173. Available from: https://www.jstor.org/stable/28970. Available from: doi: 10.1073/pnas.84.1.170.

Jagruti H. Patel, Yanping Du, Penny G. Ard, Charles Phillips, Beth Carella, Chi-Ju Chen, Carrie Rakowski, Chandrima Chatterjee, Paul M. Lieberman, William S. Lane, Gerd A. Blobel & Steven B. McMahon. (2004) The c-MYC Oncoprotein Is a Substrate of the Acetyltransferases hGCN5/PCAF and TIP60. *Molecular and Cellular Biology.* 24 (24), 10826-

10834. Available from: http://mcb.asm.org/content/24/24/10826.abstract. Available from: doi: 10.1128/MCB.24.24.10826-10834.2004.

Jain, S., Wang, X., Chang, C., Ibarra-Drendall, C., Wang, H., Zhang, Q., Brady, S. W., Li, P., Zhao, H., Dobbs, J., Kyrish, M., Tkaczyk, T. S., Ambrose, A., Sistrunk, C., Arun, B. K., Richards-Kortum, R., Jia, W., Seewaldt, V. L. & Yu, D. (2015) Src Inhibition Blocks c-Myc Translation and Glucose Metabolism to Prevent the Development of Breast Cancer. *Cancer Research.* 75 (22), 4863-4875. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26383165. Available from: doi: 10.1158/0008-5472.CAN-14-2345.

Jennifer A. Mertz, Andrew R. Conery, Barbara M. Bryant, Peter Sandy, Srividya Balasubramanian, Deanna A. Mele, Louise Bergeron & Robert J. Sims. (2011) Targeting MYC dependence in cancer by inhibiting BET bromodomains. *Proceedings of the National Academy of Sciences of the United States of America.* 108 (40), 16669-16674. Available from: https://www.jstor.org/stable/41321759. Available from: doi: 10.1073/pnas.1108190108.

John E Stone, Michael J Hallock, James C Phillips, Joseph R Peterson, Zaida Luthey-Schulten & Klaus Schulten. (Jan 1, 2016) Evaluation of Emerging Energy-Efficient Heterogeneous Computing Platforms for Biomolecular and Cellular Simulation Workloads. *The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings.* Piscataway, The Institute of Electrical and Electronics Engineers, Inc. (IEEE). pp.89Available from: https://search.proquest.com/docview/1808942296 .

K S Hatton, K Mahon, L Chin, F C Chiu, H W Lee, D Peng, S D Morgenbesser, J Horner & R A DePinho. (1996) Expression and activity of L-Myc in normal mouse development. *Molecular and Cellular Biology.* 16 (4), 1794-1804. Available from: http://mcb.asm.org/content/16/4/1794.abstract. Available from: doi: 10.1128/MCB.16.4.1794.

Karen I. Zeller, XiaoDong Zhao, Charlie W. H. Lee, Kuo Ping Chiu, Fei Yao, Jason T. Yustein, Hong Sain Ooi, Yuriy L. Orlov, Atif Shahab, How Choong Yong, YuTao Fu, Zhiping Weng, Vladimir A. Kuznetsov, Wing-Kin Sung, Yijun Ruan, Chi V. Dang & Chia-Lin Wei. (2006) Global Mapping of c-Myc Binding Sites and Target Gene Networks in Human B Cells. *Proceedings of the National Academy of Sciences of the United States of America.* 103 (47), 17834-17839. Available from: https://www.jstor.org/stable/30052549. Available from: doi: 10.1073/pnas.0604129103.

Kelly, K., Cochran, B. H., Stiles, C. D. & Leder, P. (1983) Cell-specific regulation of the c-myc gene by lymphocyte mitogens and platelet-derived growth factor. *Cell.* 35 (3), 603-610. Available from: https://www.sciencedirect.com/science/article/pii/0092867483900922. Available from: doi: 10.1016/0092-8674(83)90092-2.

Kim, S. Y., Herbst, A., Tworkowski, K. A., Salghetti, S. E. & Tansey, W. P. (2003) Skp2 Regulates Myc Protein Stability and Activity. *Molecular Cell.* 11 (5), 1177-1188. Available from: http://dx.doi.org/10.1016/S1097-2765(03)00173-4. Available from: doi: 10.1016/S1097-2765(03)00173-4.

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. & Svergun, D. I. (2003) PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography.* 36 (5), 1277-1282.

Kouzarides, T., Didelot, C., de Launoit, Y., Bernard, D., Deplus, R., Loriot, A., Brenner, C., De Smet, C., Viré, E., Giuseppe Pelicci, P., Di Croce, L., Amati, B., Boon, T., Danovi, D., Gutierrez, A. & Fuks, F. (2005) Myc represses transcription through recruitment of DNA

methyltransferase corepressor. *The EMBO Journal.* 24 (2), 336-346. Available from: http://dx.doi.org/10.1038/sj.emboj.7600509. Available from: doi: 10.1038/sj.emboj.7600509.

Kovacs, J. A. & Wriggers, W. (2016) Spatial Heat Maps from Fast Information Matching of Fast and Slow Degrees of Freedom: Application to Molecular Dynamics Simulations. *The Journal of Physical Chemistry B.* 120 (33), 8473-8484. Available from: http://dx.doi.org/10.1021/acs.jpcb.6b02136. Available from: doi: 10.1021/acs.jpcb.6b02136.

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., Xia, B., Beglov, D. & Vajda, S. (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols.* 10 (5), 733-755. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25855957. Available from: doi: 10.1038/nprot.2015.043.

Kwiatkowski, N., Zhang, T., Rahl, P. B., Abraham, B. J., Reddy, J., Ficarro, S. B., Dastur, A., Amzallag, A., Ramaswamy, S., Tesar, B., Jenkins, C. E., Hannett, N. M., McMillin, D., Sanda, T., Sim, T., Kim, N. D., Look, T., Mitsiades, C. S., Weng, A. P., Brown, J. R., Benes, C. H., Marto, J. A., Young, R. A. & Gray, N. S. (2014) Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature.* 511 (7511), 616-620. Available from: https://search.proquest.com/docview/1551986268. Available from: doi: 10.1038/nature13393.

Lawrence W. Stanton, Manfred Schwab & J. Michael Bishop. (1986) Nucleotide Sequence of the Human N-myc Gene. *Proceedings of the National Academy of Sciences of the United States of America.* 83 (6), 1772-1776. Available from: https://www.jstor.org/stable/26875. Available from: doi: 10.1073/pnas.83.6.1772.

Le, A., Lane, A., Hamaker, M., Bose, S., Gouw, A., Barbi, J., Tsukamoto, T., Rojas, C., Slusher, B., Zhang, H., Zimmerman, L., Liebler, D., Slebos, R. C., Lorkiewicz, P., Higashi, R., Fan, T. M. & Dang, C. (2012) Glucose-Independent Glutamine Metabolism via TCA Cycling for Proliferation and Survival in B Cells. *Cell Metabolism.* 15 (1), 110-121. Available from: http://dx.doi.org/10.1016/j.cmet.2011.12.009. Available from: doi: 10.1016/j.cmet.2011.12.009.

Legouy, E., DePinho, R., Zimmerman, K., Collum, R., Yancopoulos, G., Mitsock, L., Kriz, R. & Alt, F. W. (1987) Structure and expression of the murine L-myc gene. *The EMBO Journal.* 6 (11), 3359-3366. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1460-2075.1987.tb02657.x. Available from: doi: 10.1002/j.1460-2075.1987.tb02657.x.

Levine, Z. A. & Shea, J. (2017) Simulations of disordered proteins and systems with conformational heterogeneity. *Current Opinion in Structural Biology.* 43 95-103. Available from: http://dx.doi.org/10.1016/j.sbi.2016.11.006. Available from: doi: 10.1016/j.sbi.2016.11.006.

Licchesi, J. D. F., Van Neste, L., Tiwari, V. K., Cope, L., Lin, X., Baylin, S. B. & Herman, J. G. (2010) Transcriptional regulation of Wnt inhibitory factor-1 by Miz-1 c-Myc. *Oncogene.* 29 (44), 5923-5934. Available from: http://dx.doi.org/10.1038/onc.2010.322. Available from: doi: 10.1038/onc.2010.322.

Lin, C., Lovén, J., Rahl, P., Paranal, R., Burge, C., Bradner, J., Lee, T. & Young, R. (2012) Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell.* 151 (1), 56-67. Available from: https://www.sciencedirect.com/science/article/pii/S0092867412010574. Available from: doi: 10.1016/j.cell.2012.08.026.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews.* 23 (1-3), 3-25.

Liu, H., Guo, X., Han, J., Luo, R. & Chen, H. (2018) Order-disorder transition of intrinsically disordered kinase inducible transactivation domain of CREB. *The Journal of Chemical Physics.* 148 (22), 225101.

Liu, Y., Li, F., Handler, J., Huang, C. R. L., Xiang, Y., Neretti, N., Sedivy, J. M., Zeller, K. I. & Dang, C. V. (2008) Global Regulation of Nucleotide Biosynthetic Genes by c-Myc. *PloS One.* 3 (7), e2722. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18628958. Available from: doi: 10.1371/journal.pone.0002722.

Liu, Y., Huang, H., Liu, M., Wu, Q., Li, W. & Zhang, J. (2017) MicroRNA-24-1 suppresses mouse hepatoma cell invasion and metastasis via directly targeting O -GlcNAc transferase. *Biomedicine & Pharmacotherapy.* 91 731-738. Available from: https://www.clinicalkey.es/playcontent/1-s2.0-S0753332217309332. Available from: doi: 10.1016/j.biopha.2017.05.007.

Luo, L., Tang, H., Ling, L., Li, N., Jia, X., Zhang, Z., Wang, X., Shi, L., Yin, J., Qiu, N., Liu, H., Song, Y., Luo, K., Li, H., He, Z., Zheng, G. & Xie, X. (2018) LINC01638 lncRNA activates MTDH-Twist1 signaling by preventing SPOP-mediated c-Myc degradation in triple-negative breast cancer. *Oncogene.* 37 (47), 6166-6179. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30002443. Available from: doi: 10.1038/s41388-018-0396-8.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation.* 11 (8), 3696-3713.

Mark A. Gregory, Ying Qi & Stephen R. Hann. (2003) Phosphorylation by Glycogen Synthase Kinase-3 Controls c-Myc Proteolysis and Subnuclear Localization. *Journal of Biological Chemistry.* 278 (51), 51606-51612. Available from: http://www.jbc.org/content/278/51/51606.abstract. Available from: doi: 10.1074/jbc.M310722200.

Mark, P. & Nilsson, L. (2001) Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *The Journal of Physical Chemistry A.* 105 (43), 9954-9960. Available from: http://dx.doi.org/10.1021/jp003020w. Available from: doi: 10.1021/jp003020w.

Markus Welcker, Amir Orian, Jianping Jin, Jonathan A. Grim, J. Wade Harper, Robert N. Eisenman & Bruce E. Clurman. (2004) The Fbw7 Tumor Suppressor Regulates Glycogen Synthase Kinase 3 Phosphorylation-Dependent c-Myc Protein Degradation. *Proceedings of the National Academy of Sciences of the United States of America.* 101 (24), 9085-9090. Available from: https://www.jstor.org/stable/3372394. Available from: doi: 10.1073/pnas.0402770101.

McCammon, J. A. & Karplus, M. (2002) Molecular dynamics simulations of biomolecules. *Nature Structural Biology.* 9 (9), 646-652. Available from: http://dx.doi.org/10.1038/nsb0902-646. Available from: doi: 10.1038/nsb0902-646.

McGibbon, R., Beauchamp, K., Harrigan, M., Klein, C., Swails, J., Hernández, C., Schwantes, C., Wang, L., Lane, T. & Pande, V. (2015) MDTraj: A Modern Open Library for the Analysis of

Molecular Dynamics Trajectories. *Biophysical Journal.* 109 (8), 1528-1532. Available from: http://dx.doi.org/10.1016/j.bpj.2015.08.015. Available from: doi: 10.1016/j.bpj.2015.08.015.

Michael Pourdehnad, Morgan L. Truitt, Imran N. Siddiqi, Gregory S. Ducker, Kevan M. Shokat & Davide Ruggero. (2013) Myc and mTOR converge on a common node in protein synthesis control that confers synthetic lethality in Myc-driven cancers. *Proceedings of the National Academy of Sciences of the United States of America.* 110 (29), 11988-11993. Available from: https://www.jstor.org/stable/42712514. Available from: doi: 10.1073/pnas.1310230110.

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. (2011) MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry.* 32 (10), 2319-2327. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21787. Available from: doi: 10.1002/jcc.21787.

Miller, D. M., Thomas, S. D., Islam, A., Muench, D. & Sedoris, K. (2012) c-Myc and Cancer Metabolism. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research.* 18 (20), 5546-5553. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23071356. Available from: doi: 10.1158/1078-0432.CCR-12-0977.

Mladenov, Z., Heine, U., Beard, D. & Beard, J. W. (1967) Strain MC29 avian leukosis virus. Myelocytoma, endothelioma, and renal growths: pathomorphological and ultrastructural aspects. *Journal of the National Cancer Institute.* 38 (3), 251. Available from: https://www.ncbi.nlm.nih.gov/pubmed/4290185.

Morrish, F., Noonan, J., Perez-Olsen, C., Gafken, P. R., Fitzgibbon, M., Kelleher, J., VanGilst, M. & Hockenbery, D. (2010) Myc-dependent Mitochondrial Generation of Acetyl-CoA Contributes to Fatty Acid Biosynthesis and Histone Acetylation during Cell Cycle Entry. *The Journal of Biological Chemistry.* 285 (47), 36267-36274. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20813845. Available from: doi: 10.1074/jbc.M110.141606.

Mousavi, K. & Sartorelli, V. (2010) Myc-Nick: The Force Behind c-Myc. *Science Signaling.* 3 (152), pe49. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21156935. Available from: doi: 10.1126/scisignal.3152pe49.

Mu, Y., Nguyen, P. H. & Stock, G. (2005) Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins, Structure, Function, and Bioinformatics.* 58 (1), 45-52. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20310. Available from: doi: 10.1002/prot.20310.

Nicholson, L. K., Lu, K. P., Finn, G. & Lee, T. H. (2007) Prolyl cis-trans isomerization as a molecular timer. *Nature Chemical Biology.* 3 (10), 619-629. Available from: http://dx.doi.org/10.1038/nchembio.2007.35. Available from: doi: 10.1038/nchembio.2007.35.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D., Tessarollo, L., Casellas, R., Zhao, K. & Levens, D. (2012) c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. *Cell.* 151 (1), 68-79. Available from: https://www.sciencedirect.com/science/article/pii/S0092867412011014. Available from: doi: 10.1016/j.cell.2012.08.033.

Nikolay V. Dokholyan, Lewyn Li, Feng Ding & Eugene I. Shakhnovich. (2002) Topological Determinants of Protein Folding. *Proceedings of the National Academy of Sciences of the*

United States of America. 99 (13), 8637-8641. Available from: https://www.jstor.org/stable/3059068. Available from: doi: 10.1073/pnas.122076099.

Nilsson, J. A. & Cleveland, J. L. (2003) Myc pathways provoking cell suicide and cancer. *Oncogene.* 22 (56), 9007-9021. Available from: https://www.ncbi.nlm.nih.gov/pubmed/14663479. Available from: doi: 10.1038/sj.onc.1207261.

Onufriev, A. V. & Case, D. A. (2019) Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics.* 48 (1), 275-296. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30857399. Available from: doi: 10.1146/annurev-biophys-052118-115325.

Orian, A., van Steensel, B., Delrow, J., Bussemaker, H. J., Li, L., Sawado, T., Williams, E., Loo, L. W. M., Cowley, S. M., Yost, C., Pierce, S., Edgar, B. A., Parkhurst, S. M. & Eisenman, R. N. (2003) Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. *Genes & Development.* 17 (9), 1101-1114. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12695332. Available from: doi: 10.1101/gad.1066903.

Oster, S. K., Ho, C. S. W., Soucie, E. L. & Penn, L. Z. (2002) The myc oncogene: MarvelouslY Complex. *Advances in Cancer Research.* 84 81. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11885563. Available from: doi: 10.1016/S0065-230X(02)84004-0.

Pamela Mellon, Anthony Pawson, Klaus Bister, G. Steven Martin & Peter H. Duesberg. (1978) Specific RNA Sequences and Gene Products of MC29 Avian Acute Leukemia Virus. *Proceedings of the National Academy of Sciences of the United States of America.* 75 (12), 5874-5878. Available from: https://www.jstor.org/stable/68860. Available from: doi: 10.1073/pnas.75.12.5874.

Petit, J., Meurice, N., Kaiser, C. & Maggiora, G. (2012) Softening the Rule of Five—where to draw the line? *Bioorganic & Medicinal Chemistry.* 20 (18), 5343-5351. Available from: http://dx.doi.org/10.1016/j.bmc.2011.11.064. Available from: doi: 10.1016/j.bmc.2011.11.064.

Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *Journal of Applied Crystallography.* 45 (2), 342-350. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1107/S0021889812007662. Available from: doi: 10.1107/S0021889812007662.

Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. (2015) Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B.* 119 (16), 5113-5123. Available from: http://dx.doi.org/10.1021/jp508971m. Available from: doi: 10.1021/jp508971m.

Piana, S., Klepeis, J. L. & Shaw, D. E. (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology.* 24 98-105. Available from: http://dx.doi.org/10.1016/j.sbi.2013.12.006. Available from: doi: 10.1016/j.sbi.2013.12.006.

Pietrucci, F. & Laio, A. (2009) A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *Journal of Chemical Theory and*

Computation. 5 (9), 2197-2201. Available from: http://dx.doi.org/10.1021/ct900202f. Available from: doi: 10.1021/ct900202f.

Polivka, J., Jiri & Janku, F. (2014) Molecular targets for cancer therapy in the PI3K/AKT/mTOR pathway. *Pharmacology & Therapeutics.* 142 (2), 164-175. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24333502. Available from: doi: 10.1016/j.pharmthera.2013.12.004.

Posternak, V. & Cole, M. D. (2016) Strategically targeting MYC in cancer. *F1000Research.* 5 408. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27081479. Available from: doi: 10.12688/f1000research.7879.1.

Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A. & Young, R. A. (2010) c-Myc Regulates Transcriptional Pause Release. *Cell.* 141 (3), 432-445. Available from: https://www.sciencedirect.com/science/article/pii/S0092867410003181. Available from: doi: 10.1016/j.cell.2010.03.030.

Rahl, P. B. & Young, R. A. (2014) MYC and Transcription Elongation. *Cold Spring Harbor Perspectives in Medicine.* 4 (1), a020990. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24384817. Available from: doi: 10.1101/cshperspect.a020990.

Raj, P. A., Marcus, E. & Sukumaran, D. K. (1998) Structure of human salivary histatin 5 in aqueous and nonaqueous solutions. *Biopolymers.* 45 (1), 51-67. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-0282(199801)45:13.0.CO;2-Y. Available from: doi: 10.1002/(SICI)1097-0282(199801)45:13.0.CO;2-Y.

Rajan, A., Freddolino, P. L. & Schulten, K. (2010) Going beyond Clustering in MD Trajectory Analysis: An Application to Villin Headpiece Folding. *PloS One.* 5 (4), e9890. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20419160. Available from: doi: 10.1371/journal.pone.0009890.

Rambo, R. (2017) *Scatter* (3.0J) U.K., Diamond Light Source, .

Rathert, P., Roth, M., Neumann, T., Muerdter, F., Roe, J., Muhar, M., Deswal, S., Cerny-Reiterer, S., Peter, B., Jude, J., Hoffmann, T., Boryń, Ł M., Axelsson, E., Schweifer, N., Tontsch-Grunt, U., Dow, L. E., Gianni, D., Pearson, M., Valent, P., Stark, A., Kraut, N., Vakoc, C. R. & Zuber, J. (2015) Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. *Nature.* 525 (7570), 543-547. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26367798. Available from: doi: 10.1038/nature14898.

Rauscher, S., Gapsys, V., Gajda, M. J., Zweckstetter, M., de Groot, B. L. & Grubmüller, H. (2015) Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of Chemical Theory and Computation.* 11 (11), 5513-5524. Available from: http://dx.doi.org/10.1021/acs.jctc.5b00736. Available from: doi: 10.1021/acs.jctc.5b00736.

Rechsteiner, M. & Rogers, S. W. (1996) PEST sequences and regulation by proteolysis. *Trends in Biochemical Sciences.* 21 (7), 267-271. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8755249. Available from: doi: 10.1016/0968-0004(96)10031-1.

Ridderstråle, K., Wright, A. P. H., Shiue, C., Grummt, I., Wu, S., Larsson, L., Arabi, A., Fahlén, S., Hydbring, P., Söderberg, O., Fatyol, K. & Bierhoff, H. (2005) c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription. *Nature Cell Biology.* 7 (3), 303-310. Available from: http://dx.doi.org/10.1038/ncb1225. Available from: doi: 10.1038/ncb1225.

Robustelli, P., Piana, S. & Shaw, D. E. (2018) Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences - PNAS.* 115 (21), E4758-E4766. Available from: https://search.datacite.org/works/10.1073/pnas.1800690115. Available from: doi: 10.1073/pnas.1800690115.

Roe, D. R. & Cheatham, T. E. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation.* 9 (7), 3084-3095. Available from: http://dx.doi.org/10.1021/ct400341p. Available from: doi: 10.1021/ct400341p.

Roohi, A. & Hojjat-Farsangi, M. (2017) Recent advances in targeting mTOR signaling pathway using small molecule inhibitors. *Journal of Drug Targeting.* 25 (3), 189-201. Available from: http://www.tandfonline.com/doi/abs/10.1080/1061186X.2016.1236112. Available from: doi: 10.1080/1061186X.2016.1236112.

Rous, P. (1911) A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. *The Journal of Experimental Medicine.* 13 (4), 397-411. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19867421. Available from: doi: 10.1084/jem.13.4.397.

Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation.* 9 (9), 3878-3888. Available from: http://dx.doi.org/10.1021/ct400314y. Available from: doi: 10.1021/ct400314y.

Scherer, M. K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J. & Noé, F. (2015) PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation.* 11 (11), 5525-5542. Available from: http://dx.doi.org/10.1021/acs.jctc.5b00743. Available from: doi: 10.1021/acs.jctc.5b00743.

Schmidtke, P., Bidon-Chanal, A., Luque, F. J. & Barril, X. (2011) MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics.* 27 (23), 3276-3285. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21967761. Available from: doi: 10.1093/bioinformatics/btr550.

Schrodinger, L. (2010) The PyMOL molecular graphics system. *Version.* 1 (5), 0.

Schwab, M. (2004) MYCN in neuronal tumours. *Cancer Letters.* 204 (2), 179-187. Available from: https://www.sciencedirect.com/science/article/pii/S0304383503004543. Available from: doi: 10.1016/S0304-3835(03)00454-3.

Sears, R. C. (2004) The Life Cycle of C-Myc: From Synthesis to Degradation. *Cell Cycle.* 3 (9), 1131-1135. Available from: http://www.tandfonline.com/doi/abs/10.4161/cc.3.9.1145. Available from: doi: 10.4161/cc.3.9.1145.

Seoane, J., Pouponnot, C., Staller, P., Schader, M., Eilers, M. & Massagué, J. (2001) TGFbeta influences Myc, Miz-1 and Smad to control the CDK inhibitor p15INK4b. *Nature Cell Biology.* 3 (4), 400. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11283614.

Shao, J., Tanner, S. W., Thompson, N. & Cheatham, T. E. (2007) Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation.* 3 (6), 2312-2334. Available from: http://dx.doi.org/10.1021/ct700119m. Available from: doi: 10.1021/ct700119m.

Shen, Y. & Bax, A. (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR.* 48 (1), 13-22. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20628786. Available from: doi: 10.1007/s10858-010-9433-9.

Shenkin, P. S. & McDonald, D. Q. (1994) Cluster analysis of molecular conformations. *Journal of Computational Chemistry.* 15 (8), 899-916. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540150811. Available from: doi: 10.1002/jcc.540150811.

Shi, X., Mihaylova, V. T., Kuruvilla, L., Chen, F., Viviano, S., Baldassarre, M., Sperandio, D., Martinez, R., Yue, P., Bates, J. G., Breckenridge, D. G., Schlessinger, J., Turk, B. E. & Calderwood, D. A. (2016) Loss of TRIM33 causes resistance to BET bromodomain inhibitors through MYC- and TGF-β–dependent mechanisms. *Proceedings of the National Academy of Sciences of the United States of America.* 113 (31), E4558-E4566. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27432991. Available from: doi: 10.1073/pnas.1608319113.

Shi, Y., Glynn, J. M., Guilbert, L. J., Cotter, T. G., Bissonnette, R. P. & Green, D. R. (1992) Role for c-myc in activation-induced apoptotic cell death in T cell hybridomas. *Science.* 257 (5067), 212-214. Available from: http://www.sciencemag.org/cgi/content/abstract/257/5067/212. Available from: doi: 10.1126/science.1378649.

Shimamura, T., Chen, Z., Soucheray, M., Carretero, J., Kikuchi, E., Tchaicha, J. H., Gao, Y., Cheng, K. A., Cohoon, T. J., Qi, J., Akbay, E., Kimmelman, A. C., Kung, A. L., Bradner, J. E. & Wong, K. (2013) Efficacy of BET Bromodomain Inhibition in Kras-Mutant Non-Small Cell Lung Cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research.* 19 (22), 6183-6192. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24045185. Available from: doi: 10.1158/1078-0432.CCR-12-3904.

Sittel, F., Jain, A. & Stock, G. (2014) Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *The Journal of Chemical Physics.* 141 (1), 014111. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25005281. Available from: doi: 10.1063/1.4885338.

Song, D., Luo, R. & Chen, H. (2017) The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *Journal of Chemical Information and Modeling.* 57 (5), 1166-1178. Available from: http://dx.doi.org/10.1021/acs.jcim.7b00135. Available from: doi: 10.1021/acs.jcim.7b00135.

Song, D., Wang, W., Ye, W., Ji, D., Luo, R. & Chen, H. (2017) ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chemical Biology & Drug Design.* 89 (1), 5-15. Available from:

https://onlinelibrary.wiley.com/doi/abs/10.1111/cbdd.12832. Available from: doi: 10.1111/cbdd.12832.

Soucek, L., Buggy, J. J., Kortlever, R., Adimoolam, S., Monclús, H. A., Allende, M. T. S., Swigart, L. B. & Evan, G. I. (2011) Modeling Pharmacological Inhibition of Mast Cell Degranulation as a Therapy for Insulinoma. *Neoplasia.* 13 (11), 1093-IN43. Available from: https://www.clinicalkey.es/playcontent/1-s2.0-S1476558611800953. Available from: doi: 10.1593/neo.11980.

Sreerama, N. & Woody, R. W. (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Analytical Biochemistry.* 287 (2), 252-260.

Stanton, B. R., Perkins, A. S., Tessarollo, L., Sassoon, D. A. & Parada, L. F. (1992) Loss of N-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. *Genes & Development.* 6 (12A), 2235-2247. Available from: https://www.ncbi.nlm.nih.gov/pubmed/1459449. Available from: doi: 10.1101/gad.6.12a.2235.

Steven B. McMahon, Marcelo A. Wood & Michael D. Cole. (2000) The Essential Cofactor TRRAP Recruits the Histone Acetyltransferase hGCN5 to c-Myc. *Molecular and Cellular Biology.* 20 (2), 556-562. Available from: http://mcb.asm.org/content/20/2/556.abstract. Available from: doi: 10.1128/MCB.20.2.556-562.2000.

Stine, Z. E., Walton, Z. E., Altman, B. J., Hsieh, A. L. & Dang, C. V. (2015) MYC, Metabolism, and Cancer. *Cancer Discovery.* 5 (10), 1024-1039. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26382145. Available from: doi: 10.1158/2159-8290.CD-15-0507.

Sullivan, S. S. & Weinzierl, R. O. J. (2020) Optimization of Molecular Dynamics Simulations of c-MYC1-88—An Intrinsically Disordered System. *Life (Basel, Switzerland).* 10 (7), 109. Available from: https://search.proquest.com/docview/2423949083. Available from: doi: 10.3390/life10070109.

Sun, B., Mason, S., Wilson, R. C., Hazard, S. E., Wang, Y., Fang, R., Wang, Q., Yeh, E. S., Yang, M., Roberts, T. M., Zhao, J. J. & Wang, Q. (2020) Inhibition of the transcriptional kinase CDK7 overcomes therapeutic resistance in HER2-positive breast cancers. *Oncogene.* 39 (1), 50-63. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31462705. Available from: doi: 10.1038/s41388-019-0953-9.

Sun, X., Sears, R. C. & Dai, M. (2015) Deubiquitinating c-Myc: USP36 steps up in the nucleolus. *Cell Cycle.* 14 (24), 3786-3793. Available from: http://www.tandfonline.com/doi/abs/10.1080/15384101.2015.1093713. Available from: doi: 10.1080/15384101.2015.1093713.

Svergun, D. I. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography.* 25 (4), 495-503.

Swier, Lotteke J. Y. M, Dzikiewicz-Krawczyk, A., Winkle, M., Berg, A. & Kluiver, J. (2019) Intricate crosstalk between MYC and non-coding RNAs regulates hallmarks of cancer. *Molecular Oncology.* 13 (1), 26-45. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/1878-0261.12409. Available from: doi: 10.1002/1878-0261.12409.

Takahashi, K. & Yamanaka, S. (2006) Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell.* 126 (4), 663-676. Available from: https://www.sciencedirect.com/science/article/pii/S0092867406009767. Available from: doi: 10.1016/j.cell.2006.07.024.

Thomas, L., Wang, Q., Grieb, B., Phan, J., Foshage, A., Sun, Q., Olejniczak, E., Clark, T., Dey, S., Lorey, S., Alicie, B., Howard, G., Cawthon, B., Ess, K., Eischen, C., Zhao, Z., Fesik, S. & Tansey, W. (2015) Interaction with WDR5 Promotes Target Gene Recognition and Tumorigenesis by MYC. *Molecular Cell.* 58 (3), 440-452. Available from: https://www.sciencedirect.com/science/article/pii/S1097276515001422. Available from: doi: 10.1016/j.molcel.2015.02.028.

Thorsten Berg, Steven B. Cohen, Joel Desharnais, Corinna Sonderegger, Daniel J. Maslyar, Joel Goldberg, Dale L. Boger & Peter K. Vogt. (2002) Small-Molecule Antagonists of Myc/Max Dimerization Inhibit Myc-Induced Transformation of Chicken Embryo Fibroblasts. *Proceedings of the National Academy of Sciences of the United States of America.* 99 (6), 3830-3835. Available from: https://www.jstor.org/stable/3058212. Available from: doi: 10.1073/pnas.062036999.

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. (2014) PLUMED 2: New feathers for an old bird. *Computer Physics Communications.* 185 (2), 604-613. Available from: http://dx.doi.org/10.1016/j.cpc.2013.09.018. Available from: doi: 10.1016/j.cpc.2013.09.018.

Trott, O. & Olson, A. J. (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry.* 31 (2), 455-461. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334. Available from: doi: 10.1002/jcc.21334.

Tseng, Y., Moriarity, B. S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P., Reuland, B., Guenther, K., Beadnell, T. C., Essig, J., Otto, G. M., O'Sullivan, M. G., Largaespada, D. A., Schwertfeger, K. L., Marahrens, Y., Kawakami, Y. & Bagchi, A. (2014) PVT1 dependence in cancer with MYC copy-number increase. *Nature.* 512 (7512), 82-86. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25043044. Available from: doi: 10.1038/nature13311.

Tu, W. B., Helander, S., Pilstål, R., Hickman, K. A., Lourenco, C., Jurisica, I., Raught, B., Wallner, B., Sunnerhagen, M. & Penn, L. Z. (2015) Myc and its interactors take shape. *BBA - Gene Regulatory Mechanisms.* 1849 (5), 469-483. Available from: http://dx.doi.org/10.1016/j.bbagrm.2014.06.002. Available from: doi: 10.1016/j.bbagrm.2014.06.002.

Valera, A., Pujol, A., Gregori, X., Riu, E., Visa, J. & Bosch, F. (1995) Evidence from transgenic mice that myc regulates hepatic glycolysis. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology.* 9 (11), 1067-1078. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7649406. Available from: doi: 10.1096/fasebj.9.11.7649406.

Varmus, H. E. (1984) The Molecular Genetics of Cellular Oncogenes. *Annual Review of Genetics.* 18 (1), 553-612. Available from: https://www.ncbi.nlm.nih.gov/pubmed/6397126. Available from: doi: 10.1146/annurev.ge.18.120184.003005.

Vazquez, A., Markert, E. K. & Oltvai, Z. N. (2011) Serine Biosynthesis with One Carbon Catabolism and the Glycine Cleavage System Represents a Novel Pathway for ATP

Generation. *PloS One.* 6 (11), e25881. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22073143. Available from: doi: 10.1371/journal.pone.0025881.

Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002) Small-world view of the amino acids that play a key role in protein folding. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics.* 65 (6 Pt 1), 061910. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12188762. Available from: doi: 10.1103/PhysRevE.65.061910.

Wahlström, T. & Arsenian Henriksson, M. (2015) Impact of MYC in regulation of tumor cell metabolism. *BBA - Gene Regulatory Mechanisms.* 1849 (5), 563-569. Available from: https://www.sciencedirect.com/science/article/pii/S1874939914001928. Available from: doi: 10.1016/j.bbagrm.2014.07.004.

Walz, S., Lorenzin, F., Morton, J., Wiese, K. E., von Eyss, B., Herold, S., Rycak, L., Dumay-Odelot, H., Karim, S., Bartkuhn, M., Roels, F., Wüstefeld, T., Fischer, M., Teichmann, M., Zender, L., Wei, C., Sansom, O., Wolf, E. & Eilers, M. (2014) Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature.* 511 (7510), 483-487. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25043018. Available from: doi: 10.1038/nature13473.

Wang, H., Chauhan, J., Hu, A., Pendleton, K., Yap, J. L., Sabato, P. E., Jones, J. W., Perri, M., Yu, J., Cione, E., Kane, M. A., Fletcher, S. & Prochownik, E. V. (2013) Disruption of Myc-Max Heterodimerization with Improved Cell-Penetrating Analogs of the Small Molecule 10074-G5. *Oncotarget.* 4 (6), Available from: doi: 10.18632/oncotarget.1108.

Wang, X., Cunningham, M., Zhang, X., Tokarz, S., Laraway, B., Troxell, M. & Sears, R. C. (2011) Phosphorylation Regulates c-Myc's Oncogenic Activity in the Mammary Gland. *Cancer Research.* 71 (3), 925-936. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21266350. Available from: doi: 10.1158/0008-5472.CAN-10-1032.

Wang, Y., Zhang, T., Kwiatkowski, N., Abraham, B., Lee, T., Xie, S., Yuzugullu, H., Von, T., Li, H., Lin, Z., Stover, D., Lim, E., Wang, Z., Iglehart, J. ., Young, R., Gray, N. & Zhao, J. (2015) CDK7-Dependent Transcriptional Addiction in Triple-Negative Breast Cancer. *Cell.* 163 (1), 174-186. Available from: http://dx.doi.org/10.1016/j.cell.2015.08.063. Available from: doi: 10.1016/j.cell.2015.08.063.

Wei, W., Zhao, X., Wu, S., Zhao, C., Zhao, H., Sun, L. & Cui, Y. (2018) Dihydroartemisinin triggers c-Myc proteolysis and inhibits protein kinase B/glycogen synthase kinase 3β pathway in T-cell lymphoma cells. *Oncology Letters.* 16 (5), 6838-6846. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30405828. Available from: doi: 10.3892/ol.2018.9450.

Whitfield, J. R., Beaulieu, M. & Soucek, L. (2017) Strategies to Inhibit Myc and Their Clinical Applicability. *Frontiers in Cell and Developmental Biology.* 5 10. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28280720. Available from: doi: 10.3389/fcell.2017.00010.

Whitmore, L. & Wallace, B. A. (2004) DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Research.* 32 (Web Server issue), W668-W673. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15215473. Available from: doi: 10.1093/nar/gkh371.

Whitmore, L. & Wallace, B. A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers.* 89 (5), 392-400. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.20853. Available from: doi: 10.1002/bip.20853.

Wickham, H. (2016) *ggplot2: elegant graphics for data analysis. [google]* Springer.

Wiegering, A., Uthe, F. W., Jamieson, T., Ruoss, Y., Hüttenrauch, M., Küspert, M., Pfann, C., Nixon, C., Herold, S., Walz, S., Taranets, L., Germer, C., Rosenwald, A., Sansom, O. J. & Eilers, M. (2015) Targeting Translation Initiation Bypasses Signaling Crosstalk Mechanisms That Maintain High MYC Levels in Colorectal Cancer. *Cancer Discovery.* 5 (7), 768-781. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25934076. Available from: doi: 10.1158/2159-8290.CD-14-1040.

William P. Tansey. (2014) Mammalian MYC Proteins and Cancer. *New Journal of Science.* 2014 1-27. Available from: http://dx.doi.org/10.1155/2014/757534. Available from: doi: 10.1155/2014/757534.

Williams, T., Kelley, C., Bersch, C., Bröker, H., Campbell, J., Cunningham, R., Denholm, D., Elber, G., Fearick, R. & Grammes, C. (2018) *Gnuplot 5.2* (5.2.4) .

Wriggers, W., Stafford, K. A., Shan, Y., Piana, S., Maragakis, P., Lindorff-Larsen, K., Miller, P. J., Gullingsrud, J., Rendleman, C. A., Eastwood, M. P., Dror, R. O. & Shaw, D. E. (2009) Automated Event Detection and Activity Monitoring in Long Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation.* 5 (10), 2595-2605. Available from: http://dx.doi.org/10.1021/ct900229u. Available from: doi: 10.1021/ct900229u.

Xiang, J., Yin, Q., Chen, T., Zhang, Y., Zhang, X., Wu, Z., Zhang, S., Wang, H., Ge, J., Lu, X., Yang, L. & Chen, L. (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Research.* 24 (5), 513-531. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24662484. Available from: doi: 10.1038/cr.2014.35.

Xin Li, Xin A. Zhang, Xiaoqing Li, Wei Xie & Shiang Huang. (2015) MYC-Mediated Synthetic Lethality for Treating Tumors. *Current Cancer Drug Targets.* 15 (2), 99-115. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1568-0096&volume=15&issue=2&spage=99. Available from: doi: 10.2174/1568009615666150121162921.

Xu, D. & Zhang, Y. (2013) Toward optimal fragment generations for ab initio protein structure assembly. *Proteins: Structure, Function, and Bioinformatics.* 81 (2), 229-239. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24179. Available from: doi: 10.1002/prot.24179.

Xu, D. & Zhang, Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics.* 80 (7), 1715-1735. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24065. Available from: doi: 10.1002/prot.24065.

Xue, G., Yan, H., Zhang, Y., Hao, L., Zhu, X., Mei, Q. & Sun, S. (2015) c-Myc-mediated repression of miR-15-16 in hypoxia is induced by increased HIF-2α and promotes tumor angiogenesis and metastasis by upregulating FGF2. *Oncogene.* 34 (11), 1393-1406. Available

from: https://www.ncbi.nlm.nih.gov/pubmed/24704828. Available from: doi: 10.1038/onc.2014.82.

Yao, W., Yue, P., Khuri, F. R. & Sun, S. (2015) The BET bromodomain inhibitor, JQ1, facilitates c-FLIP degradation and enhances TRAIL-induced apoptosis independent of BRD4 and c-Myc inhibition. *Oncotarget.* 6 (33), 34669-34679. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26415225. Available from: doi: 10.18632/oncotarget.5785.

Yap, J. L., Wang, H., Hu, A., Chauhan, J., Jung, K., Gharavi, R. B., Prochownik, E. V. & Fletcher, S. (2013) Pharmacophore identification of c-Myc inhibitor 10074-G5. *Bioorganic & Medicinal Chemistry Letters.* 23 (1), 370-374. Available from: http://dx.doi.org/10.1016/j.bmcl.2012.10.013. Available from: doi: 10.1016/j.bmcl.2012.10.013.

Ye, J., Fan, J., Venneti, S., Wan, Y., Pawel, B. R., Zhang, J., Finley, L. W. S., Lu, C., Lindsten, T., Cross, J. R., Qing, G., Liu, Z., Simon, M. C., Rabinowitz, J. D. & Thompson, C. B. (2014) Serine Catabolism Regulates Mitochondrial Redox Control during Hypoxia. *Cancer Discovery.* 4 (12), 1406-1417. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25186948. Available from: doi: 10.1158/2159-8290.CD-14-0250.

Yin, X., Giap, C., Lazo, J. S. & Prochownik, E. V. (2003) Low molecular weight inhibitors of Myc-Max interaction and function. *Oncogene.* 22 (40), 6151-6159. Available from: http://dx.doi.org/10.1038/sj.onc.1206641. Available from: doi: 10.1038/sj.onc.1206641.

Zhang, N., Ichikawa, W., Faiola, F., Lo, S., Liu, X. & Martinez, E. (2014) MYC interacts with the human STAGA coactivator complex via multivalent contacts with the GCN5 and TRRAP subunits. *BBA - Gene Regulatory Mechanisms.* 1839 (5), 395-405. Available from: http://dx.doi.org/10.1016/j.bbagrm.2014.03.017. Available from: doi: 10.1016/j.bbagrm.2014.03.017.

Zhang, Q., West-Osterfield, K., Spears, E., Li, Z., Panaccione, A. & Hann, S. R. (2017) MB0 and MBI Are Independent and Distinct Transactivation Domains in MYC that Are Essential for Transformation. *Genes.* 8 (5), 134. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28481271. Available from: doi: 10.3390/genes8050134.

Zhang, X., Zhao, X., Fiskus, W., Lin, J., Lwin, T., Rao, R., Zhang, Y., Chan, J., Fu, K., Marquez, V., Chen-Kiang, S., Moscinski, L., Seto, E., Dalton, W., Wright, K., Sotomayor, E., Bhalla, K. & Tao, J. (2012) Coordinated Silencing of MYC-Mediated miR-29 by HDAC3 and EZH2 as a Therapeutic Target of Histone Modification in Aggressive B-Cell Lymphomas. *Cancer Cell.* 22 (4), 506-523. Available from: http://dx.doi.org/10.1016/j.ccr.2012.09.003. Available from: doi: 10.1016/j.ccr.2012.09.003.

Zhong, C., Fan, L., Yao, F., Shi, J., Fang, W. & Zhao, H. (2014) HMGCR is necessary for the tumorigenecity of esophageal squamous cell carcinoma and is regulated by Myc. *Tumor Biology.* 35 (5), 4123-4129. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24390662. Available from: doi: 10.1007/s13277-013-1539-8.

Supplementary Information

Figure S1 compares the total secondary structure content created by the trajectories with the ones obtained by NMR.

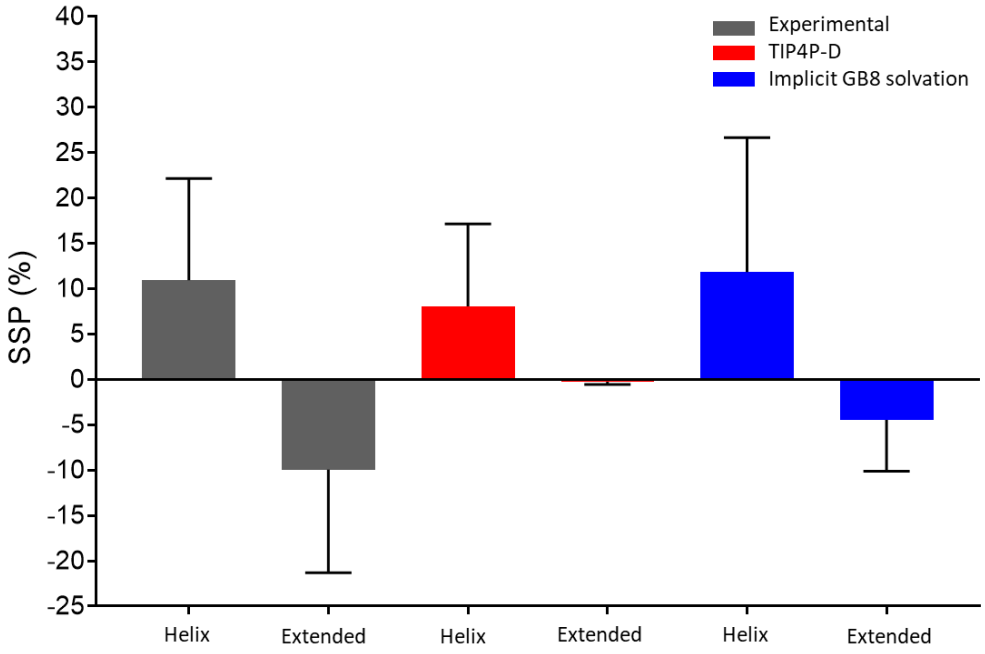

Figure S1. Comparison of the helical and extended SSPs between the NMR-determined SSP and the SSP values predicted from the explicit TIP4P-D and the implicit GB8 solvation models.

From the SSP results it is evident that whilst both GB8 and TIP4P-D accurately replicate the helical content of MYC88, only GB8 correctly describes the extended content and TIP4P-D dramatically underestimates β-sheet propensities (p-value <0.0001).

Figure S2 shows the free energy landscape PCA plot for the MYC88 trajectory structures defined in terms of their RMSD, Rg, SASA, distance end-to-end and hydrogen bond number.

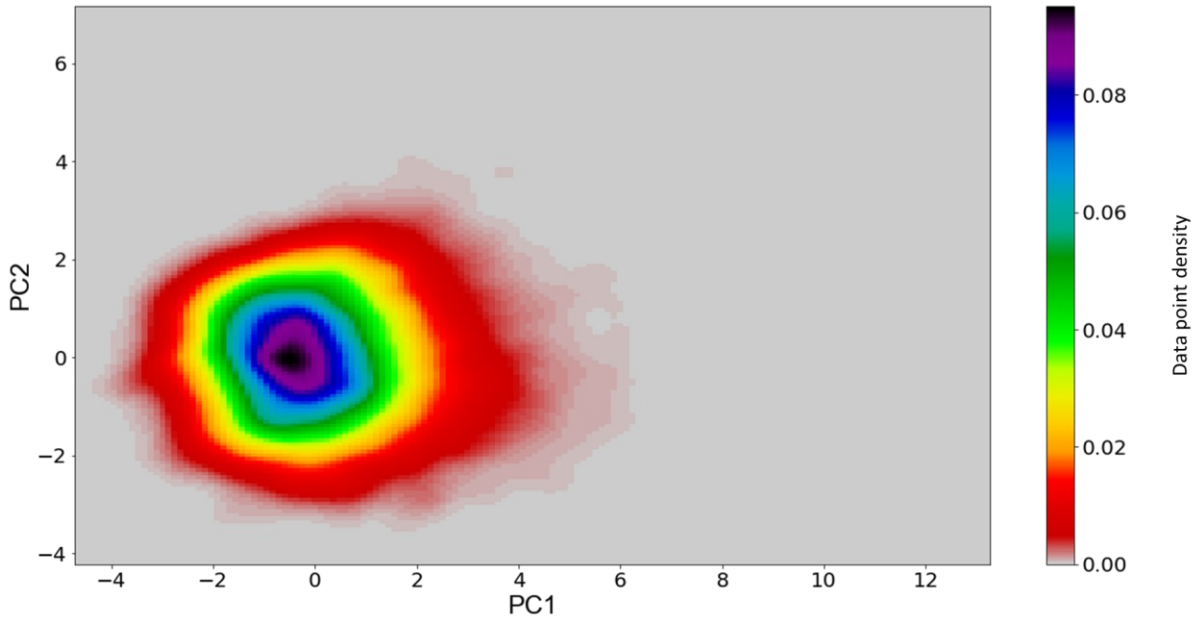

Figure S2. PCA plot for the MYC88 MD simulation defined by simple geometrics: RMSD, Rg, SASA, distance end-to-end and the number of hydrogen bonds.

It is evident that these metrics do not a PCA landscape with discernible clusters. Therefore, **Figure S3** shows the PCA plot for the MYC88 MD simulation described in terms of its secondary structure content.

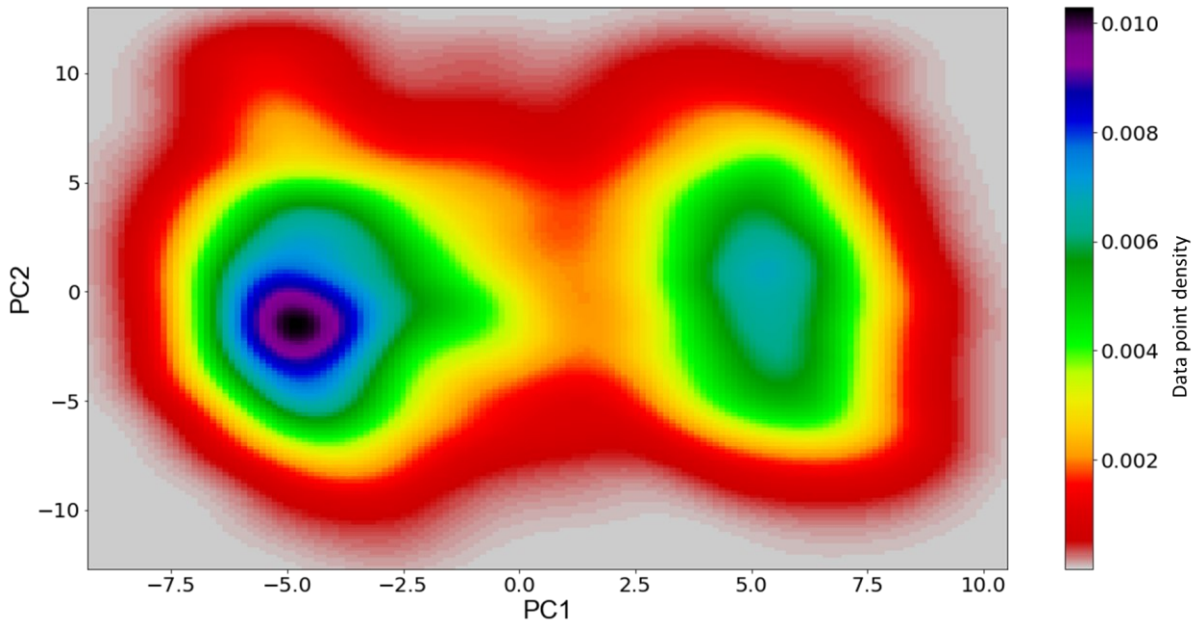

Figure S3. PCA plot for the MYC88 MD simulation defined by its secondary structure content.

Although the PCA plot for MYC88 secondary structure displays two large cluster of structures, the explanative power of the principal components (**Table S1**) is so low that these clusters cannot be used to make predictive conclusions about the data - cumulatively, the first two PCs account for only 7% of the explained variance.

Table S1. MYC88 secondary structure PCA - explained variance and cumulative explained variance for the first 6 principal components.

	Variance (%)	Cumulative (%)
PC 1	3.8	3.8
PC 2	3.2	7.0
PC 3	3.1	10.1
PC 4	2.8	12.9
PC 5	2.6	15.5
PC 6	2.4	17.9

The internal coordinates of inter-atom distances between alpha carbons were also used to 'featurise' the PCA calculation **(Figure S4)**. However, even this metric does not yield a PCA landscape with distinct clusters.

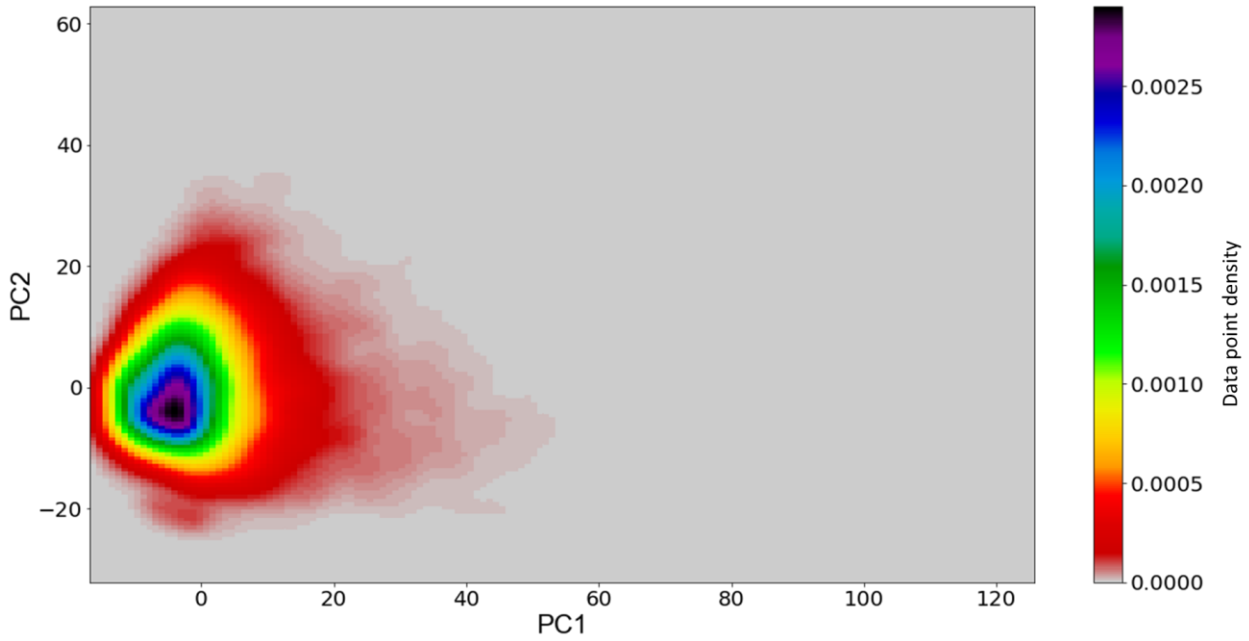

Figure S4. PCA plot for the MYC88 MD simulation defined by its inter-atom alpha carbon distances.

Figure S5 displays the chemical structures for the 6 ligands identified by the drug discovery process to bind MYC88 with great affinity and stability.

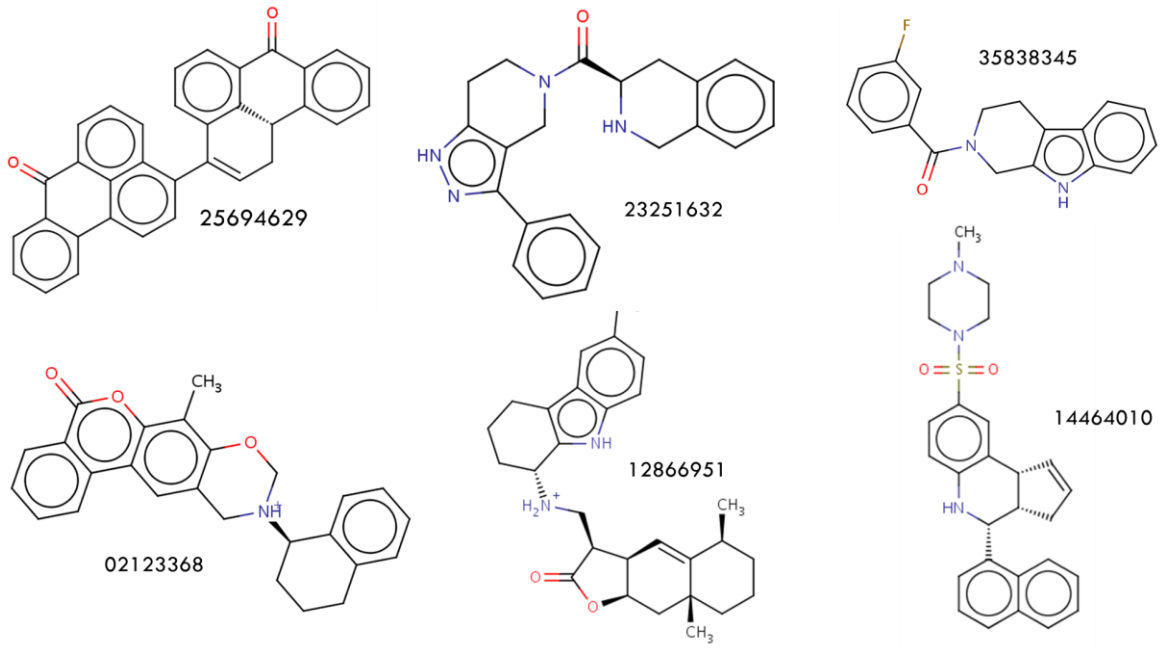

Figure S5. Chemical structures of the 6 identified ligands.

Figure S6 shows he implied timescales used to derive the lag time.

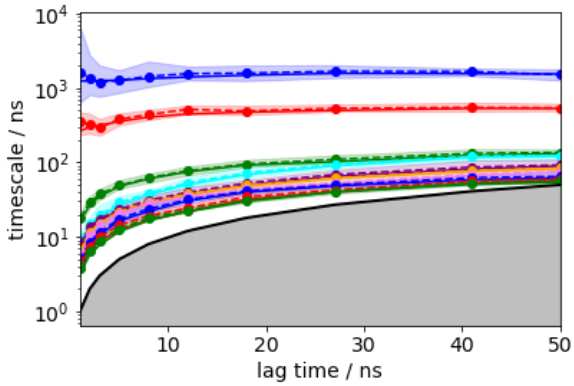

Figure S6. Implied timescales plot showing the MSM timescales converge to the true relaxation timescales with increasing lag time.

Figure S7 shows a contour plots of each PCCA++ metastable states assignments.

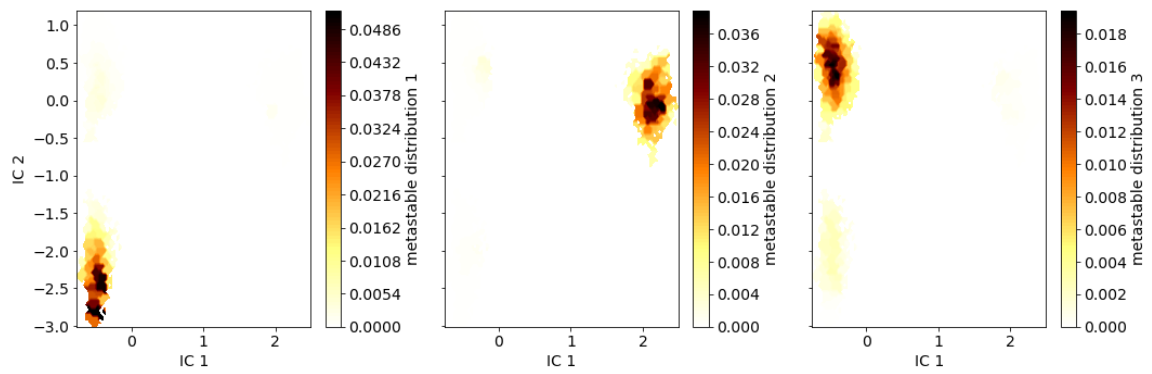

Figure S7. PCCA++ metastable states.

Figure S8 shows an example of a TLeap input file which parameterises the simulation to use the ff14SB force field solvated with the explicit the TIP3-P water model.

```
#Call force field and water model
source leaprc.protein.ff14SB
 source leaprc.gaff2
 source leaprc.water.tip3p

 #Load input file
 test = loadpdb “test.pdb"
 check test

 #Adding the periodic solvation box with a boundary of 15 angstroms
solvatebox test TIP3PBOX 15.0

#Neutralisation set
 addions complex NA 0
 addions complex CL 0

#Adding ions to a final concentration of 150nM
 addions complex NA 76
 addions complex CL 76

#Saving files as topology and input
 saveamberparm complex histatin5.prmtop histatin5.inpcrd
 quit
```

Figure S8. TLeap system preparation example file.

Figure S9 shows examples for the input files used to run conventional MD simulations (cMD), including the two minimisations followed by two production runs. And **Figure S10** shows the input file for accelerated MD simulation.

```
#MIN1
initial minimisation solvent + ions
 &cntrl
  imin   = 1,
  maxcyc = 10000,
  ncyc   = 5000,
  ntb    = 1,
  ntr    = 1,
  cut    = 10.0
 /
Hold the protein fixed
500.0
RES 1 1000
END
END

#MIN2
initial minimisation whole system
&cntrl
 imin   = 1,
 maxcyc = 2500,
 ncyc   = 1000,
 ntb    = 1,
 ntr    = 0,
 cut    = 10.0
/
```

```
#MD1
100ps MD with res on protein
 &cntrl
  imin   = 0,
  irest  = 0,
  ntx    = 1,
  ntb    = 1,
  cut    = 10.0,
  ntr    = 1,
  ntc    = 2,
  ntf    = 2,
  tempi  = 0.0,
  temp0  = 310.0,
  ntt    = 3,
  gamma_ln = 1.0,
  nstlim = 50000, dt = 0.002,
  ntpr = 100, ntwx = 100, ntwr = 1000
 /
Keep protein fixed with weak restraints
10.0
RES 1 1000
END
END

#MD2
 taf2: 1000ns MD
 &cntrl
  imin = 0, irest = 1, ntx = 7,
  ntb = 2, pres0 = 1.0, ntp = 1,
  taup = 2.0, ig=-1, ioutfm=1,
  cut = 10.0, ntr = 0,
  ntc = 2, ntf = 2,
  tempi = 310.0, temp0 = 310.0,
  ntt = 3, gamma_ln = 1.0,
  nstlim = 500000000, dt = 0.002,
  ntpr = 25000, ntwx = 25000, ntwr = 25000
 /
```

Figure S9. Examples of cMD input files used to perform conventional MD simulations.

```
#aMD
1000ns aMD
 &cntrl
  imin = 0, irest = 1, ntx = 5,
  dt = 0.002, ntc = 2, ntf = 2, tol=0.000001, iwrap=1,
  ntb = 2, cut = 8.0, ntp=1, igb=0, ntwprt=0, ioutfm=1,
  ntt = 3, temp0=310.0, gamma_ln = 1.0, ig=-1,
  ntpr = 500000, ntwx = 500000, ntwr = -10000000, nstlim = 500000000,
  iamd=3,
  EthreshD=369, alphaD=18,
  EthreshP=-265282, alphaP=16857,
 /
&ewald
dsum_tol=0.000001,
 /
```

Figure S10. aMD input file example. Note that EthreshD, alphaD, EthreshP and alphaP values are simulation specific and obtained from the cMD simulation.

Figure S11 contains an example of analysis script for CPPTRAJ.

```
#load trajectory

parm test.pdb
trajin test.dcd

reference ref.pdb [strut]

#Align trajectory to its first frame
rms first !@H=

#Strip trajectory of water molecules and ions
strip :WAT
strip :NA
strip :CL

#Calculate Rg, distance, RMSD and RMSF
radgyr :1-88@CA out gyration_test.dat mass nomax
distance end-to-end :1 :88 out distance_test.dat
rms Strut ref [strut] :1-88&!@H= out rmsd_test.dat mass
atomicfluct out RMSF_test.dat :1-88@CA byres

#Calculate solvent accessible surface area
surf :1-88 out surf_test.dat

#secondary structure calculation
secstruct :1-88 out dssp_test.dat assignout assign_test.dat sumout
sum_test.dat

#calculating backbone hydrogen bonds
hbond All out hbond_test.dat

#Calculation of dihedral angles
multidihedral phi psi resrange 1-88 out phipsi_diheds_test.dat
```

Figure S11. CPPTRAJ input file example with instructions for RMSD, Rg, distance end-to-end, solvent accessible surface area, secondary structure, hydrogen bond and dihedral angles calculation.

Figure S12 shows an example of PCA calculation in R using the BIO3D package.

```
#load library
library(bio3d)

#Load trajectory and topology
dcd <- read.dcd("test.dcd")
pdb <- read.pdb("test.pdb")

#Select c-alpha atoms for calculation
ca.inds <- atom.select(pdb, elety="CA")
xyz <- fit.xyz(fixed=pdb$xyz, mobile=dcd,
          fixed.inds=ca.inds$xyz,
          mobile.inds=ca.inds$xyz)

#PCA calculation
pc <- pca.xyz(xyz[,ca.inds$xyz])
plot(pc, col=bwr.colors(nrow(xyz)) )

#Output PC1 and PC2 maximum amplitude motion
p1 <- mktrj.pca(pc, pc=1, b=pc$au[,1], file="pc1.pdb")
p2 <- mktrj.pca(pc, pc=2,b=pc$au[,2], file="pc2.pdb")
```

Figure S12. R script for PCA calculation using the BIO3D analysis package.

Figure S13 shows an example of Markov-chain Monte Carlo simulation input file.

```
#Setting up a simulation with the amino acid sequence as input
#With profasi-cached as the force field
#Allowing uniform sidechain movement
#Using Metropolis-Hastings acceptance algorithm
#Output observables aligned to reference structure

./phaistos --aa-file test.aa --energy profasi-cached --move pivot-uniform sidechain-
uniform --threads 25  --monte-carlo metropolis-hastings --observable backbone-dbn
rmsd[reference-pdb-file:test.pdb] --observable xtc-trajectory
```

Figure S13. Markov-chain Monte Carlo simulation input.

Figure S14 shows an example of the maximum and minimum peak calculation for Rg linear data using Python and argrelextrema package.

```
#Load libraries and data
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from scipy.signal import argrelextrema


n=500 # Rolling window of number of points to be checked before and after
# Find local peaks
df1['min'] = df1.iloc[argrelextrema(df1.Rg.values, np.less_equal, order=n)[0]]['Rg']
df1['max'] = df1.iloc[argrelextrema(df1.Rg.values, np.greater_equal, order=n)[0]]['Rg']

#Export data to csv file
export = df1.to_csv(r'Rg_min_max.csv')

# Plot results
f, ax = plt.subplots(figsize=(15, 8))
plt.scatter(df1.index, df1['min'], c='r', s=100)
plt.scatter(df1.index, df1['max'], c='g', s=100)
plt.plot(df1.index, df1['Rg'])
plt.xticks(np.arange(0, 24000, step=1000))
plt.xticks(fontsize=24, rotation='vertical')
plt.yticks(fontsize=24)
#plt.grid(True)
plt.ylabel('Rg (Å)', fontsize=24)
plt.xlabel('Time (ns)', fontsize=24)
plt.show();
```

Figure S14. Python script for peak detection while removing local noise.

Figure S15 shows the different stages for the calculation of TICA landscape and extraction of the representative structures.

```
#load libraries
import pyemma
from pyemma.util.contexts import settings
plt.matplotlib.rcParams.update({'font.size': 16})

#Load the data files
pdb = 'test.pdb'
files = 'test.dcd'

#Featurise the data, in this case choosing backbone torsions
feat = pyemma.coordinates.featurizer(pdb)
feat.add_backbone_torsions(cossin=True, periodic=False)
data = pyemma.coordinates.load(files, features=feat)

#Calculate and plot TICA's first two IC's
tica = pyemma.coordinates.tica(data, lag=20, dim=2)
tica_output = tica.get_output()
tica_concatenated = np.concatenate(tica_output)
pyemma.plots.plot_free_energy(*np.concatenate(tica_output).T, legacy=False)

#Calculate and assign structures to clusters, map over TICA landscape
cluster = pyemma.coordinates.cluster_kmeans(
    tica_output, k=200, max_iter=50, stride=5)
dtrajs_concatenated = np.concatenate(cluster.dtrajs)
fig, ax = plt.subplots(figsize=(5, 4))
pyemma.plots.plot_density(
    *tica_concatenated[:, :2].T, ax=ax, cbar=False, alpha=0.3)
ax.scatter(*cluster.clustercenters[:, :2].T, s=5, c='C1')
ax.set_xlabel('IC 1')
ax.set_ylabel('IC 2')
fig.tight_layout()

#Implied timescales calculation
its = pyemma.msm.its(cluster.dtrajs, lags=50, nits=10, errors='bayes')
pyemma.plots.plot_implied_timescales(its, units='ns', dt=1);
msm = pyemma.msm.bayesian_markov_model(cluster.dtrajs, lag=18, dt_traj='1 ns')
print('fraction of states used = {:.2f}'.format(msm.active_state_fraction))
print('fraction of counts used = {:.2f}'.format(msm.active_count_fraction))

#Chapman-Kolmogorov validation
nstates = 3
cktest = msm.cktest(nstates, mlags=6)
pyemma.plots.plot_cktest(cktest, dt=1, units='ns');
```

```
#Spectral analysis
def its_separation_err(ts, ts_err):
    """
    Error propagation from ITS standard deviation to timescale
separation.
    """
    return ts[:-1] / ts[1:] * np.sqrt(
        (ts_err[:-1] / ts[:-1])**2 + (ts_err[1:] / ts[1:])**2)


nits = 15

timescales_mean = msm.sample_mean('timescales', k=nits)
timescales_std = msm.sample_std('timescales', k=nits)

fig, axes = plt.subplots(1, 2, figsize=(10, 4))

axes[0].errorbar(
    range(1, nits + 1),
    timescales_mean,
    yerr=timescales_std,
    fmt='.', markersize=10)
axes[1].errorbar(
    range(1, nits),
    timescales_mean[:-1] / timescales_mean[1:],
    yerr=its_separation_err(
        timescales_mean,
        timescales_std),
    fmt='.',
    markersize=10,
    color='C0')

for i, ax in enumerate(axes):
    ax.set_xticks(range(1, nits + 1))
    ax.grid(True, axis='x', linestyle=':')
axes[0].axhline(msm.lag * 0.1, lw=1.5, color='k')
axes[0].axhspan(0, msm.lag * 0.1, alpha=0.3, color='k')
axes[0].set_xlabel('implied timescale index')
axes[0].set_ylabel('implied timescales / ns')
axes[1].set_xticks(range(1, nits))
axes[1].set_xticklabels(
    ["{:d}/{:d}".format(k, k + 1) for k in range(1, nits + 2)],
    rotation=45)
axes[1].set_xlabel('implied timescale indices')
axes[1].set_ylabel('timescale separation')
fig.tight_layout()


fig, axes = plt.subplots(1, 2, figsize=(14, 6), sharex=True,
sharey=True)
pyemma.plots.plot_contour(
    *tica_concatenated[:, :2].T,
    msm.pi[dtrajs_concatenated],
    ax=axes[0],
    mask=True,
    cbar_label='stationary distribution')
pyemma.plots.plot_free_energy(
    *tica_concatenated[:, :2].T,
    weights=np.concatenate(msm.trajectory_weights()),
    ax=axes[1],
    legacy=False)
for ax in axes.flat:
    ax.set_xlabel('IC 1')
axes[0].set_ylabel('IC 2')
axes[0].set_title('Stationary distribution', fontweight='bold')
axes[1].set_title('Reweighted free energy surface',
fontweight='bold')
fig.tight_layout()

eigvec = msm.eigenvectors_right()
print('The first eigenvector is one: {} (min={}, max={})'.format(
    np.allclose(eigvec[:, 0], 1, atol=1e-15), eigvec[:, 0].min(),
eigvec[:, 0].max()))

fig, axes = plt.subplots(1, 3, figsize=(15, 3), sharex=True,
sharey=True)
for i, ax in enumerate(axes.flat):
    pyemma.plots.plot_contour(
        *tica_concatenated[:, :2].T,
        eigvec[dtrajs_concatenated, i + 1],
        ax=ax,
        cmap='PiYG',
        cbar_label='{}. right eigenvector'.format(i + 2),
        mask=True)
    ax.set_xlabel('IC 1')
axes[0].set_ylabel('IC 2')
fig.tight_layout()
```

```
#Perron cluster analysis and metas
structures
msm.pcca(nstates)

fig, axes = plt.subplots(1, 3, figsize=
sharey=True)
for i, ax in enumerate(axes.flat):
    pyemma.plots.plot_contour(
        *tica_concatenated[:, :2].T,
        msm.metastable_distribution
        ax=ax,
        cmap='afmhot_r',
        mask=True,
        method='nearest',
        cbar_label='metastable distrib
    ax.set_xlabel('IC 1')
axes[0].set_ylabel('IC 2')
fig.tight_layout()

metastable_traj =
msm.metastable_assignments[dtr

fig, ax = plt.subplots(figsize=(5, 4))
_, _, misc = pyemma.plots.plot_sta
    *tica_concatenated[:, :2].T, meta
ax.set_xlabel('IC 1')
ax.set_ylabel('IC 2')
misc['cbar'].set_ticklabels([r'$\mat
                    for i in range(nsta
fig.tight_layout()

pcca_samples =
msm.sample_by_distributions(ms
50)
torsions_source = pyemma.coordi
features=feat)
pyemma.coordinates.save_trajs(
    torsions_source,
    pcca_samples,
    outfiles=['pcca{}.pdb'.format(n +
            for n in range(msm.n_met
```

Figure S15. TICA analysis script using the Pyemma package.