Investigating the Ligandability of Plant Homeodomains

David Bowkett

Keble College

Trinity Term 2015

A thesis submitted to the Medical Sciences Division, University of Oxford in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

Investigating the Ligandability of Plant Homeodomains

David Bowkett, Keble College

Submitted for the degree of Doctor of Philosophy, Trinity 2015

Plant homeodomains (PHDs) are protein domains which bind to the *N*-terminal region of histone 3 (H3) in a manner dependent on the post-transitional modification state of residues near the *N*-terminus of H3. For example some PHDs will only bind H3 if lysine 4 (K4) is trimethylated, whereas other PHDs will only bind H3 if K4 is unmodified. PHDs play a crucial role in gene regulation in cells and mutations and translocations of PHDs have been linked to disease states.

A search of the human proteome revealed the presence of 173 PHDs. These were aligned using structural information and the alignment used to create a PHD phylogenetic tree, the first time such a tree has been reported for this domain family.

All PHD structures in the Protein Data Bank were computationally analysed using SiteMap in order to assess ligandability. Single PHDs were generally found to be poorly ligandable, but some were predicted to be at the lower threshold of ligandability. The study was expanded and found that PHDs found in tandem with another PHD or other epigenetic reader domain were generally more ligandable due to the formation of ligandable pockets at domain-domain interfaces.

Based on the results from the SiteMap analysis, experimental work was carried out to discover ligands of the tandem PHD-PHD of double PHD finger protein 2 (DPF2) and the PHD-JmjC of PHD finger protein 8 (PHF8). Assays were designed and both virtual and experimental screens conducted. No hits were identified for DPF2, but a single fragment hit for PHF8 was found with an IC_{50} of 260 μM. Analogues of this fragment were synthesised and tested, but none were more potent than the initial hit.

Acknowledgments

Many, many people provided help, support, and encouragement that made the completion of this thesis possible. I make no apologies for the length of this section, or for getting a bit emotional near the end.

I would first like to thank Paul Brennan for his supervision and guidance during the course of this project, and for the promoting a friendly and welcoming attitude within the group. I would also like to thank Paul for the many hours he spend proof reading my work; however, his “half a pint per silly mistake” rule has resulted in me racking up quite large bar bill.

Brian Marsden helped to spark my interest in PHDs and provided help and supervision for the work presented in Chapter 2. I am very grateful to Brian for this, and also for the time he spend proof reading this chapter and his well thought-out suggestions.

I would like to thank Gordon Bruton for warm words and thoughtful insight, which proved invaluable when choosing the direction of my studies. I would also like to thank Chris Schofield for making sure I felt welcome in his lab and at his group meetings, as well as for the time spent proof reading my work.

All members of the Med Chem group (past and present) were generous with their help and also their friendship. Alessio, Andrea, Tamas, Filippo, Melissa, Alex, Stephane, Kat, Stephen, Katherine, George, Mattias, Milan, Miranda, Moses, Finn, Oakley, Aicha, Julia, Dong, Lin, Pavel, and Phil, you’ve all helped create the great working atmosphere that I’ve enjoyed so much over the last three years. I will miss the highly regimented routine of 12 pm lunch followed swiftly by 12.40 coffee, and the high standard of food brought in for group meetings.

Anthony Bradley, Jia Tsing Ng, and David Damerell provided invaluable help and advice with coding, which really helped the work presented in Chapters 2 and 3. I am very grateful for their patience in the face of relentless basic questions.

I would like to thank all those who helped teach me how to produce protein. Protein production was only possible thanks to the members of the SGC who answered my questions and passed on their experience. In particular Eidarus Salah, Sebastian Mathea, Kamal Abdul Azeez, Pavel Savitsky, Cynthia Tallant, Chela Nunez Alonso, Beth Jose, Radek Nowak, Stephanie Oerum, and Carina Gileadi were all generous with the time, wisdom, and reagents.

I am very grateful to Oleg Fedorov, Octovia Monteiro, Sarah Martin, Tony Tumber, Giuseppe Scozzafava, and James Bennett for making me feel welcome in their screening lab and for being generous with their advice (and buffers). I would particularly like to thank Oleg for his help interpreting results, and Octovia for showing me how all the pieces of equipment work.

Others who deserve thanks include Rod Chalk and George Berridge for their help with mass spectrometry; Tobias Krojer, Patrick Collins, and Romain Talon for sharing with me the secrets of protein crystallisation and helping me collect and process X-ray diffraction data; Daniel Ebner and Alison Howarth for sharing their compound library and for helping me pick out the compounds I needed; Adam Hardy and Ivan Leung for helping me attempt to produce ^{15}N-labelled protein for NMR work; Finn Wolfreys for his help with PHF8 assays; and finally Andrea Keyte, Ling Jinks, Omar Rivas Rosiles, Ross MacRae and all the others who help keep the building standing and the lab running smoothly.

I would like to thank Charlotte Deane, and all the staff and students at the Doctoral Training Centre who helped me learn so much during my first few months of my DPhil.

I am grateful to the EPSRC and GSK for funding my stipend and paying my fees, and I'm also grateful to Keble College for providing travel grants.

I would like to thank all the clever, talented people I have met over the last four years through the DTC, SGC, TDI, and Keble College. You have helped create a fantastic environment in which to learn and conduct research, and were great people to play football, cricket, and go to the pub

with. I have met many people who I hope will stay friends for life. In particular I would like to thank Jia, Anthony, and Radek for being excellent companions during the DPhil journey.

Finally, I would like to thank family for their love and support, and for providing so many outstanding role models in life. I would particularly like to thank my father for always believing in me and always supporting me in whatever I chose to do. Last, but not least, I would like to thank my partner, Jenny, who has been their through all the highs and lows, and never wavered in her support. I wouldn't have been able to do this without her.

List of Technical Contributions to This Thesis Made by Others

Construct design and cloning for the tandem PHDs of DPF2 and PHF10 was performed by Pavel Savitsky. DNA plasmids were transformed into Rosetta competent cells and provided as stocks frozen in 25% glycerol.

Screening of peptides for use in the DPF2 AlphaScreen assay using Biolayer Interferometry was carried out alongside Oleg Fedorov.

Construct design and cloning for DPF3b was performed by the Jinrong Min group (Structural Genomics Consortium, University of Toronto). This group provided plasmid DNA of the relevant construct, information on crystallisation conditions, and a structure of DPF3b which was used to solve the structure of DPF3b by molecular replacement.

Rosetta and Mach1 competent E.coli cells used in plasmid DNA and protein production respectively were prepared and provided by Kamal Abdul Azeez.

Attempted production of ^{15}N-labelled DPF2 was performed under the supervision of Ivan Leung and Adam Hardy. A final attempt at production of ^{15}N-labelled DPF2 using labelled rich media was performed solely by Ivan Leung.

Construction of a DPF2 library informed by molecular dynamics and virtual screening was performed by Jan Domanski.

PHF8 protein for AlphaScreen and RapidFire assays was prepared by Finn Wolfreys. The PHF8 AlphaScreen assay used in Chapter 5 was designed and optimised by Octovia Monteiro. The RapidFire mass spectrometry assay used in Chapter 5 was designed and optimised by Finn Wolfreys. RapidFire assays were carried out alongside Finn Wolfreys.

Crystallisation experiments, crystal mounting, data collection, and data analysis of DPF3b were carried out by the author under the guidance of Tobias Krojer and Patrick Collins and with the assistance of Radek Nowak.

Table of Contents

List of Abbreviations

2-Oxoglutarate	2OG
3-[(3-Cholamidopropyl)Dimethylammonio]-1-Propanesulfonate	CHAPS
4-(2-Hydroxyethyl)-1-Piperazineethanesulfonic Acid	HEPES
5-Carboxylcytosine	5-caC
5-Formylcytosine	5-fC
5-Hydroxymethylcytosine	5-hmC
5-Methylcytosine	5-mC
6-Methoxy-(8-*P*-Toluenesulfonamido)Quinolone	TSQ
Absent Small Or Homeotic-Like 1	ASH1L
Absent Small Or Homeotic-Like 2	ASH2L
Acetyl Co-Enzyme A	Ac-CoA
Acute Myeloid Leukaemia	AML
Adsorption, Distribution, Metabolism, Excretion, And Toxicity	ADMET
Alpha-Thalassemia And Mental Retardation, X-Linked Syndrome	ATRX
Arginine Methyltransferase	PRMT
ATPase Family AAA Domain Containing 2	ATAD2
ATRX-DNMT3-DNMT3L	ADD
Autoimmune Polyendocrinopathy Candidiasis Ectodermal Dystrophy	APECED
Autoimmune Polyglandular Syndrome Type 1	APS-1
Autoimmune Regulator Protein	AIRE
Basic Local Alignment Search Tool	BLAST
Biolayer Interferometry	BLI
Borjeson-Forssman-Lehmann Syndrome	BFLS
Bovine Serum Albumin	BSA
Bromodomain Adjacent To Zinc Finger Domain 1B	BAZ1B
Bromodomain And PHD Finger Containing Protein 1	BRPF1
Bromodomain And PHD Transcription Factor	BPTF
Bromodomain Protein 1	BRD1
Camp Response Element-Binding Protein-Binding Protein	CREBBP
Chromatin Immunoprecipitation	ChIP
Chromobox Homologue 7	CBX7
Chromodomain Helicase DNA Binding Protein 1	CHD1
Chromodomain Helicase DNA Binding Protein 4	CHD4
Chromodomain Helicase DNA Binding Protein 5	CHD5
Chromodomain Protein 1	Chp1
Class Switch Recombination	CSR
Dimethyl Sulfoxide	DMSO
DNA Methyltransferase	DNMT
DNA Methyltransferase 3-Like	DNMT3L
Double PHD Finger Protein 1	DPF1
Double PHD Finger Protein 2	DPF2
Double PHD Finger Protein 3b	DPF3B
E1A Binding Protein P300	EP300
Ethyl	Et

Euchromatic Histone-Lysine *N*-Methyltransferase 2	EHMT2
European Synchrotron Radiation Facility	ESRF
Flavin Adenine Dinucleotide	FAD
US Food and Drug Administration	FDA
Formaldehyde Dehydrogenase	FDH
Gcn5-Related *N*-Acetyltransferase	GNAT
Glutathione S-Transferase	GST
Haematopoietic Stem Cell	HSC
Hepatoma-Derived Growth Factor-Related Protein 2	HDGF2
Hepatoma-Derived Growth Factor-Related Protein 2	HDGFRP2
High Resolution Mass Spectrometry	HRMS
High Throughput Screening	HTS
Histone 2A	H2A
Histone 2B	H2B
Histone 3	H3
Histone 4	H4
Histone Acetyl Transferase	HAT
Histone Deacetylase	HDAC
Histone Methyltransferase	HMT
Horseradish Peroxidase	HRP
Inhibitor Of Growth 5	ING5
Inhibitor Of Growth 1	ING1
Isopropyl B-D-1-Thiogalactopyranoside	IPTG
Isothermal Calorimetry	ITC
Jumonji Domain 2	JMJD2
Lethal (3) Malignant Brain Tumour-Like 1	L3MBTL1
Liquid Chromatography/Mass Spectrometry	LC/MS
Lysine Acetyl Transferase	KAT
Lysine Demethylase	KDM
Lysine Specific Demethylase	LSD
Lysine-Specific Demethylase 7A	KDM7A
Malignant Brain Tumour	MBT
Mass Spectrometry	MS
Messenger RNA	mRNA
Metal Response Element Binding Transcription Factor 2	MTF2
Methanesulfonyl Chloride	MsCl
Methyl	Me
Mixed-Lineage Leukaemia	MLL
Mixed-Lineage Leukaemia 2	MLL2
Mixed-Lineage Leukaemia 3	MLL3
Mixed-Lineage Leukaemia 4	MLL4
Mixed-Lineage Leukaemia 5	MLL5
Molecular Dynamics	MD
Molecular Evolutionary Genetics Analysis	MEGA
Mouse Embryonic Fibroblast	MEF
Moz, Ybf2, Sas2, Tip60	MYST

Multiple Sequence Alignment	MSA
National Center For Biotechnology Information	NCBI
Nickel-Nitrilotriacetic Acid	Ni-NTA
Nicotinamide Adenine Dinucleotide	NAD^+
N-Oxalyl-D-Cysteine	DNOC
N-Oxalylglycine	NOG
Nuclear Magnetic Resonance	NMR
Nuclear Overhauser Effect	NOE
Nuclear Receptor Binding SET Domain Protein 1	NSD1
Nucleoporin 98	NUP98
Pan-Assay Interference Compound	PAIN
PHD Finger Protein 1	PHF1
PHD Finger Protein 10	PHF10
PHD Finger Protein 19	PHF19
PHD Finger Protein 2	PHF2
PHD Finger Protein 20	PHF20
PHD Finger Protein 20-Like	PHF20L
PHD Finger Protein 6	PHF6
PHD Finger Protein 8	PHF8
Phosphate Buffered Saline	PBS
Plant Homeodomain	PHD
Polyethylene Glycol	PEG
Polymerase Chain Reaction	PCR
Position-Specific Iterated Basic Local Alignment Search Tool	PSI-BLAST
Protein Data Bank	PDB
Protein Kinase C Binding Protein	PRKCBP1
Protein-Protein Interaction	PPI
Pygopus Family PHD Finger 1	PYGO1
Pygopus Family PHD Finger 2	PYGO2
Pyridine-2,4-Dicarboxylic Acid	2,4-PDCA
Random Access Memory	RAM
Rapid Elimination Of Swill	REOS
Really Interesting New Gene	RING
Recombination Activating Gene 1	RAG1
Recombination Activating Gene 2	RAG2
Recruits DNA Methyltransferase 3 Alpha	DNMT3A
Relative Humidity	RH
Ribosomal RNA	rRNA
RING Finger Protein 168	RNF168
RNA Interference	RNAi
Rubenstein-Taybi Syndrome	RTS
S-Adenosyl Methionine	SAM
Set Domain Bifurcated 1	SETDB1
Severe Combined Immunodeficiency	SCID
Simple Modular Architecture Research Tool	SMART
Spindlin 1	SPIN1

Squamous Cell Carcinoma	SSC
Structural Genomics Consortium	SGC
Structure Activity Relationship	SAR
Surface Plasmon Resonance	SPR
Target Discovery Institute	TDI
T-Cell Acute Lymphoblastic Leukaemia	T-ALL
Ten-Eleven Translocation 5-Methylcytosine Dioxygenase	TET
tert-Butyloxycarbonyl	Boc
Tobacco Etch Virus	TEV
Tripartite Motif Containing 33	TRIM33
Tris(2-Carboxyethyl)Phosphine	TCEP
Tumour Protein P53 Binding Protein 1	TP53BP1
Ubiquitin-Like With PHD And Ring Finger Domains 1	UHRF1
Whelan And Goldman	WAG
Wolf-Hirschhorn Syndrome Candidate 1	WHSC1
Wolf-Hirschhorn Syndrome Candidate 1-Like 1	WHSC1L1
X-Linked Mental Retardation	XLMR
Zinc Finger, MYND-Type Containing 11	ZMYND11

Chapter 1- Introduction

A wide variety of cell types within the human perform differing functions despite containing identical nuclear DNA. Therefore there must be a level of control above that of nuclear DNA which allows a wide range of specialised cell types to develop from a single zygote. These variations between cell types are a result of differing patterns of gene expression between cell types.[1]

The term “epigenetic” is used to describe the mechanisms which control how the same nuclear DNA sequence is interpreted differently in different cell types within the same organism. This term epigenetic derives from the Greek for "above-genetics", and refers to the storage of cellular information at a level above that of simple DNA sequence.[2] This epigenetic information can be stored in the form of DNA-methylation and the post-translational modification of histone proteins, which act as a spool which linear DNA is wrapped around.[3] These modifications control access to the underlying DNA and hence play a role in controlling gene expression, as well as affecting other cellular processes.[4] The breakdown of epigenetic signalling has been associated with a wide range of disease states, and therefore the discovery of molecules that can manipulate epigenetic processes is of growing interest as a means of therapeutic intervention.[5]

Chromatin

Each nucleus containing cell in the human body contains approximately 2 m of nuclear DNA across 46 chromosomes. In order to fit such a large amount of linear DNA inside the nucleus, DNA is wrapped around successive octomers of histone proteins, containing two each of the histone proteins, histone 2A (H2A), histone 2B (H2B), histone 3 (H3), and histone 4 (H4) (Figure 1). Each octomer of histone proteins has 146 base-pairs of DNA wrapped around it to form a body called a nucleosome. The successive nucleosomes along a sequence of DNA are often described as being analogous to beads on a string. This string of nucleosomes, along with

histone 1, forms a fibrous material called chromatin that in turn is organised to form chromosomes.[6]

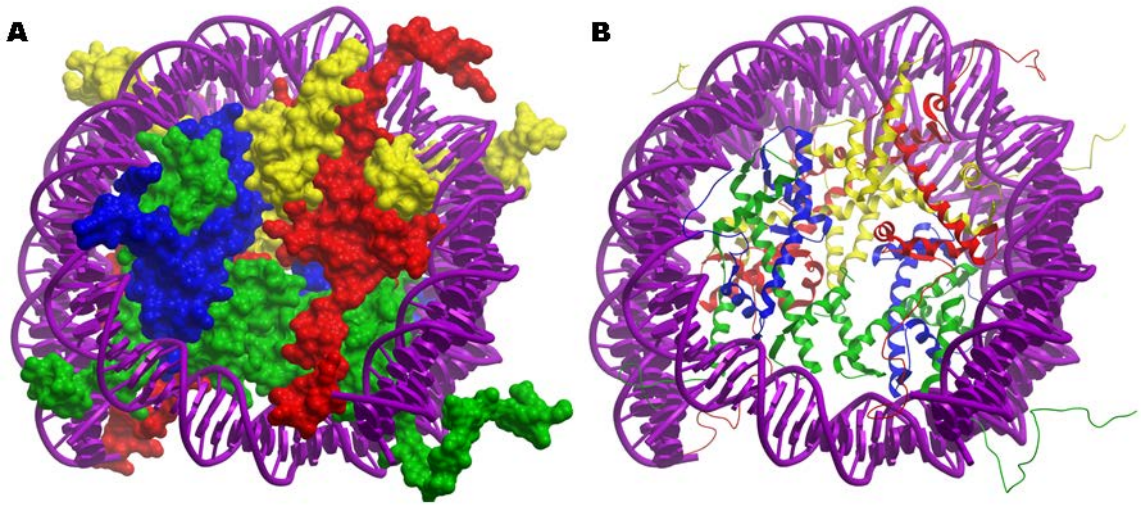


Figure 1. A. nucleosome contains two copies each of histone 2A (red), histone 2B (yellow), histone 3 (green), and histone 4 (blue) and a 146 base pair sequence of DNA (magenta) wrapped around the histone octomer. **A** Histone protein surfaces. **B.** Backbone ribbon of the histone proteins. The structure shown is a 2.5 Å x-ray crystal structure of a chicken nucleosome (*Gallus gallus*).[7] PDB ID: 1EQZ.

The structure of the nucleosome was elucidated in 1997[8] and shows that the negatively charged DNA is wrapped around the positively charged surface of the histone proteins. Importantly, the *N*-termini of both histone 3 and both histone 4 proteins are unstructured and protrude from the nucleosome. These *N*-termini are often referred to as "histone tails", and are the sites of some of the most well understood post-translational modifications of histones.

Chromatin Structure Can Control DNA Templated Functions

Chromatin can be densely packed into a form known as heterochromatin. This densely packed form of chromatin prevents access of transcriptional machinery to the DNA sequence within it, and is hence associated with genomically silent regions. Alternatively chromatin can form a loosely packed structure known as euchromatin which allows greater access to the underlying DNA and is hence associated with actively transcribed regions of the genome (Figure 2).[9] Factors

that control whether chromatin adopts a heterochromatin or euchromatin structure are discussed below.

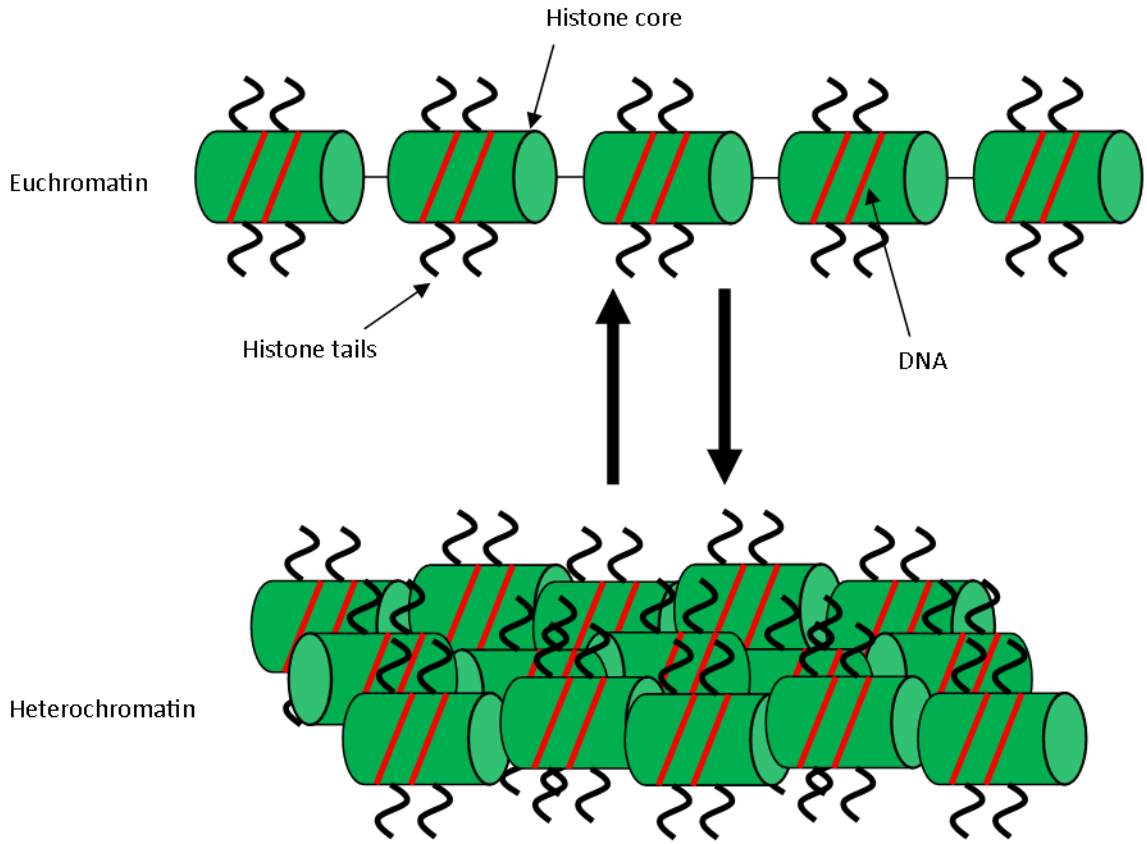


Figure 2. Euchromatin is loosely packed and therefore allows access to the underlying DNA. Heterochromatin is densely packed and is associated with transcriptionally silent parts of the genome. Green barrels represent the histone octomer at the core of the nucleosome. DNA is shown wrapped around the histone octomer, and histone tails are shown protruding from the nucleus.

Epigenetics

Chromatin can be covalently modified in a specific and reversible manner. These modifications occur both on the histone proteins and on DNA itself, and influence whether local regions of chromatin adopt an active euchromatin state or an inactive heterochromatin state.[4,10] These modifications can effect chromatin structure by recruiting cellular machinery which remodels chromatin in a process coupled to ATP hydrolysis,[9] or by physically changing the interaction between the DNA and histone proteins. For example, acetylation of a lysine converts it from

being a positively charged residue to a neutral residue. This will affect the interactions between the histone in question and negatively charged DNA by reducing the electrostatic attraction.

Covalent chromatin modifications act in a combinatorial manner, with DNA modifications influencing the addition or removal of nearby histone modifications, and vice-versa.[11–13] For example, the enzymatically inactive protein DNA methyltransferase 3-like (DNMT3L) binds to an *N*-terminal region of histone 3 in a manner dependant on the absence of methylation state of lysine 3. This in turn recruits DNA methyltransferase 3 alpha (DNMT3A) which triggers *de novo* DNA methylation.[11] This example is demonstrative of the cooperative nature of covalent modifications of histones and DNA.

DNA Methylation

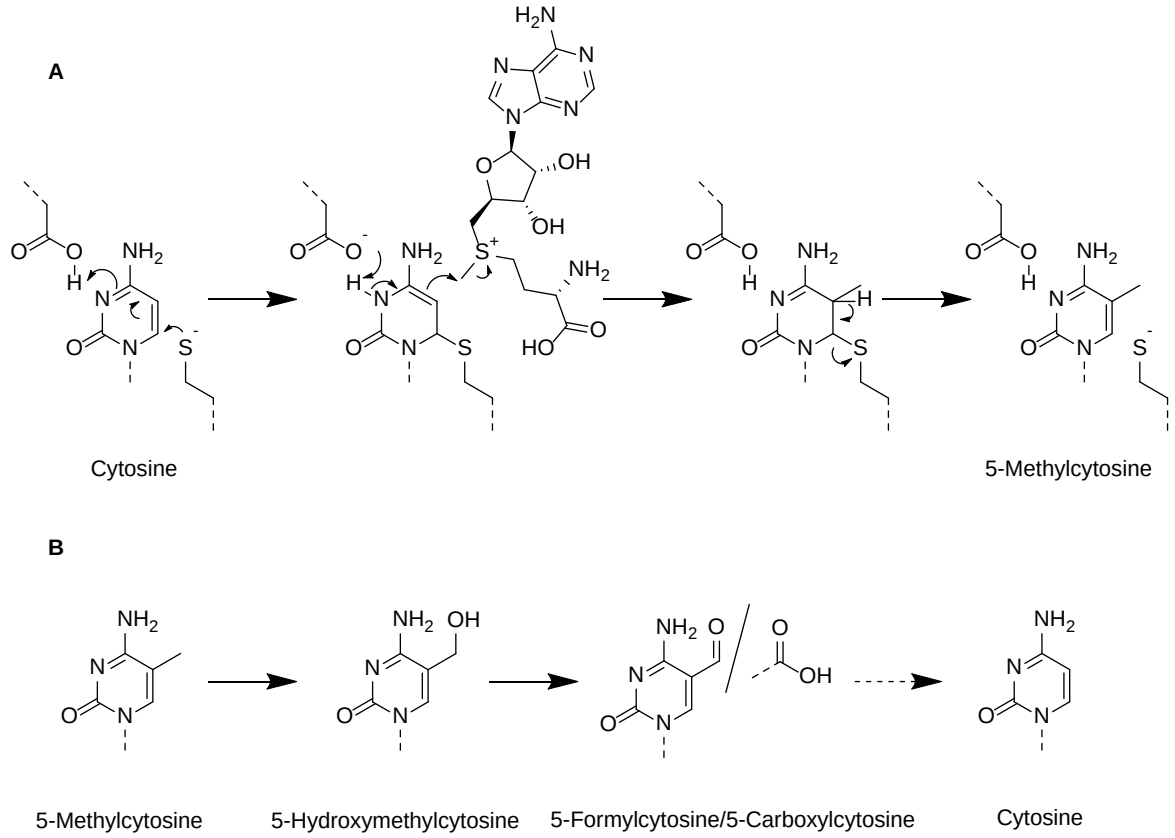


Figure 3. A. The mechanism by which a DNA methyltransferase (DNMT) enzyme transfers a methyl group from SAM to a cytosine residue. **B.** 5-Methylcytosine and oxidised variants. 5-Methylcytosine is associated with transcriptionally silent regions of chromatin, whereas the functional role of its oxidised forms are yet to be fully established.

One of the molecular mechanisms that controls whether chromatin is actively transcribed or not is DNA methylation. The 5-position of a cytosine residue followed immediately by a guanine residue (CpG motif) is the dinucleotide with the highest degree of methylation.[14,15] The resultant 5-methylcytosine (5-mC) is associated with silenced regions of the genome. DNA methylation is controlled via DNA methyltransferase enzymes, which transfer a methyl group from *S*-adenosyl methionine (SAM) to the 5 position of cytosine (Figure 3A). The modified base 5-methylcytosine can be converted to 5-hydroxymethylcytosine (5-hmC), and on to 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC) (Figure 3B). The ten-eleven translocation 5-methylcytosine dioxygenase (TET) family of enzymes are responsible for the oxidative steps in this pathway. It has been hypothesised that decarboxylases may exist that would convert 5-carboxylcytosine back to unmodified cytosine; however, this hypothesis is yet to be confirmed.[16]

Histone Tail Modifications

A wide range of histone tail modifications have been observed, including ubiquitination, SUMOylation, phosphorylation, and citrullination.[17] The most well studied examples of histone modification are acetylation and methylation. Acetylation is observed on lysine residues of histone 2A, histone 2B, histone 3, and histone 4;[18] methylation has been observed on lysine residues in histone 3 and histone 4,[19] and arginine residues in histone 2A, histone 3, and histone 4.[20]

Histone Acetylation

Lysine acetylation is associated with euchromatin and actively transcribed regions of the genome. This was originally attributed to physiochemical changes caused by changing from a charged lysine residue to a neutral acetyl lysine residue. This loss of charge would weaken the interaction between negatively charged DNA and the unmodified lysine side chains in histone, and cause chromatin relaxation. This would allow greater access to the underlying DNA and hence increased levels of gene expression. It is now known that in addition to a change in ionic

interactions between DNA and histones, acetyl lysine marks also recruit proteins containing acetyl-lysine read domains. The recruitment of an acetyl-lysine reader domain containing protein to an acetylated lysine can start a signalling cascade, leading to various cellular processes, including gene expression.

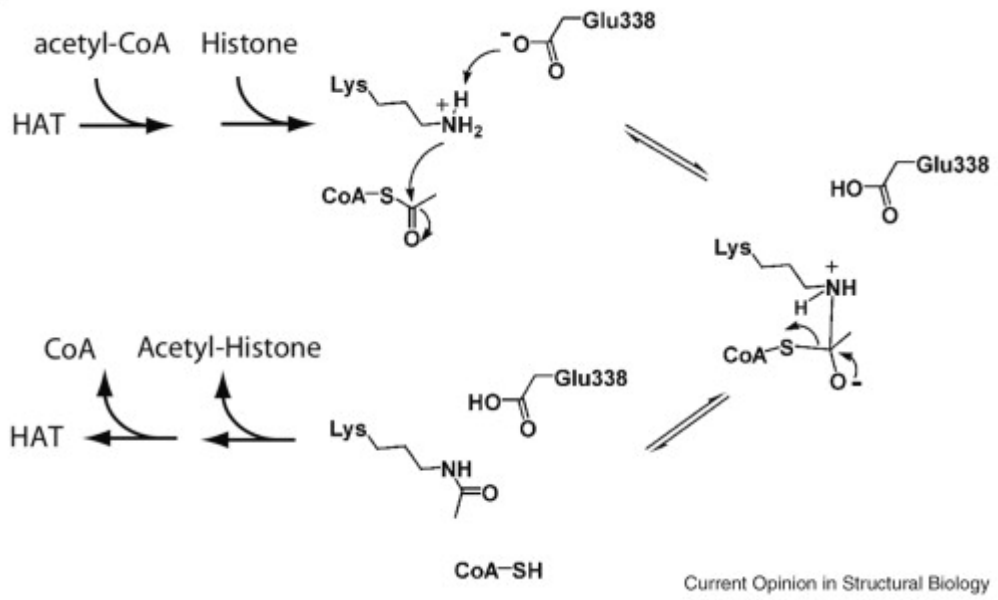


Figure 4. Proposed mechanism of acetyl transfer. The KAT contains a deprotonated acidic residue that acts as a general base catalyst. This figure is taken from *Catalysis and substrate selection by histone/protein lysine acetyltransferases.*[21]

A series of reader, writer and, eraser protein domains are associated with histone lysine acetylation. Lysine acetyl transferases (KATs) transfer an acetyl group from the acetyl donor acetyl co-enzyme A (Ac-CoA) to an unmodified lysine. This process is subject to general base catalysis from a nearby acidic residue (Figure 4). The human proteome contains 29 KATs, which are generally divided up into three sub families.[18,22] These are the Gcn5-related *N*-acetyltransferase (GNAT) family, the MOZ, Ybf2, Sas2, Tip60 (MYST) family, and the orphan family.

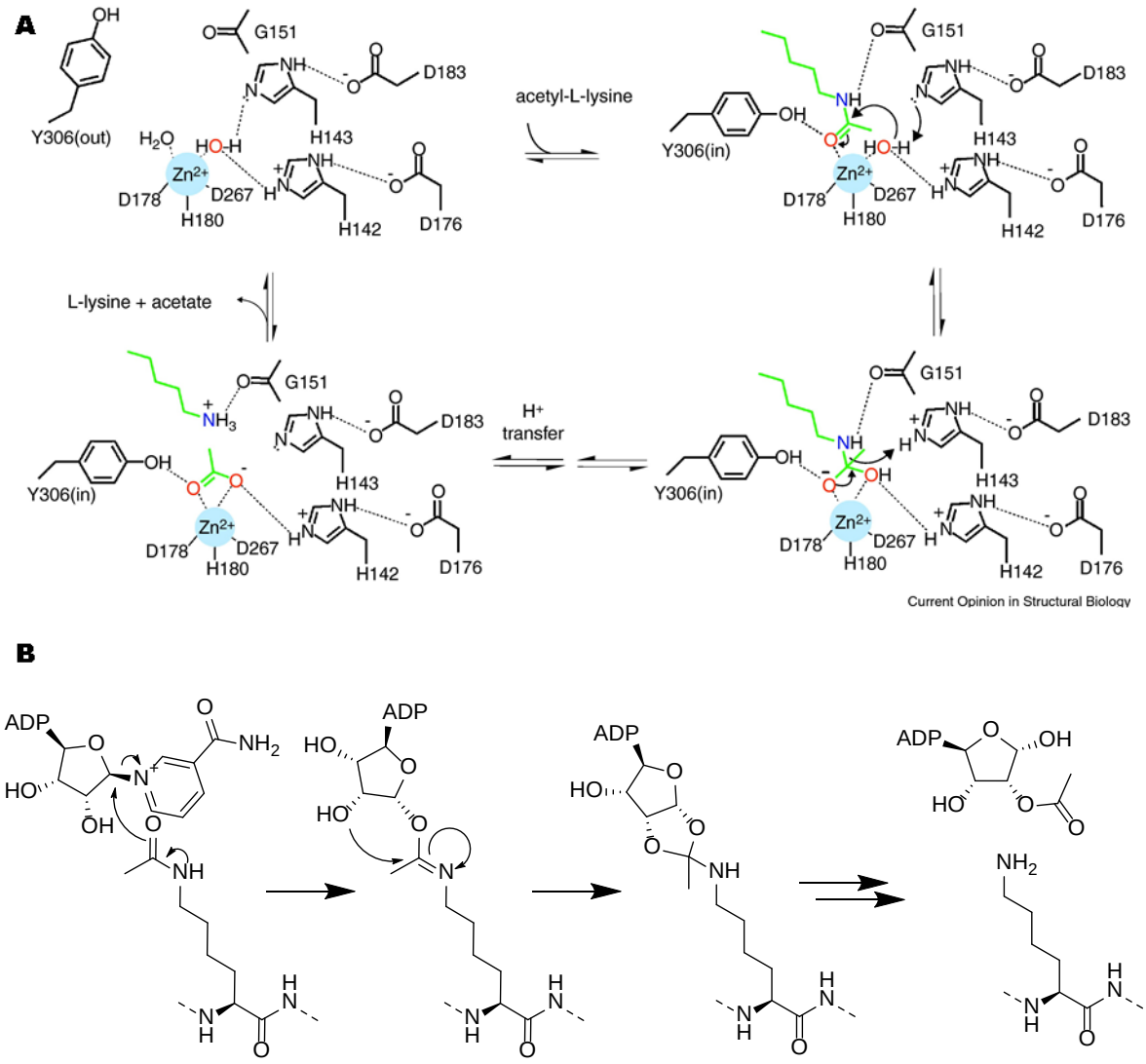


Figure 5. A. Mechanism of action of a Zn(II) dependent HDAC (class I, II, and IV). This figure is taken from *Structure, mechanism, and inhibition of histone deacetylases and related metalloenzymes.*[23] **B.** Mechanism of action of a NAD^+ dependent class III HDAC.

Histone deacetylases (HDACs) remove the acetyl group from acetyl lysine, restoring the acetyl lysine to an unmodified state. There are 18 HDAC enzymes in the human proteome, and are commonly grouped into four classes. Class I, class II, and class IV HDACs are zinc-containing metalloenzymes. The coordinated Zn(II) ion acts as a Lewis acid, and activates the amide to hydrolysis by water (Figure 5A). Class III HDACs (also known as sirtuins) use nicotinamide adenine dinucleotide (NAD^+) as a co-factor during deacetylation (Figure 5B). There has been much interest in the development of HDAC inhibitors as drugs. Two HDAC inhibitors, vorinostat

and romidepsin, are currently approved for the treatment of cutaneous T-cell lymphoma and in 2014 Belinostat was approved by the FDA for peripheral T-cell lymphoma.[24] Two further HDAC inhibitors have been approved in 2015, Panobinostat has been approved for use for multiple-myeloma,[25] and Chidamide has been approved in China only for pancreatic cancer.[26] These are the currently the only drugs to be approved that target a gene product involved in reading, writing, or erasing histone post-translational modifications.

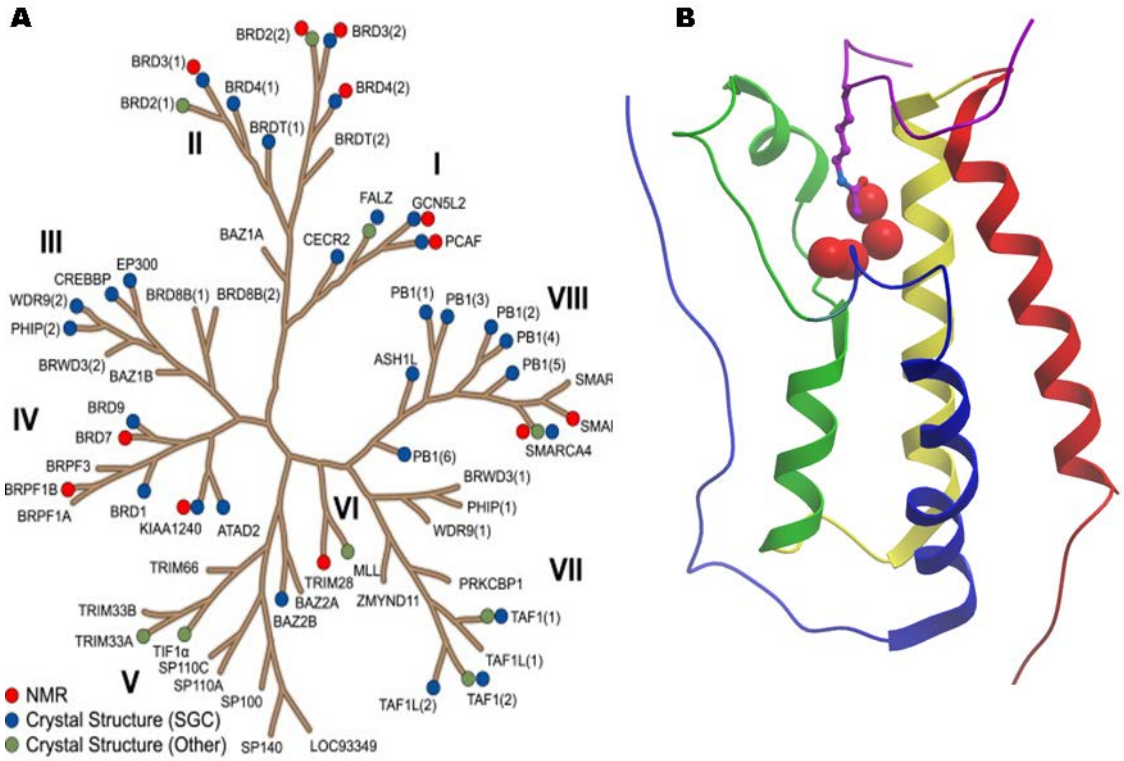


Figure 6. A. Bromodomains can be grouped into eight sub-families. All of these sub-families are well studied and many bromodomain structures have been solved. Figure taken from *Histone recognition and large-scale structural analysis of the human bromodomain family.*[27] **B.** All bromodomains have a well conserved fold consisting of four alpha helices. These helices are commonly known as Z (blue), A (green), B (yellow), and C (red). A bound H4 peptide (magenta) is shown with an acetylated (H4K5ac, magenta stick) shown in the bromodomains binding site. This site contains four water molecules (red spheres) that are found across the bromodomain family. Structure shown is the first bromodomain of BRD4[27] PDB ID: 3UVW.

The most well studied acetyl-lysine reader domain is the bromodomain. The human proteome has 61 bromodomains in 46 diverse proteins, some of which contain up to six individual bromodomains (Figure 6A). Bromodomains have a conserved fold containing four alpha-helices,

connected by varying loop regions (Figure 6B). These varying regions allow bromodomains to exhibit a wide variety of substrate specificity for many different potential acetyl lysine sites.[27]

Due to their links with pathogenic states, there has been intensive research into small-molecule bromodomain inhibitors (Figure 7). There is now a wide range of selective bromodomain inhibitors covering most branches of the bromodomain phylogenetic tree. Some of these inhibitors are currently in clinical trials,[28] and many have found use as chemical probes, helping to elucidate the role played by bromodomains in both healthy and diseased cells.[29]

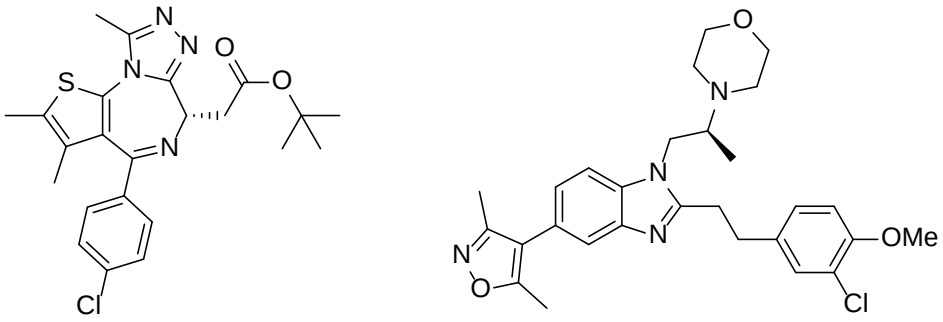


Figure 7. (+)-JQ1 is a competitive inhibitor of bromodomains belonging to the bromodomain and extra-terminal (BET) family.[30] SGC CP30 is a competitive inhibitor of the bromodomains of CREBBP and EP300.[31]

Histone Methylation

As well as acetylation, post-translational methylation of histones is used to control chromatin templated processes. Methylation can occur at lysines or arginines; lysine methylation can be either mono-, di-, or trimethylated, and arginine can be mono-methylated, symmetrically dimethylated, or asymmetrically dimethylated (Figure 8). Methylation is controlled by a series of demethylase and methyltransferases, and various methyl lysine reader domains translate methylation patterns into downstream effects.

Histone Modification	Effect on Gene Transcription
H3K4me1-3	Activating
H3K36me3	Activating
H3K79me3	Activating
H3K9me1	Activating
H3K27me1	Activating
H3K9me2-3	Silencing
H3K27me2-3	Silencing

Table 1. Summary of the effects of histone 3 lysine methylation of gene transcription.

Histone methylation can be either activating or deactivating depending on the exact residue and level of methylation. Generally mono-, di-, and tri-methylation at H3K4, tri-methylation of H3K36, and tri-methylation of H3K79 are considered to be activating.[32] Monomethylation of H3K9 and H3K27 have been suggested to be associated with active genes,[32] whereas di- and tri-methylation at these positions are associated with silenced genes (Table 1).

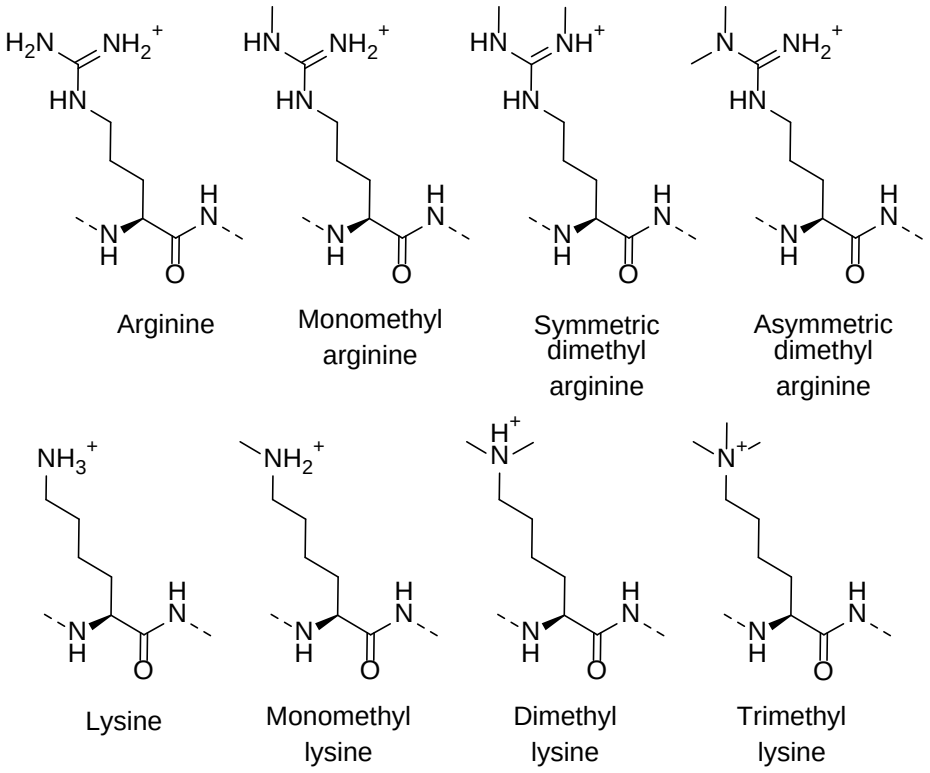


Figure 8. Arginine and lysine are found in multiple methylated states in histones.

Histone methylation is mediated by histone methyltransferases (HMTs) which transfer a methyl group from *S*-adenosyl methionine (SAM) to either lysine or arginine. There are three classes of HMT: SET-domain containing lysine methyltransferases, non-SET lysine methyltransferases, and arginine methyltransferases (PRMT). Of lysine methyltransferases in humans, the only known non-SET lysine methyltransferase is DOT1L, which methylates H3K79.[33] The mechanism of histone lysine methylation is contested, with various possible models for lysine deprotonation. Kipp et al propose a mechanism where lysine is deprotonated by bulk solvent while bound to the enzyme (Figure 9).[34] This occurs via a channel which provides bulk water with access to the active site. Other studies have proposed models where a tyrosine in the active site acts as a base,[35] or where lysine is deprotonated by bulk solvent prior to binding the active site.[36]

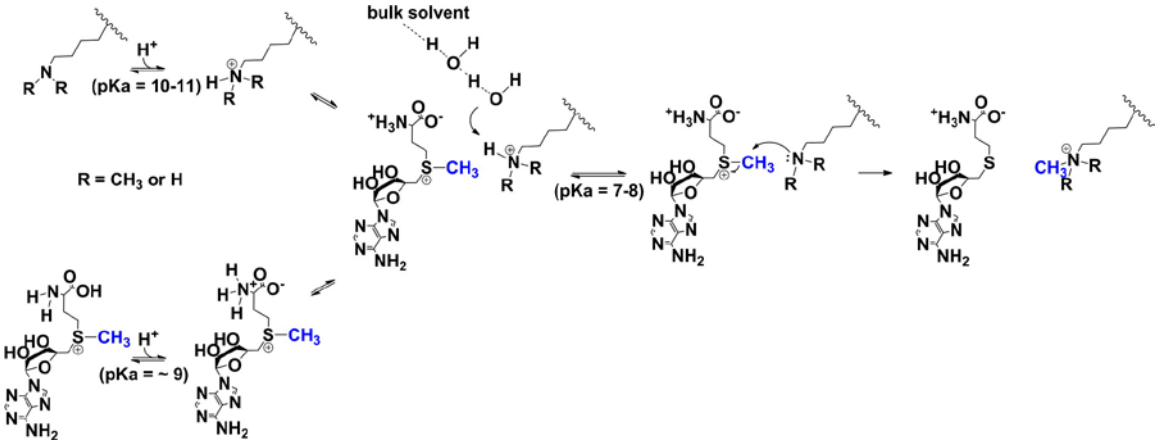


Figure 9. A proposed model for the mechanism of histone lysine methylation. Lysine is deprotonated by bulk solvent prior to reacting with methyl donor S-adenosyl methionine (SAM). Figure taken from *Enzyme-dependent lysine deprotonation in EZH2 catalysis*.[34].

There are two major classes of lysine demethylases, lysine specific demethylases (LSD), and JmjC containing lysine demethylases (KDM2-7). LSDs were first discovered in 2004,[37] and can remove a methyl group from either monomethyl or dimethyl lysine. LSDs are flavin adenine dinucleotide (FAD) dependent enzymes and work by removing two hydrogens from the methylated lysine to give $FADH_2$ and an iminium ion. This iminium is then hydrolysed to give formaldehyde and the lysine in a lower methylation state (Figure 10A). LSDs cannot demethylate tri-methylated lysine, as it lacks a nitrogen bound hydrogen; this means the oxidative loss of hydrogen is not possible.

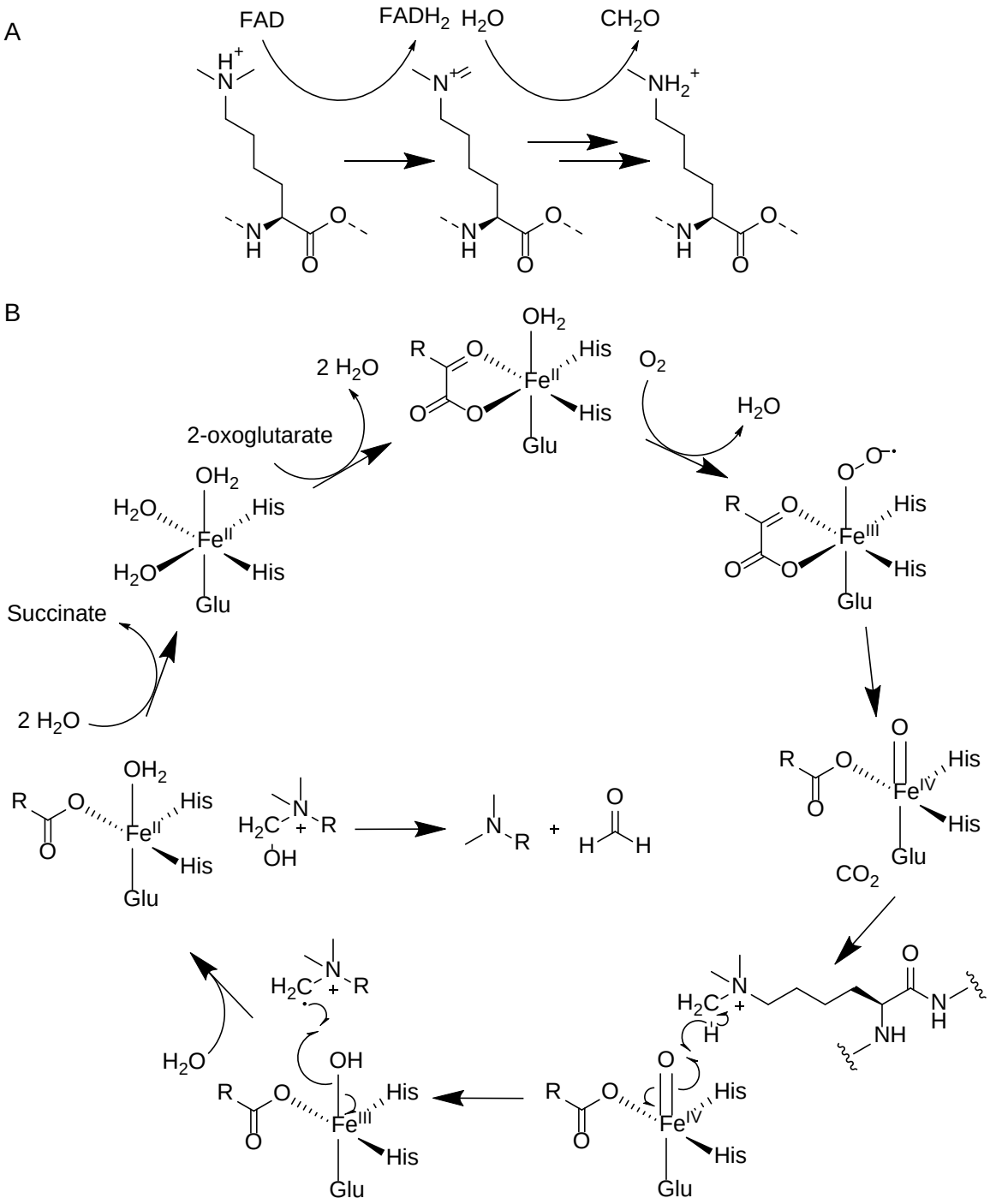


Figure 10. There are two enzyme families that demethylate lysines. These two enzyme families use distinctive mechanisms of demethylation. **A.** Lysine specific demethylases (LSDs) are flavin adenine dinucleotide (FAD) dependent enzymes. LSDs are only capable of demethylating mono- and dimethyl lysine. **B.** Lysine demethylases (KDM2-7) are 2-oxoglutarate dependent metalloenzymes. They are capable of removing a methyl group from mono-, di, or tri-methyl lysine. Figure taken from *Investigations on the oxygen dependence of a 2-oxoglutarate histone demethylase*[38]

Tri-methylated lysine as well as lower methylation states can be demethylated by KDMs (KDM2-7), which were first discovered in 2007.[39] These are 2-oxoglutarate dependent Fe(II) containing enzymes, which catalyse the hydroxylation of the methyl group. This creates a hemi-aminal that is hydrolysed to give formaldehyde and a lysine with one less methyl group (Figure 10B). There is also evidence that JmjC enzymes act as histone arginine demethylases.[40]

Methyl Lysine Reader Domains

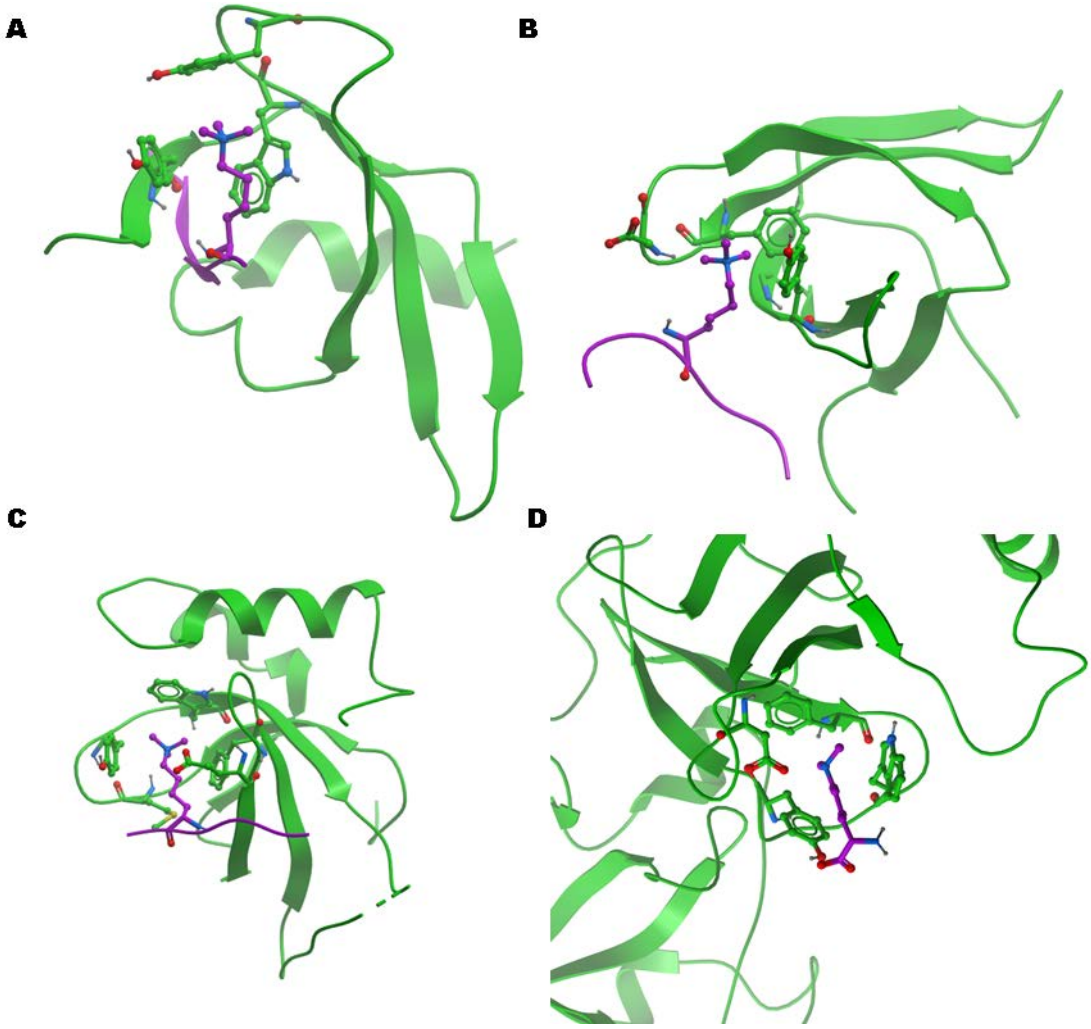


Figure 11. The royal family of methyl lysine reader domains use aromatic-cages to recognise methylated lysine via cation-pi and hydrophobic interactions. Protein structures are shown in green with methylated lysines shown in magenta. **A.** The chromodomain of Chromodomain protein 1 (Chp1) from *Schizosaccharomyces pombe* in complex with K3K9me3. PDB ID: 3G7L. **B.** The Tudor domain of PHD finger protein 1 (PHF1) in complex with H3K36me. PDB ID: 2M0O. **C.** The PWWP domain of hepatoma-derived growth factor-related protein 2 (HDGFRP2) in complex with H3K79me3. PDB ID: 3QJ6. **D.** The MBT of lethal (3) malignant brain tumour-like 1 (L3MBTL1) in complex with dimethyl lysine. PDB ID: 2RHX.

There are many domain families that read methyl lysines.[41,42] They include Tudor domains, chromodomains, PWWP domains, and malignant brain tumour (MBT) domains, which are collectively known as the royal family and all contain a beta-barrel fold. They bind to methyl lysine using several aromatic residues that form an 'aromatic cage' which makes cation-pi interactions with methyl lysine (Figure 11). The other major domain family capable of methyl lysine recognition are the plant homeodomains (PHDs). The prevalence of these methyl lysine readers within the human genome is summarised in Table 2.

Methyl Lysine Reader Domain	Prevalence in Human Genome
Tudor domain	71
Chromodomain	43
PWWP domain	28
MBT domain	29
PHD	173

Table 2. A summary of methyl lysine reader domains and their prevalence within the human genome.

Plant Homeodomains

Plant homeodomain are protein domains which bind to histone tails in a lysine-methylation state dependant manner. There are 173 PHDs in the human proteome in 103 proteins (Figure 12). PHDs typically bind histone 3 at the *N*-terminus in a manner that is dependent on the methylation state of histone 3 lysine 4 (H3K4). PHDs that are specific for methylated H3K4 and unmodified H3K4 are known.

PHDs typically contain 50-70 amino acids and have eight conserved zinc binding residues that bind two Zn(II) ions that stabilise the domain. These zinc binding domains are typically found in a Cys4-His-Cys3 pattern, with zinc binding residues 1-2 and 5-6 binding one Zn(II) ion, and zinc binding residues 3-4 and 5-6 binding the other Zn(II) ion (Figure 13A and B). This zinc binding mode is sometimes described as "cross-braced", and is also found in the related Really Interesting New Gene (RING) domain.[43]

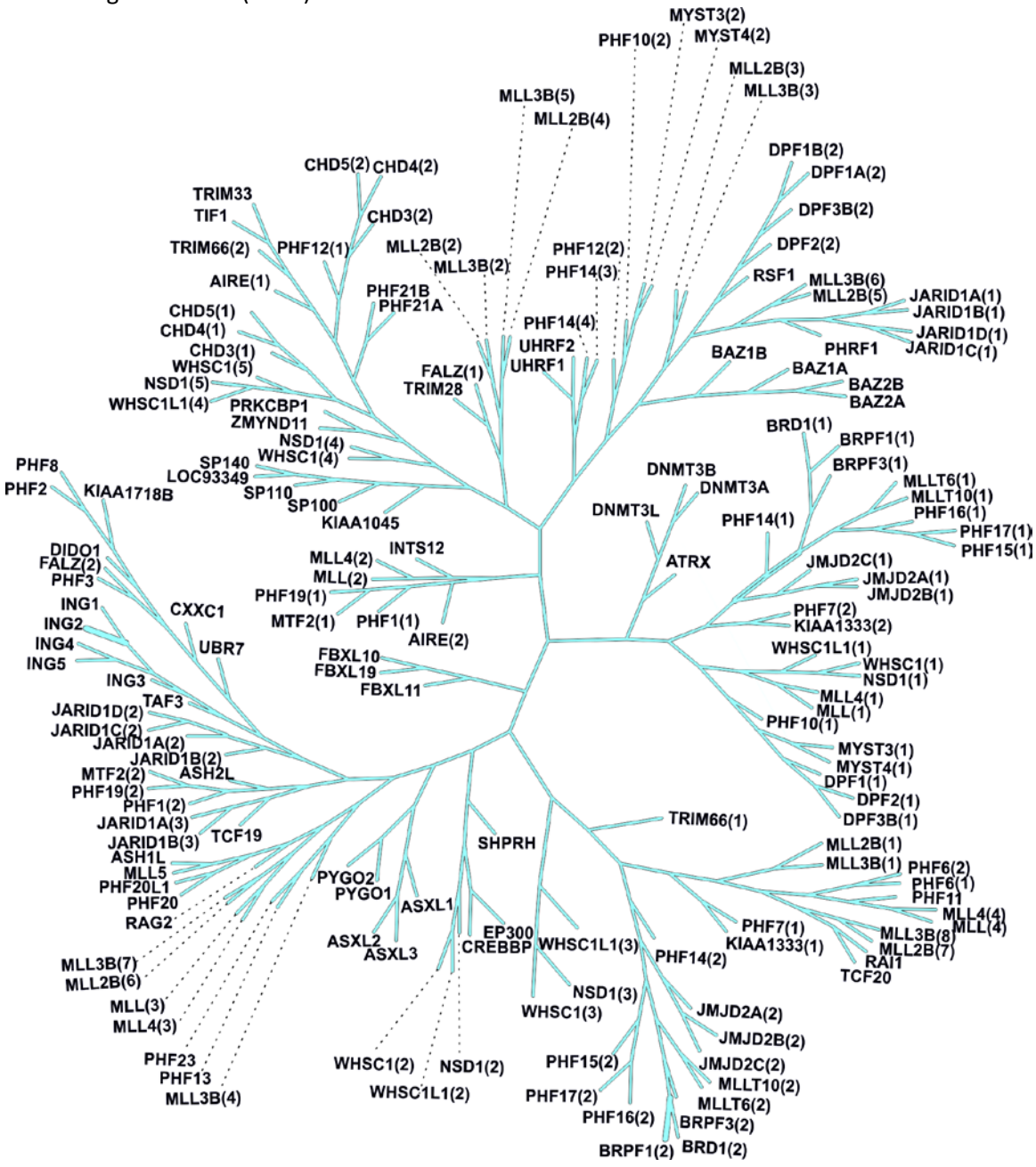


Figure 12. Phylogenetic tree of PHDs based on sequence. There are 173 PHDs in the human proteome. They have a well conserved fold but exhibit great sequence diversity. This tree is described in detail in Chapter 2.

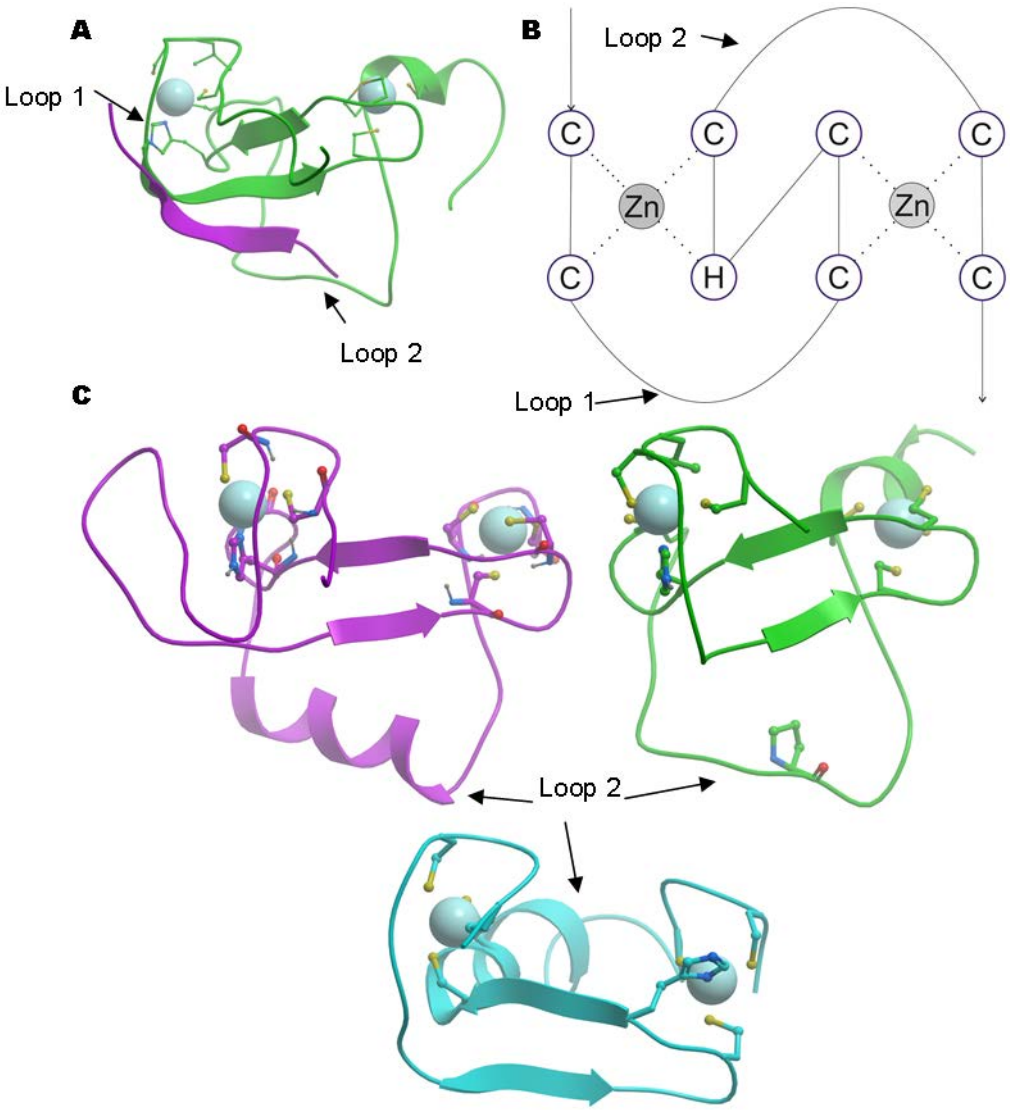


Figure 13. PHDs have a well conserved fold and bind two Zn(II) ions. **A.** The first PHD of autoimmune regulator protein (AIRE) has a canonical PHD fold (green) and engages histone tails (magenta). PDB ID: 1XWH. **B.** PHDs have a cross braced zinc binding motif, with the zinc binding residues typically in a Cys4-His-Cys3 pattern. The sequence of Loop 1 and Loop 2 can vary greatly between PHDs. **C.** Based on analysis of available structures; approximately 40 % of PHDs contain an alpha-helix in Loop 2. The Loop 2 region is also the area that distinguishes a PHD from a RING domain. The first PHD of Double PHD Finger Protein 3 B (DPF3B) contains an alpha helix in loop 2 (magenta). PDB ID: 2KWJ. The PHD of autoimmune regulator protein (AIRE) contains a proline in Loop 2 (green). This is found in most PHDs without a loop 2 alpha helix and is likely to prevent alpha-helix formation. PDB ID: 1XWH. RING domains are E3 ubiquitin ligases and share structural similarities in the central beta-sheet and zinc binding regions with PHDs; however, they differ in the Loop 2 region (c). The RING domain shown is the RING domain of RING finger protein 168 (RNF168). PDB ID: 4GB0.

Between the two zinc binding sites, PHDs contain a well conserved 2-strand anti-parallel beta-sheet. In many instances this is the only secondary structure motif found in a PHD; however, some PHDs contain an alpha-helix in an extended loop section between zinc binding residue 6 and zinc binding residue 7 (Figure 13C). It is the sequence and structure of this region that distinguishes PHDs from RING domains, as RING domains contain a short alpha-helix immediately after the 6th zinc binding residue that acts as an E3 ubiquitin ligase (Figure 13).

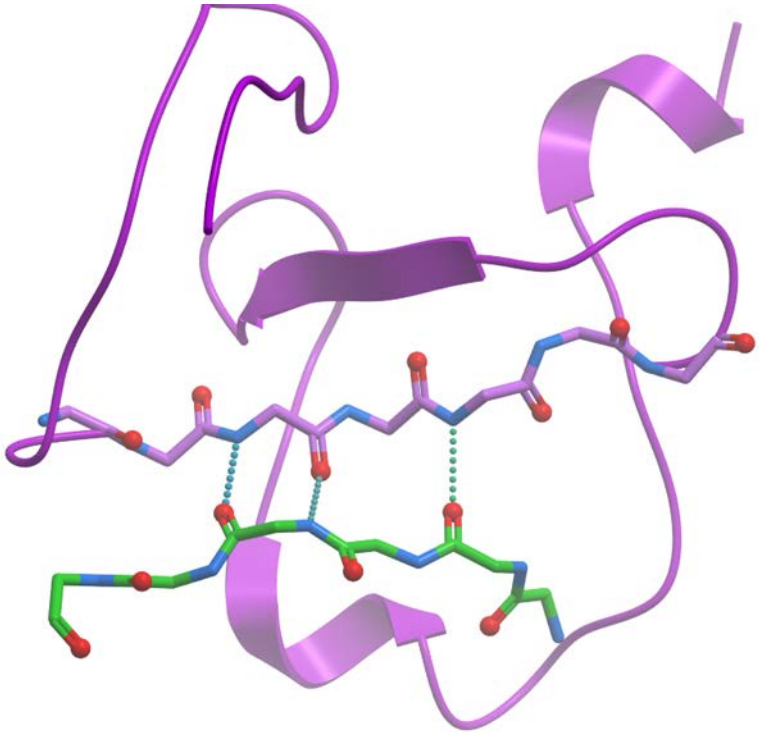


Figure 14. Histone 3 forms an extended beta-sheet on binding to a PHD, making several hydrogen bonding interactions via the peptide backbone. The PHD shown is the PHD of bromodomain and PHD transcription factor (BPTF) PDB ID: 2FSA.

PHDs engage with resides 1-4 of histone 3 via the formation of an extended anti-parallel beta-sheet (Figure 14). Residues 1-4 of histone 3 form a beta-strand that makes backbone contacts with the anti-parallel beta-sheet found at the core of the PHD. This beta-sheet formation places the *N*-terminus of histone 3 near a backbone carbonyl of the PHD allowing hydrogen bond formation (Figure 15A). This engagement of the *N*-terminus is important for controlling substrate specificity, as it directs H3K4 towards the PHDs lysine binding site. Other features of

the PHD domains that ensure substrate specificity are the hydrophobic pocket that accommodates the side chain of H3A1, and the methyl group of H3T3 (Figure 15B), and an acidic region that interacts with H3R2 (Figure 15C). These interactions ensure that H3K4 is directed to the lysine binding site of the PHD domain.

The composition of this lysine binding site determines whether the PHD binds to methylated H3K4 or unmethylated H3K4. PHDs that bind methylated H3K4 have between two and four aromatic residues that form an aromatic cage that engages methylated H3K4 via cation-pi interactions (Figure 15D). PHDs that bind to unmodified H3K4 do not contain an aromatic cage; they contain acidic residues near the H3K4 binding site that form salt bridges with H3K4 (Figure 15E). Some PHDs will preferentially bind histone 3 containing methylated H3K9, this occurs through an orthogonal binding mode to methylated H3K4 recognition (Figure 15F).[44–46] Residues 1-4 are bound in a conventional manner, with H3K9 positioned near to an aromatic residue on the surface on the domain. Therefore H3K4 and H3K9 methylation state recognition is orthogonal, and it is possible for a PHD to selectively bind histone 3 unmethylated at H3K4 and methylated at H3K9.[44]

PHDs and Disease

PHDs have been implicated in numerous disease states (Table 3). A PHD can contribute to a disease state by one of the two mechanisms. The first is through a loss of function, such as that caused by a missense mutation which inactivates the PHD or prevents it from folding correctly. This mechanism been implicated in numerous disease states. The second is through inappropriate activity of a PHD. This can be caused by over-expression of a PHD containing protein, or via a fusion protein caused by a chromosomal translocation. Examples of PHDs which are implicated in disease states are described below.

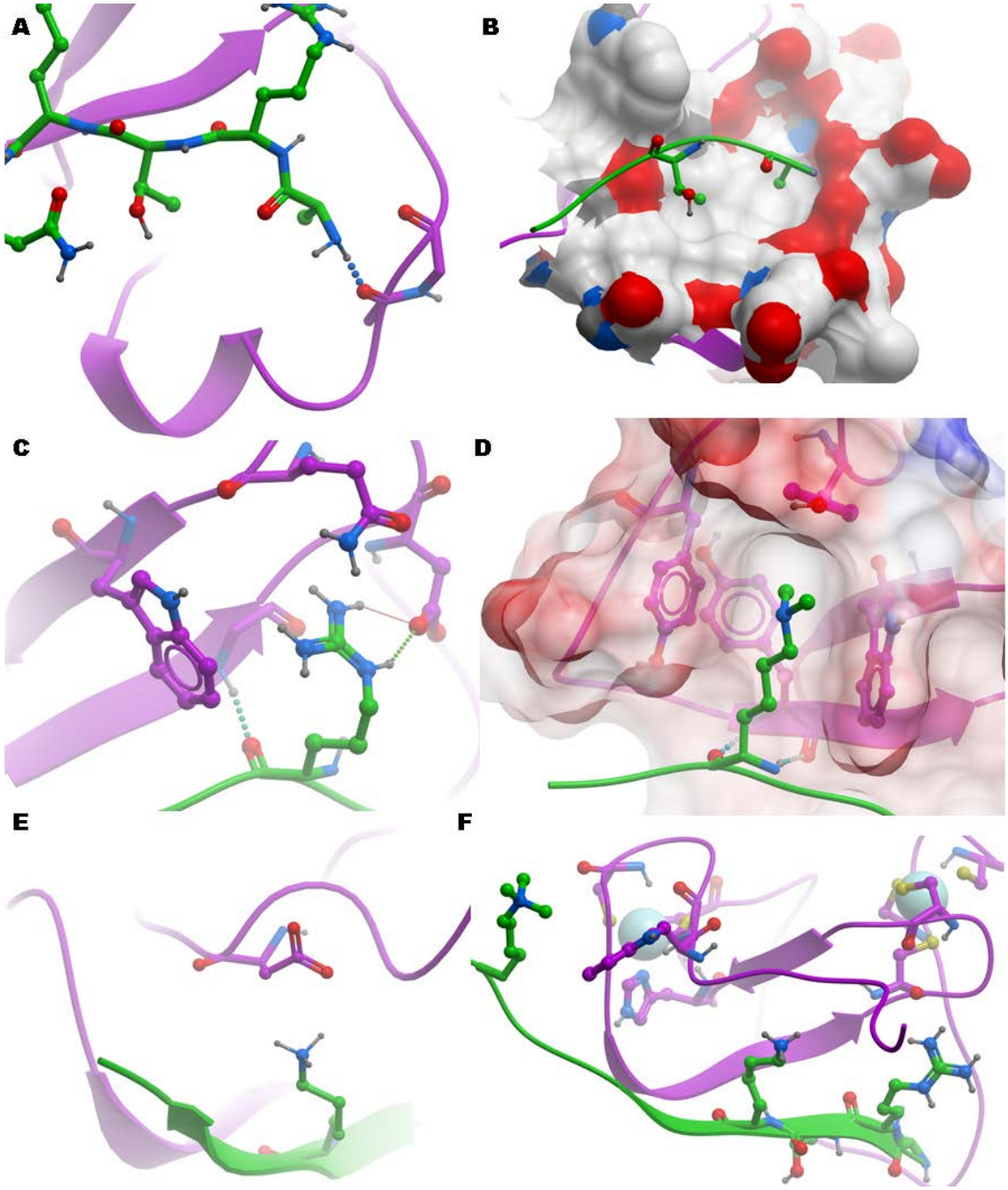


Figure 15. A. The *N*-terminus of histone 3 forms a hydrogen bond with a backbone carbonyl of the PHD. **B.** The methyl groups of H3A1 and H3T3 are accommodated in a hydrophobic cleft. **C.** H3R2 is bound in an acidic patch separated from the H3K4 binding channel by a tryptophan residue. **D.** PHDs which bind to methylated lysine have an aromatic cage that engages methylated lysine via cation-pi interactions. **E.** PHDs which bind non-methylated lysine have acidic residues that engage lysine via Coulombic forces. **F.** Tri-methylated histone 3 lysine 9 (H3K9me3) forms a cation-pi interaction with a tryptophan found on the surface of the PHD. PHDs which interact with H3K9me3 do so via an orthogonal mode to H3K4me3 recognition. **A-D** shows the PHD of bromodomain and PHD transcription factor (BPTF). PDB ID: 2FSA and **E.** shows the PHD of autoimmune regulator protein (AIRE). PDB ID: 1XWH. **F.** Shows the PHD of tripartite motif containing 33 (TRIM33). PDB ID: 3U5N.

Protein	Disease association	PHD finger misregulation	Ligand for PHD
Immune disorders			
RAG2	T-B-NK+ Severe Combined Immunodeficiency (SCID) and Omenn Syndrome	Germline mutations	H3K4me3
AIRE	Autoimmune polyendocrinopathy Candidiasis ectodermal dystrophy (APECED), also known as autoimmune polyglandular syndrome type 1 (APS-1)	Germline mutations (both PHDs)	H3K4me0 (PHD1) Unknown (PHD2)
Cancer			
ING1	Breast cancer, melanoma, esophageal squamous cell carcinoma (SSC), head and neck SSC	Somatic mutation	H3K4me3,2
JARID1A (KDM5A)	Myeloid leukaemia	Translocation of the third PHD	Unknown (PHD3)
PHF23	Myeloid leukaemia	Translocation	Unknown
NSD1, NSD2, NSD3	Myeloid leukaemia	Translocation of PHD fingers	Unknown
MLL	T-cell lymphoblastic leukaemia	Internal deletion of first PHD	Unknown
PHF1	Endometrial stromal sarcoma	Translocation	Unknown
Neurological disorders			
NSD1	Childhood overgrowth syndromes such as Sotos syndrome and Weaver syndrome	Germline mutations in all 5 PHDs; truncation; micro-deletion	Unknown
ATRX	Various X-linked mental retardation disorders, including Alpha-Thalassemia and Mental Retardation, X-linked (ATRX) Syndrome	Germline mutations (>26 distinct PHD mutations reported)	Unknown
CBP	Rubenstein-Taybi Syndrome (RTS)	Germline mutation	Unknown
PHF6	Borjeson-Forssman-Lehmann Syndrome (BFLS)	Germline mutations (PHD1 only)	Unknown

Table 3. Mutations in genes encoding PHD finger proteins are associated with a wide variety of human diseases. This table is taken from *PHD fingers in human diseases: disorders arising from misinterpreting epigenetic marks*.[47]

RAG2

Recombination Activating Gene 2 (RAG2) is a PHD containing protein involved in V(D)J recombination.[48] The PHD of RAG2 recognizes histone 3 methylated at both lysine-4 and arginine-2. RAG2 acts in combination with Recombination Activating Gene 1 (RAG1) to create double strand breaks required for V(D)J recombination. This process is vital for adaptive immune response, and deletion of RAG2 in mice leads to a disruption of V(D)J recombination and hence a compromised immune system.

Point mutations in human RAG2 are also associated with immunocompromised phenotypes. Specifically, the conditions Omenn Syndrome and Severe Combined Immunodeficiency (SCID) are associated with RAG2 mutations.[49] There are 24 known mutations in RAG2 linked to either Omenn Syndrome or SCID, six of which are missense mutations located in the PHD. Two of these mutations are in zinc binding residues (C478Y and H481P), and these mutations are likely to cause disruption to the PHD fold. Another mutation, W453R, is found in a residue that forms part of the aromatic cage. It has been shown that this mutation abrogates binding to peptides containing H3K4me3. It is interesting to note that the W453R mutation, which causes loss of function but is unlikely to affect the structure of the PHD of RAG2, is associated with less severe forms of Omenn Syndrome. Whereas the C478Y and H481P mutations which are likely to cause a complete loss of structure are associated with more severe forms of Omenn Syndrome.[50]

ING1

Inhibitor of growth 1 (ING1) is a tumour suppressor gene which contains a C-terminal PHD known to bind tri-methylated lysine 4 in histone 3.[51] A C253 stop mutation which results in a truncated PHD, and a C215S mutation that disrupts zinc coordination[47] have been linked to skin cancer.[52]

Fusion to NUP98 in Blood Cancers

There are examples of PHD containing proteins being fused to the *N*-terminal FG-repeat domain of nucleoporin 98 (NUP98).[53] NUP98 is a nuclear pore complex component, which is known to form many fusion proteins associated with hematopoietic diseases such as acute myeloid leukaemia (AML). The FG-repeats of NUP98 recruit the HATs p300 and CBP.[54] Fusion of these FG-repeats to another chromatin binding domain such as a PHD could feasibly lead to misregulation of HAT activity.

Two examples of PHD containing NUP98 fusion proteins are NUP98-JARID1A[55] and NUP98-PHF23.[56] Although JARDI1A and PHF23 have different functions (JARID1A is a histone demethylase whereas PHF23 has no known enzymatic function), their C-terminal sections that form fusion proteins with NUP98 are similar. Both contain a nuclear localisation signal and a PHD domain that is likely to bind to H3K4me3.

Another class of PHD containing NUP98 fusion proteins involve nuclear receptor binding SET domain protein 1 (NSD1) and its close relative NSD3.[57] Both NSD1 and NSD3 contain five PHDs, all of which are incorporated into the malignant NUP98 fusion protein (Figure 16).

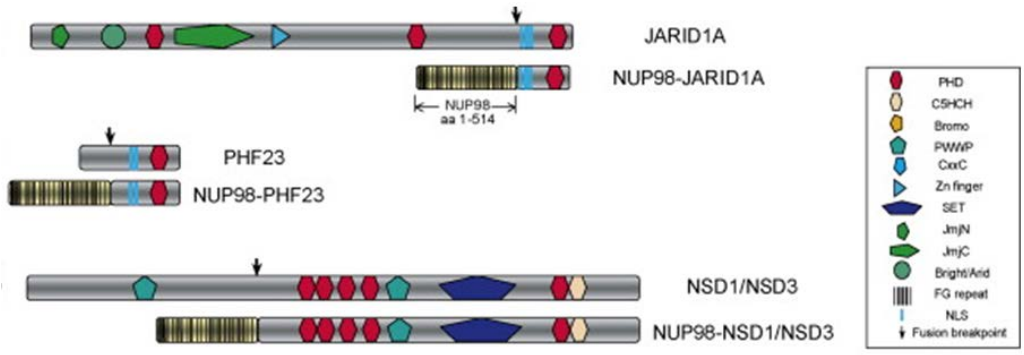


Figure 16. PHD containing fusion proteins of NUP98 associated with human disease. Figure adapted from *PHD fingers in human diseases: disorders arising from misinterpreting epigenetic marks.*[47] NUP98 forms PHD containing fusion proteins with JARID1A, PHF23, NSD1, and NSD3.

NSD1

Missense mutations in the PHDs, SET, and PWWP of NSD1 are associated with Sotos syndrome, a disease which causes excessive growth in early childhood. Many of these mutations are found in zinc binding residues of the one of the five PHDs of NSD1.[58] These mutations are likely to disrupt the PHD fold and any histone binding activity (Figure 17).

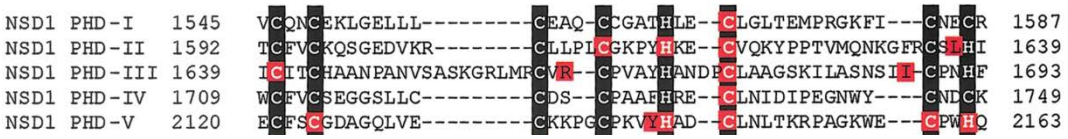


Figure 17. Mutations in PHDs of NSD1 are linked to Sotos syndrome. Figure adapted from *Genotype-Phenotype Associations in Sotos Syndrome: An Analysis of 266 Individuals with NSD1 Aberrations*[58]. Disease associated mutations – highlighted in red - in NSD1 are concentrated in zinc binding residues of PHDs (black).

ATRX and Mental Retardation Syndrome

Alpha thalassemia/mental retardation syndrome, X-linked (ATRX) is a PHD containing chromatin remodelling protein. Mutations in ATRX are associated with ATRX syndrome, a developmental disease characterised by distinctive facial features, such as widely spaced eyes and short nose. The disease is also associated with severe developmental delays and intellectual disability. Twenty-six disease causing mutations have been identified in the PHD of ATRX.[59]

PHF6

Borjeson–Forssman–Lehmann syndrome is a recessive, X-linked disease characterised by feeding problems in infancy, mild to moderate intellectual disability, and truncal obesity.[60] The disease is caused by mutations in PHD finger protein 6 (PHF6). One disease causing missense mutation, C99F, affects the fourth zinc binding residue of the first PHD domain of PHF6 with the likely effect that the mutation will be destructive to the fold of the PHD domain.

Inactivating and deleterious mutations of PHF6 have been linked with T-cell acute lymphoblastic leukaemia (T-ALL). Several of the disease causing mutations in PHF6 are missense mutations

found in the second PHD. These include the C280Y and C283R mutations found in first and second zinc binding residues, these mutations are likely to be destructive to the fold of the second PHD of PHF6, (Figure 18).[61]

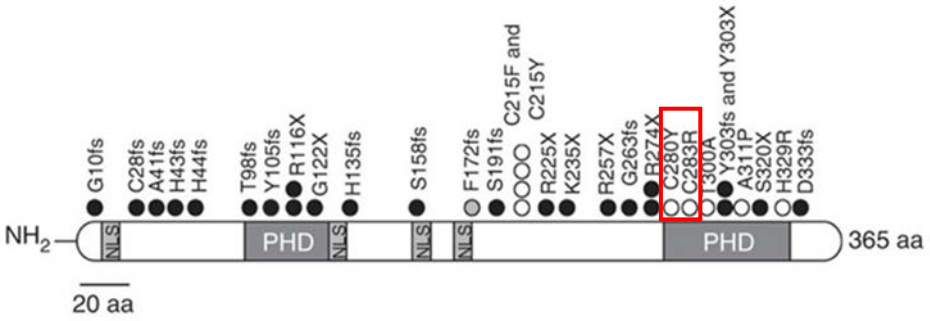


Figure 18. Mutations of PHF6 associated with T-cell acute lymphoblastic leukaemia (T-ALL). Figure adapted from *PHF6 mutations in T-cell acute lymphoblastic leukemia.*[61] Filled circles represent nonsense or frameshift mutations, unfilled circles represent missense mutations. The first and second zinc binding residues of the second PHD of PHF6 (highlighted) are subject to missense mutations which are likely to be disruptive to the PHD fold.

PHF11

PHF11 plays a role in class switch recombination (CSR) to IgE in B cells. Due to the role that IgE plays in allergic responses, it has been shown that excess expression of PHF11 lead to increased allergic reactions *in vivo* using a murine model. The same study also shows that siRNA silencing of PHF11 inhibits CSR to IgE in B cells suggesting that targeting PHF11 may lead to novel treatments for allergic diseases.[62]

CHD5

Chromodomain helicase DNA binding protein 5 (CHD5) is a tumour supressing protein that binds to unmodified H3K4 through its two PHDs.[63] This PHD-mediated binding has been shown to be essential to CHD5 ability to inhibit proliferation of mouse embryonic fibroblasts (MEFs). Mutations in the first PHD (G355A, D361A) and in the second PHD (D415A, C432W, D434A) of CHD5 have been shown to abrogate the ability of CHD5 to bind unmodified H3. CHD5 containing

these mutations shows a reduced ability to inhibit proliferation in MEFs compared to wild-type CHD5. This implies a crucial role for the PHDs of CHD5 in tumour suppression.[63]

Chemical Probes

Chemical probes are cell permeable molecules that inhibit a proteins function with a high degree of specificity.[64–66] Chemical probes can be used to aid pre-clinical target validation, by providing evidence that specific inhibition of a given target does produce a biological effect.[67] It has been suggested that improving target selection is a key approach to reducing phase II attrition rates.[68] Chemical probes provide a complimentary method to RNA interference (RNAi), which can be used to decrease the translation of a given protein by binding to a specific sequence of messenger RNA (mRNA) and targeting it for degradation. This prevents or reduces the translation of a given protein, and hence inhibits all its functions. A chemical probe based approach can be provide orthogonal information as often only a small subset of a proteins functions are inhibited. This could include inhibiting a single domain of a multi-domain protein,[31] or the possibility that a probe inhibits a proteins enzymatic activity without effecting roles relating to its structure.

The process of developing a chemical probe shares similarities with the process of developing a drug. Both involve the selection of an appropriate lead molecule and optimising its potency and selectivity; however, a chemical probe differs from a drug in both their purpose and their desired properties. A molecule designed as a chemical probe may have a higher degree of specificity than one designed as a drug.[69] A drug may achieve its desired effect by inhibiting multiple targets,[70] whereas a chemical probe is intended to study the biological effects of inhibiting a single or small number of targets. Therefore any off-target activity will prevent clear understanding of the biological role of the intended target. Although a chemical probe may have a higher degree of selectivity than a drug and be cell permeable, it is not required to be

optimised for its *in vivo* ADMET (Adsorption, Distribution, Metabolism, Excretion, and Toxicity) properties.

The assays and inhibitors discussed in Chapter 4 and Chapter 5 are intended to act as starting points for the development of chemical probes for the PHD targets described.

Methyl Lysine Reader Domain Inhibitors

Although the research into inhibitors of methyl lysine reader domains is not as advanced as research into bromodomain or kinase inhibition, several methyl lysine reader domain families have at least one reported inhibitor.[42] There are several reported inhibitors of MBT domains and a peptidomimetic inhibitor of the chromodomain of Chromobox Homologue 7 (CBX7).[71,72] A small-molecule inhibitor of the third PHD of the demethylase JARID1A has been reported,[73] and a systematic ligandability assessment of the PHD of pygopus family PHD finger 1 (PYGO1) has identified a series of benzimidazole ligands.[74] There are currently no reported inhibitors of Tudor domains or PWWP domains.

MBT Domain Inhibition

Inhibitors for the MBT domain of lethal (3) malignant brain tumour-like 1 (L3MBTL1) and the closely related L3MBTL3 have been developed by Frye and co-workers (Figure 19).[71,75–77] An initial hit **1** with a K_D of 37 µM for L3MBTL1 was based on a peptide that mimics the natural H4K20me2 ligand. The residue H4R19 was changed to a phenylalanine in order to simplify synthesis, and a series of methylated lysine mimics tested, with a pyrrolidine proving to be the most potent. The desire to identify a non-peptidic inhibitor with good predicted cell permeability and suitability for analogue synthesis using simple chemical techniques lead to the discovery of compound **2**. ITC measurements revealed that compound **2** showed an unfavourable loss of entropy on binding, which lead to the design of compound **3** with a more rigid linker between the pyrrolidine and the core. ITC measurements showed that although the ΔH of binding to L3MBTL1 was reduced from -15 kcal/mol to -13 kcal/mol on changing from

compound **2** to compound **3**, the TΔS of binding increased from -9 kcal/mol to -6 kcal/mol. This enthalpy-entropy compensation leads to a 5-fold drop in K_D. In order to improve potency a series of molecules that incorporated two pyrrolidine methyl-lysine mimetics was synthesised and lead to UNC1215, a chemical probe with nM affinity for L3MBTL3. This compound has been shown to antagonise the localisation of L3MBTL3 to methyl-lysine in cells and showed 75-fold selectivity for L3MBTL3 over L3MBTL1. X-ray crystallography revealed that UNC1215 binds in an unusual 2:2 binding mode with one pyrrolidine interacting with the first MBT domain of L3MBTL3 while the other pyrrolidine interacts with the second MBT domain of a different molecule of L3MBTL3. A second molecule of UNC1215 binds in an identical manner, occupying the first MBT of one molecule of L3MBTL3, and the second MBT of the other molecule of L3MBTL3 (Figure 20).

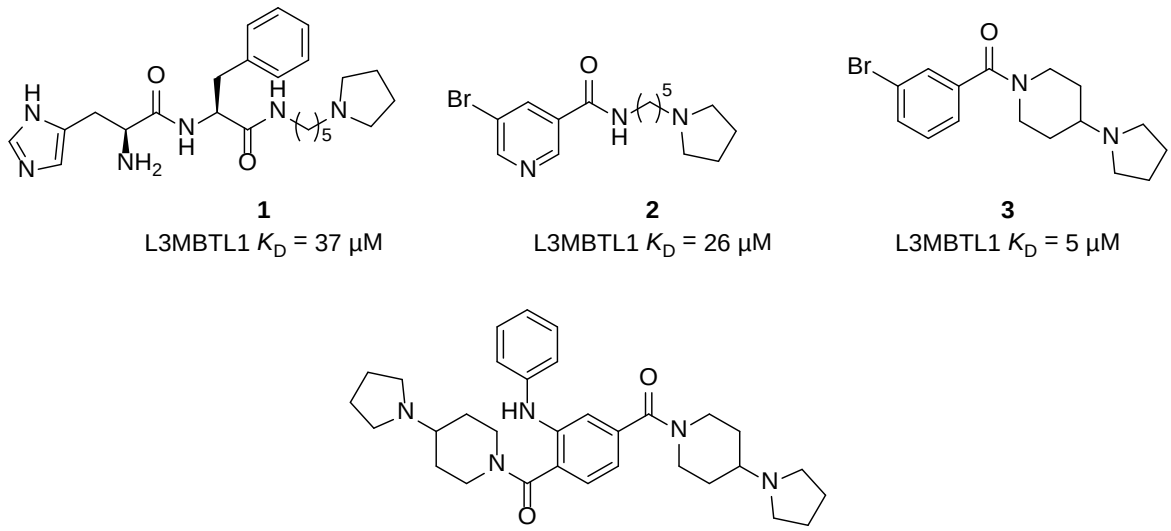


Figure 19. The evolution of inhibitors of L3MBTL1. An initial peptidic inhibitor **1** was developed into the non-peptidic inhibitor **2**. Replacement of the flexible alkyl chain with a piperidine led to the more rigid inhibitor **3,** which showed a reduced loss of entropy on binding compared to compound **2.** The addition of a second pyrrolidine group lead to UNC1215, a potent and selective inhibitor of L3MBTL3.

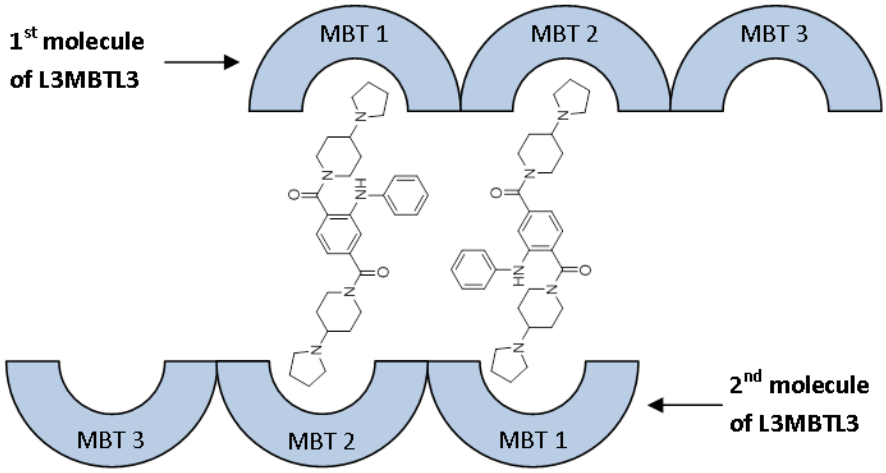


Figure 20. The co-crystal structure of the triple MBT domain of L3MBTL3 with UNC1215 shows an unusual 2:2 binding mode. One molecule of UNC1215 interacts with the first MBT of one molecule of L3MBTL3with one of its pyrrolidines, and the second MBT of another molecule of L3MBTL3 with its other pyrrolidine. A second molecule of UNC1215 completes the 2:2 complex by binding in an identical manner.

Chromodomain Inhibition

The first and only reported inhibitor of a chromodomain is a peptidomimetic inhibitor of the chromodomain of CBX7 compound **4** (Figure 21). This inhibitor was designed by modifying the *N*-terminus, *C*-terminus, and selected residues of a 5-mer peptide inhibitor. The initial 5-mer was chosen to mimic the natural substrate of CBX7. This peptide consisted of residues 24-28 of histone 3, with a trimethyl lysine at position 27. The final inhibitor **4** has a K_D of 200 nM and contains a tri-methylated lysine residue. This inhibitor shows 10-fold selectivity over CBX8 (88% sequence identity), but only 1.5-fold selectivity over CBX4. No cell permeability or activity data is provided for this peptidomimetic inhibitor.

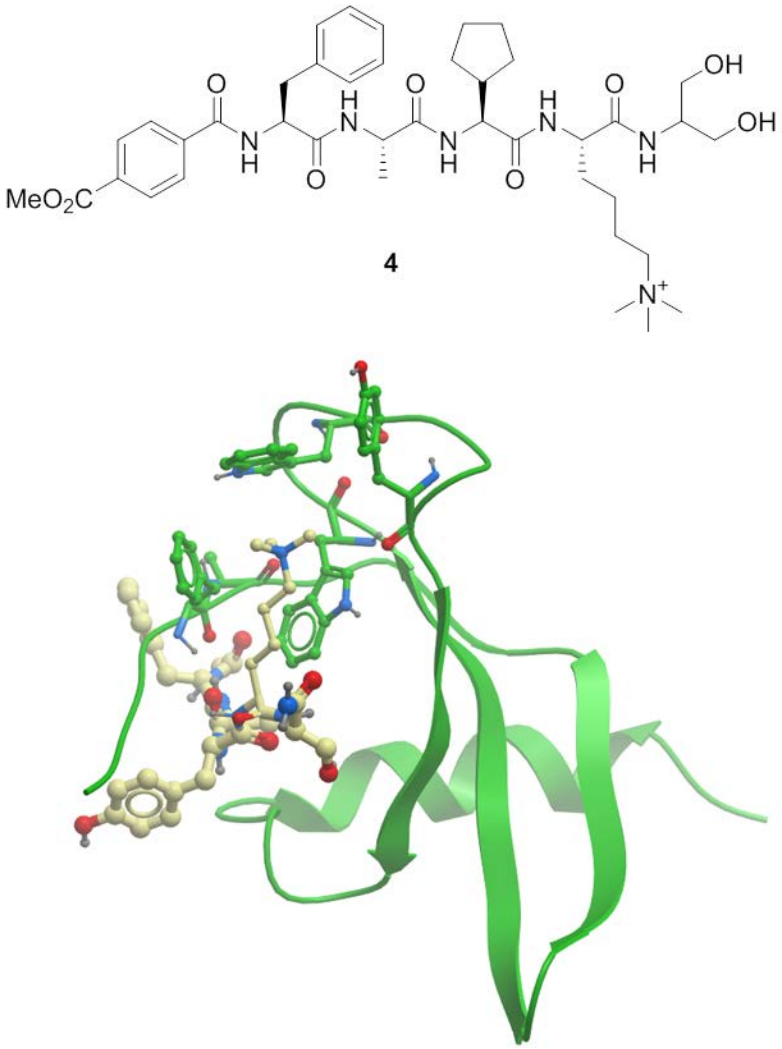


Figure 21. The modified peptide **4** is an inhibitor of chromobox 7 (CBX7). It binds with the trimethylated lysine in the aromatic cage of CBX7. PDB ID: 4MN3.

PHD Inhibition

An analysis of the ligandability of a range of methyl lysine binding sites has been carried out by Santiago et al.[78] This study suggests that PHDs are likely to be less ligandable than other families of methyl lysine reader domain, such as chromodomains and WDR domains. However, this study only covers a small fraction of known PHD structures, and therefore it is possible that it has missed out some more ligandable PHDs. A complete analysis of all available PHD structures is described in Chapter 3.

There are currently two publications describing an attempt to discover an inhibitor of a PHD. The first published work by Wagner et al.[73] describes the use of a luminescence assay to screen the NIH clinical collection I[a] against the third PHD of the lysine demethylase JARID1A. The assay makes use of a HaloTag linker to covalently attach the protein to a chloroalkane coated surface. The immobilised protein is then incubated with a biotinylated histone 3 peptide and the compound under investigation. The plates were vigorously washed to remove unbound peptide. The plates were then incubated with a streptavidin conjugated horseradish peroxidase (HRP) and again washed vigorously. The plates were then read using a plate reader, a luminescent signal indicated that the streptavidin conjugated HRP was bound to the biotinylated histone 3 peptide, which in turn was bound to the protein (Figure 22A). Loss of signal implied that the peptide was no longer bound to the protein due to compound inhibition. After eliminating false positives from their initial screen, amiodarone was the only confirmed inhibitor of JARID1A PHD3, (Figure 22B). Testing of analogies of amiodarone and known metabolites identified di-*N*-desethylamiodarone as the most potent inhibitor, with an IC_{50} of 26 ± 15 μM (Figure 22C). Mutagenesis of residue D1624 and D1629 indicate that these residues are important for the binding of di-*N*-desethylamiodarone to the PHD. Figure 22 shows potential binding modes identified by a docking experiment, with the positively charged amine interacting with either D1629 (Figure 22D) or D1624 (Figure 22E).

[a] 446 compounds that have been through phase I-III clinical trials, encompassing a broad range of clinical indications. The biosafety and bioavailability of these compounds have been highly characterised.

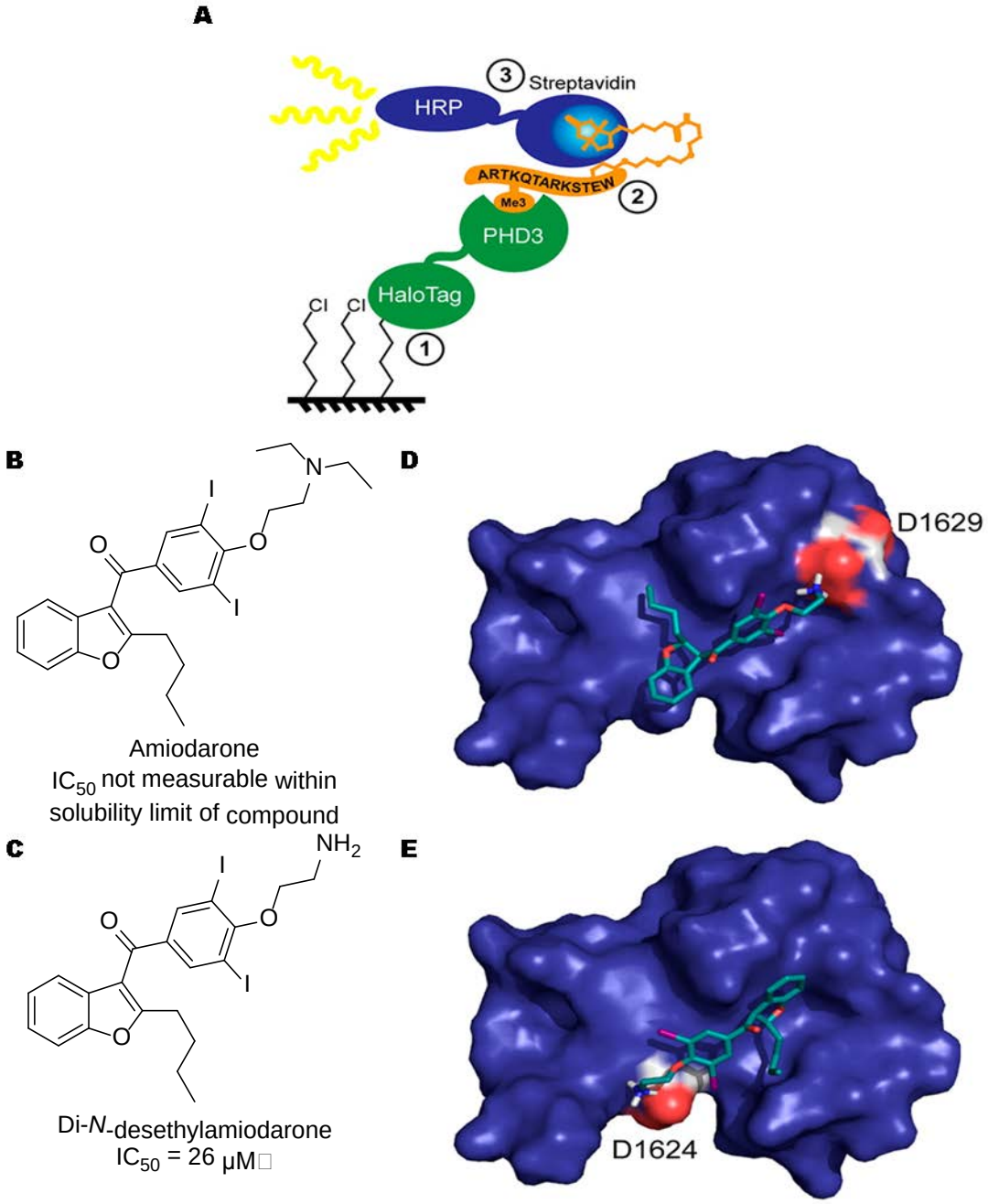
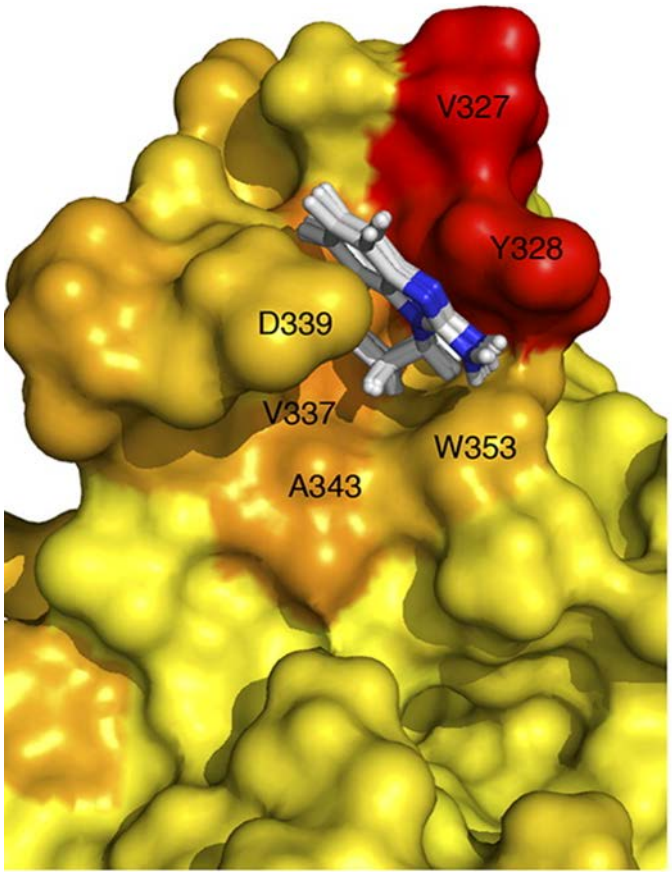
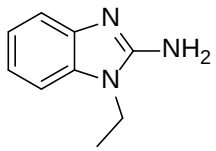


Figure 22. A. Depiction of the HaloTag assay design. **B.** Amiodarone was initially identified as an inhibitor of the third PHD of JARDI1A. **C.** di-*N*-desethylamiodarone was identified as the most potent analogue of Amiodarone with an IC_{50} of 26 μM. **D.** and **E.** Two potential binding modes of di-*N*-desethylamiodarone with the third PHD of JARID1A. Residue D1629, which is predicted to make an ionic interaction with the amine of di-*N*-desethylamiodarone, is highlighted. The PHD is shown in a different orientation in each of the two binding modes. Adapted from *Identification and Characterization of Small Molecule Inhibitors of a Plant Homeodomain Finger*[73].

Miller et al carried out a systematic analysis of the ligandability of the PHD of pygopus family PHD finger 1 (PYGO1), and have identified a series of benzimidazole ligands that bind into the H3K4me3 binding pocket (Figure 23).[74] The compounds were identified by an initial NMR fragment screen, and studied further by NMR. Although the compounds identified in this study only have binding affinities in the micromolar range, they demonstrate an ability to competitively inhibit the binding of the PHD to a peptide containing H3K4me2 (ARTKme2Q).





CF16
K_d = 7.3 ± 2.0 mM

Figure 23. Cluster of top five poses of CF16 (in stick representation; blue, nitrogen) docked into the distal K4me pocket of PHD−HD1 (4UP0, in surface representation), as calculated by HADDOCK, based on unambiguous restraints derived from 33 intermolecular NOEs between the compound and protein, and ambiguous restraints derived from chemical shift perturbations (colouring thresholds: yellow <0.04 ppm; orange <0.1 ppm; red <0.15 ppm). Key interacting residues of hPygo2 are labelled. Figure adapted from *Competitive Binding of a Benzimidazole to the Histone-Binding Pocket of the Pygo PHD Finger.*[74]

Aim of Project

This project aimed to assess the potential of PHDs to bind small molecule inhibitors using computational analysis of published PHD structures. As well as providing an assessment of the ligandability of PHDs, this project will also provide a family wide analysis of human PHDs.

Initially a search of the RefSeq database[79] was performed to identify all human PHDs. A structure -based sequence alignment was used to create a PHD family tree, which provides a valuable resource for studying the PHD family. The Protein Data Bank (PDB)[80] was mined for all available protein structures. These were analysed *in silico* to identify potential small molecule binding sites. This knowledge was combined with the PHD family tree in order to identify sub groups of PHDs that are more likely to be amenable to small molecule inhibition. This work included a special focus on PHDs that form part of a multi-domain complex, as the analysis described in Chapter 3 highlighted these as being the most ligandable.

The identification of two examples of ligandable PHDs was followed by developing suitable assays to screen small molecule libraries against these PHDs. The characterisation of initial hits from small molecule screens provides a starting point for the development of selective small molecule chemical probes.

Chapter 2 - PHD Family Analysis

Although many literature reviews of human PHDs are available,[81–84] there is no current systematic analysis of the human PHD family. In fact, there is no currently reported work which attempts to define the human PHD family. A definitive list of PHD family members, with an analysis of sequential relationships and family determining features would be an invaluable resource to researchers. Such an analysis would facilitate prediction of histone substrate specificity for unstudied PHDs, and also allow researchers developing PHD inhibitors to design appropriate selectivity panels. Work described in this chapter provides a definitive list of human PHDs, and the construction of a phylogenetic tree with an analysis of key residues and features that allows PHDs to be grouped into sub families.

Identifying PHD Family Members

There has been no previous attempt to define a comprehensive list of all human PHDs. To this end, we have enacted a thorough search of human proteins in the RefSeq database using HMMER.[85]

Choosing a Suitable Protein Sequence Database

There are numerous available databases for protein sequences including RefSeq[79] and GenBank[86] maintained by the National Center for Biotechnology Information (NCBI), and UniProt[87] and SwissProt[88] maintained by the UniProt consortium. The RefSeq database was chosen as it is a non-redundant, curated database with a single entry for each protein sequence in the human proteome.[89] The best available data for each protein sequence is stored in the database, with sequences updated if new information becomes available. This differs from the GenBank database which can contain multiple records for a single protein, as new sequence data is added to the database alongside existing data.[89] The UniProt consortium maintained databases UniProt and SwissProt have a similar relationship to each other as the two NCBI maintained databases describes above. UniProt is a non-curated database similar to GenBank,

and SwissProt is a non-redundant, curated database similar to RefSeq. RefSeq was chosen for this work, as a curated database with the most up-to-date sequence information is most appropriate. The SwissProt database would also have proved suitable; however, the availability of colleagues with experience of using the RefSeq database was a determining factor in this choice.

Choosing Suitable Database Searching Software

There are multiple tools available to search for sequence homologues within a sequence database, the commonly used tools that were considered for use in this work were HMMER, and Basic Local Alignment Search Tool (BLAST) and its derivative Position-Specific Iterated BLAST (PSI-BLAST). These three tools are summarised briefly below.

HMMER is a hidden Markov model (HMM) based tool for searching sequence databases for homologues to an input multiple sequence alignment (MSA).[90] It can be used for both amino acid and nucleotide sequences. The input alignment is initially converted into a profile HMM, which describes the probability of finding any given residue (or a gap) for each column of the input alignment. This type of method is known as a position-dependent method.

Basic Local Alignment Search Tool (BLAST) is a pairwise alignment search tool.[91] The BLAST algorithm initially breaks down the query sequence into shorted sequences known as words, where the default word length is 3. The algorithm calculates all possible words for the query sequence and also generates a complete set of neighbourhood-words, which are words that meet a predefined similarity score with a word in the initial set. The algorithm searches for matches for the query words in the database of sequences. These words are then extended in both directions provided the alignment score between the query sequence and hit sequence remains above a predefined threshold. BLAST uses a single sequence query, which differs from HMMER which uses a HMM built from a multiple sequence alignment. Another difference between HMMER and BLAST is that the substitution matrix used to calculate the alignment

score between the query sequence and the hit sequence in BLAST is position-independent. Position-Specific Iterative BLAST (PSI-BLAST) is a version of BLAST that facilitates the use of a multiple sequence alignment as a search query rather than a single sequence, and is therefore a position-dependent method similar to HMMER.[92]

For this work PSI-BLAST and HMMER were considered, as they allow a multiple sequence alignment of known PHDs to be input as a search query. HMMER was chosen as this treats gaps in a position dependent manner, whereas PSI-BLAST applies set gap opening and extension penalties independent of the position in the sequence.

Search Process for Identifying PHD Family Members

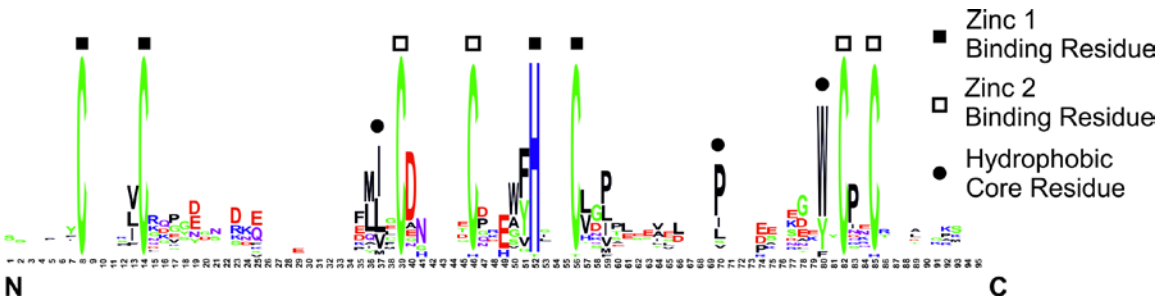


Figure 24. Sequence logo of the multiple sequence alignment of the forty PHDs identified from the PDB. Strong consensus of the Zn(II) binding residues can be seen. Zinc binding residues are indicated with a square, and residues which comprise the hydrophobic core of the domain are marked with circles.

A multiple sequence alignment of known PHDs was constructed for use as the initial query in the search of the RefSeq database. The Protein Data Bank (PDB)[80] was searched for known PHDs using a simple text search, and human PHDs were extracted from the search results. The decision to use only PHDs with known structures in the initial alignment allows a high degree of confidence that no non-PHDs have been included, and also facilitates the construction of a multiple sequence alignment based on knowledge of the relative spatial positioning of residues. This use of structural information to build a multiple sequence alignment has previously been used for the bromodomain family.[27] This initial search of the PDB yielded forty PHDs, which were

aligned based on structural similarities. (Appendix 2.1) (Figure 24). Any *N*-terminal or *C*-terminal domains present in the structures were removed.

HMMER was used to build this initial alignment into a profile HMM that could be used to search the RefSeq database for additional PHDs. The output sequences from the first search were used to create a profile HMM for the second search, and the process iterated until the number of hits stabilised at 194 after four searches. Redundant sequences from alternative isoforms were removed to give 130 unique human PHDs (Appendix 2.2). Between each search the list was inspected and any sequence too short to be a full PHD was removed. Similarly, any sequences not containing eight Zn(II) binding residues were removed. For this reason, the list of PHDs only contains four PHDs from Wolf-Hirschhorn syndrome candidate 1-like 1 (WHSC1L1), rather than the five seen in the close homologues Wolf-Hirschhorn syndrome candidate 1 (WHSC1) and nuclear receptor binding SET domain protein 1 (NSD1).[b] This is because the putative fourth PHD of WHSC1L1 is missing the first Zn(II) binding residue. As it is known that mutation of a single Zn(II) binding residue can be deleterious to PHD structure and function (Chapter 1), it was decided to not include this putative PHD in the complete list of PHD sequences until suitable evidence can be produced that it adopts a canonical PHD fold. It should be noted that there are literature examples of the putative 4th PHD of WHSC1L1 being described as a PHD.[93]

Inspection of this list revealed that some known atypical PHDs were missing. For example, the PHDs of E1A binding protein p300 (EP300) and cAMP response element-binding protein-binding protein (CREBBP), which have since been shown to have an unusual insert between the first and second pair of Zn(II) binding residues[94] were not picked up by this search. This prompted a search of Pfam[95] and Simple Modular Architecture Research Tool (SMART)[96] databases, as well as a search of scientific literature which identified a further forty-one PHDs.

[b] WHSC1 and WHSC1L1 are also commonly known as NSD2 and NSD3 respectively.

In order to give added confidence that no PHDs had been missed during this search, each individual PHD was used as a query in a BLAST search of the RefSeq database. This search was intended to identify any PHD that may have a high sequence identity with one member of the family, but not fit well with the profile HMM used in the HMMER search. This search yielded a further two PHDs, giving a final total of 173 (Appendix 2.3) (Figure 25). This number is less than the "> 200" claimed by a 2012 review.[81] However, the review does not state how this number was calculated.

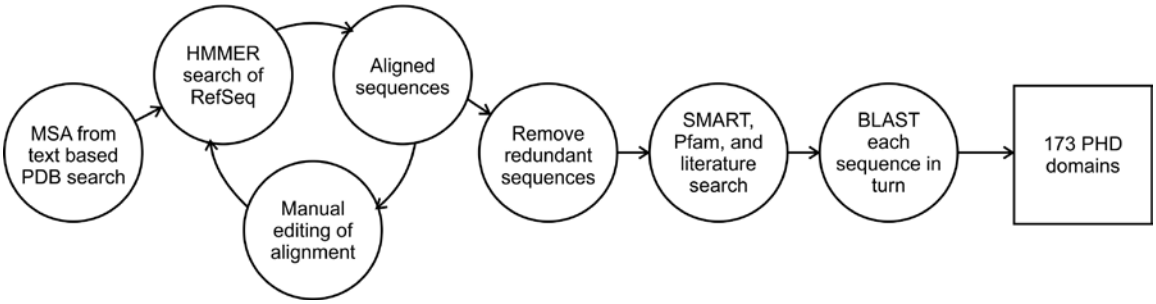


Figure 25. Schematic representation of the identification of PHD family members. MSA = Multiple Sequence Alignment.

In order to test for false positives, a similar search was conducted for the closely related RING domains (Chapter 1). This search yielded 206 putative RING domains (Appendix 2.4), with no overlap with the list of 173 putative PHDs. This indicates that the rate of false positives from the RING family in the list of 173 putative PHDs is likely to be low.

Defining PHD Domain Boundaries

PHDs identified by the search method described above varied in sequence length. In order to remove noise from the multiple sequence alignment it was necessary to define the *N*-terminal and *C*-terminal boundaries of these PHD domains. Structural superimposition of available PHD structures indicated that there was little structural similarity for residues more than four residues *N*-terminal of the first zinc-binding residue, therefore it was decided to trim the identified sequences such that only three residues were left to the *N*-terminal side of the first

zinc-binding residue. A similar structural analysis lead to the removal of all residues more than three residues *C*-terminal of the last zinc-binding residue. Although this method is biased by the construct design of the structural biologists who solved the analysed structures, inspection of the identified sequences suggested there was no sequence conservation outside of the identified domain boundaries.

Structural Based Alignment of PHD Sequences

A multiple sequence alignment of all identified PHDs is required to construct a phylogenetic tree for the PHD domain family. Numerous programs for determining multiple sequence alignments are available.[97,98] Due to the importance of the Zn(II) binding to the fold of the PHD, it was decided that the minimum requirement for a PHD multiple sequence alignment was that the Zn(II) binding residues were correctly aligned. Trials with a range of automated alignment programs[c] failed to produce multiple sequence alignments with correct alignment of the Zn(II) binding residues; therefore it was decided to manually align the PHD sequences based on structural information.

The structure based multiple sequence alignment of the forty PHDs used for the initial HMMER search was used as a seed alignment. Each new sequence was added to this alignment individually. For each new sequence the individual sequence within the alignment with the highest sequence identity was identified. The new sequence was aligned to the identified sequence ensuring that Zn(II) binding residues were correctly aligned.

Construction of PHD Tree

The construction of phylogenetic trees is a common technique for analysing evolutionary relationships between proteins.[99] In this instance, the construction of a PHD phylogenetic tree is primarily to allow the grouping of PHDs into families predicted to have similar histone binding

[c] The programs used were: Molsoft ICM, the MAAFT web server, and the ClustalW web server.

sites. This will direct future medicinal chemistry efforts to develop selective PHD inhibitors and predict the function of those PHDs that have not yet been studied.

Tree Construction Methods

Methods used for phylogenetic analysis of multiple sequence alignments fall into three categories: distance, parsimony, and likelihood.[100] Distance methods work by grouping sequences based on pairwise sequence similarities. They are computationally fast, but not regarded as reliable as parsimony and likelihood methods.[100] Parsimony based methods search for the simplest evolutionary model that explains the input sequence data. Advocates for parsimony based tree building methods claim that a simple model requiring fewer assumptions is more reliable a more complex alternative model.[101] However, this argument has been criticised as being statistically unsound.[102] Likelihood methods are used to find a hypothetical model that best predicts the observed data. In the case of phylogenetic tree construction, the observed data is the input multiple sequence alignment, and the phylogenetic tree represents the evolutionary model. Therefore likelihood based methods calculate the phylogenetic tree most likely to have given rise to the input multiple sequence alignment. Likelihood methods are more statistically rigorous than distance or parsimony based methods, but this also makes them the most computationally expensive method of the three.

This work used the method of tree construction described by Hall.[103] Hall sets out a clear strategy for the construction of maximum likelihood phylogenetic trees using the Molecular Evolutionary Genetics Analysis (MEGA) software package.[104] Although many other tree building software packages are available, the choice of MEGA for this work was based on its ease of use.

It is necessary to select a suitable substitution model when using a maximum likelihood method to estimate a phylogenetic tree. The substitution model dictates the probability of a given residue mutating to another residue at each point in the sequence. MEGA contains built-in functionality to calculate the best substitution model to use for a given dataset.[105] In this

instance the analysis of the PHD multiple sequence alignment by MEGA suggested that a Whelan and Goldman (WAG) model[106] with the assumption of the existence of evolutionary rate variation among sites and the presence of invariant sites. This model is listed in MEGA 6.0 as WAG + G + I.

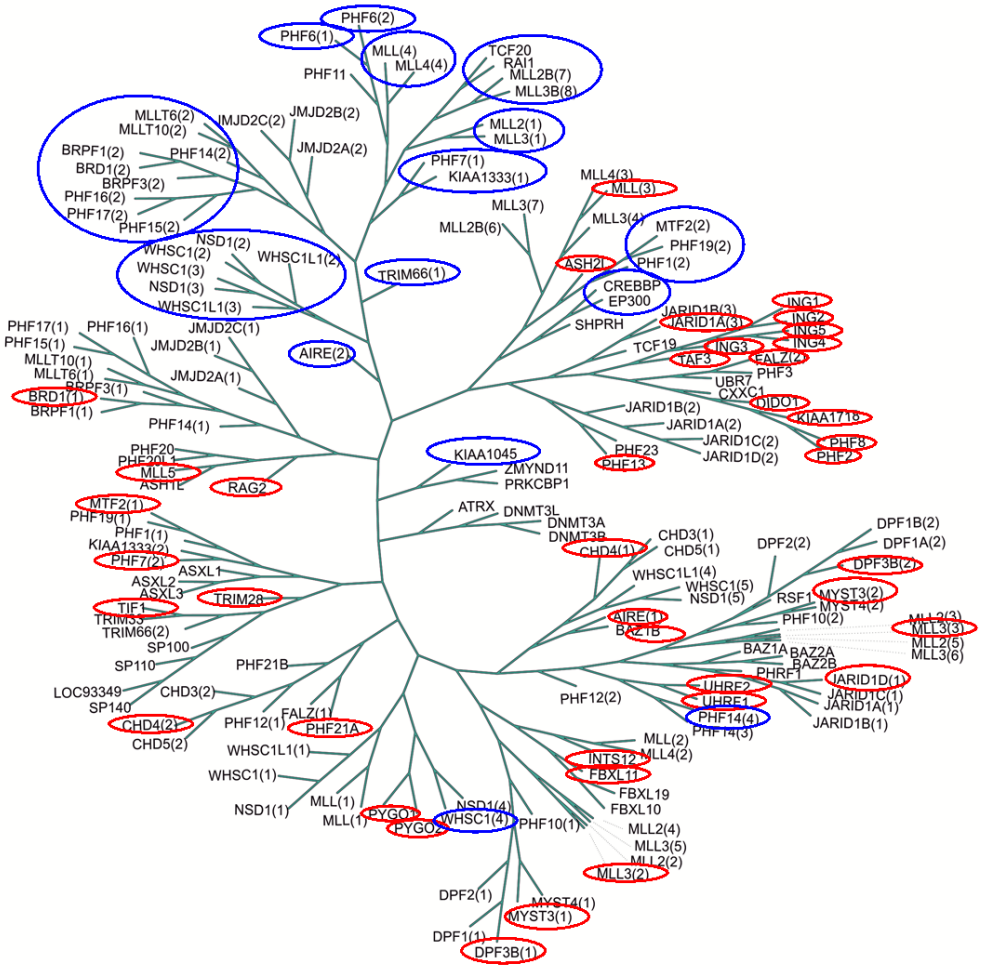


Figure 26. PHD tree showing relationships between human PHDs**.** In the case where multiple PHDs are found within the same protein, the domain number is indicated in following brackets. E.g. the second PHD of the protein AIRE is labelled as AIRE(2). PHDs used to build the initial sequence alignment for the HMMER search indicated by red circles. PHDs added after the HMMER search as a result of searches of SMART, Pfam, and scientific literature are highlighted by blue circles. The wide distribution of these PHDs within the tree indicates that they were a suitable sub-set of PHDs to use for the HMMER Search.

The main aim of the construction of the PHD phylogenetic tree was to group PHDs into sub-families; therefore it was decided to create a bootstrap consensus tree. A bootstrap tree is built by running the tree building software multiple times (in this instance 500), and combining the multiple outputs (which will show slight differences) into a single consensus tree. This method has the disadvantage of creating a tree with uniform branch-lengths. The phylogenetic tree for the PHD family constructed by the methods described above is shown as an un-rooted tree in Figure 26.

Position within Tree of Initial Forty PHDs

The tree was analysed to identify the positions of the forty PHDs used to build the initial profile hidden Markov model (Figure 26). It is apparent that the initial forty PHDs are well dispersed around the tree. The majority of PHDs added as a result of further searches of SMART, Pfam, and the scientific literature are found in regions of the tree where no members of the original initial sequence alignment are found. This suggests that these further steps were necessary for discovering all PHDs.

Analysis of PHD Tree

As previously discussed PHDs bind histones in a manner dependant on the pattern of post-translational modifications on the histone tail, and are found in a wide range of multi-domain proteins (Chapter 1). For example some PHDs possess an aromatic cage which binds to H3K4me3, whereas others do not contain this residue and will not bind H3K4me3. This section analyses the PHD tree to identify patterns of histone binding specificity, domain architecture, and PHD structural features within the PHD tree. The presence of PHDs with similar properties in the same sub-families of the tree provides a level of validation for the tree, and also allows predictions to be made about the structure and function of PHDs for which there is currently no experimental data.

Comparing Structural Features of Related PHDs

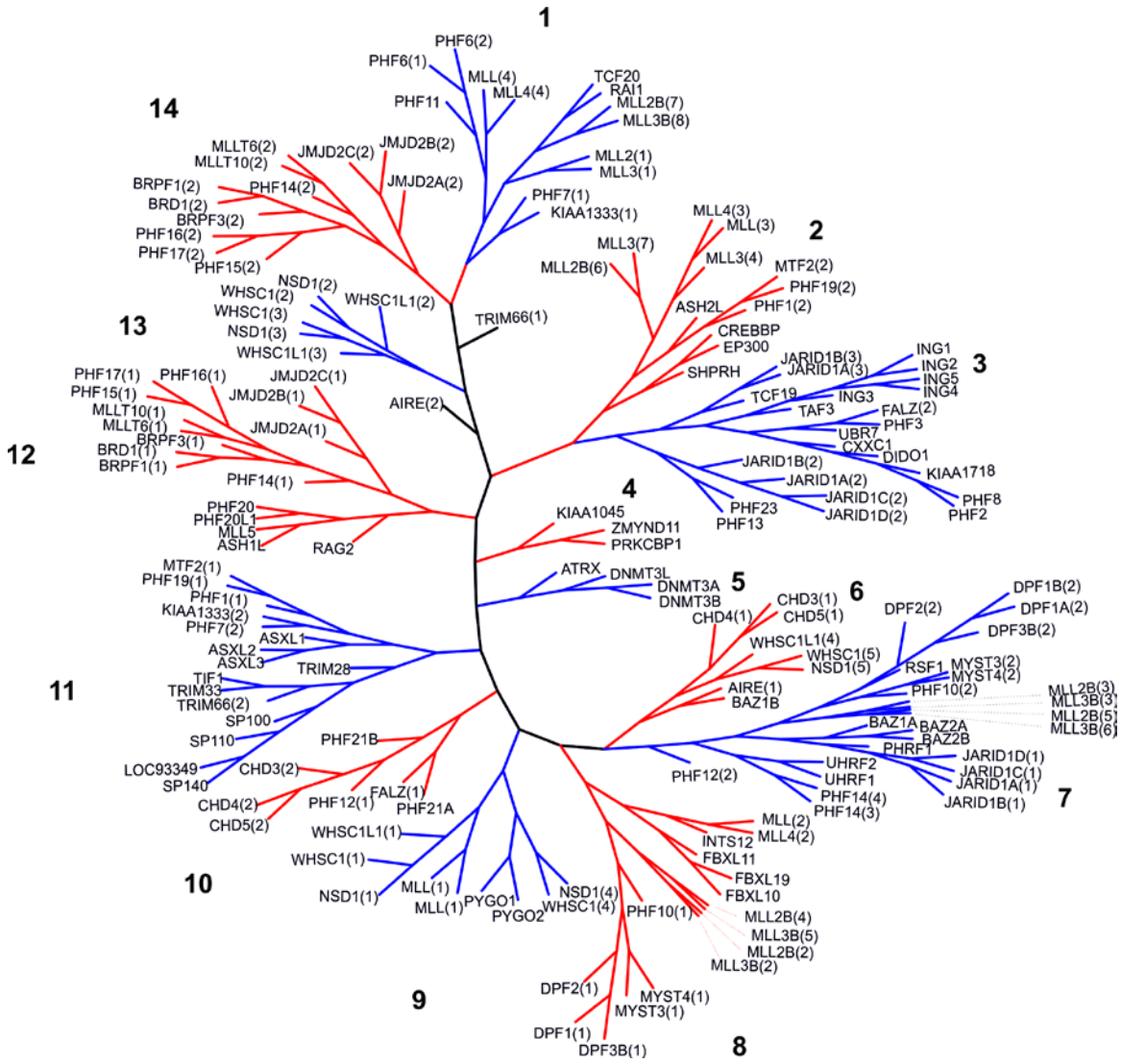


Figure 27. The 14 sub-families of the PHD tree are labelled and alternatively coloured red and blue. TRIM66(1) and AIRE(2) are atypical and not part of any sub-family.

Inspection of the tree suggests that it can be split into 14 sub-families, with two atypical PHDs that do not fit into any of these sub-families (Figure 27). These will be discussed in turn in the following section. Specifically, the presence of conserved sequential and structural features, histone binding specificity, and domain architecture will be analysed.

Sub-Family 1

Sub-family 1 contains 13 PHDs, and is one of the few sub-families in which no structural data is available. A notable feature of this sub-family is the complete conservation of a GA motif in loop 1; it is possible that is involved in the formation of a sharp turn. Other notable features include the presence of a histidine as the eighth Zn(II) binding residue, in place of the usual cysteine.

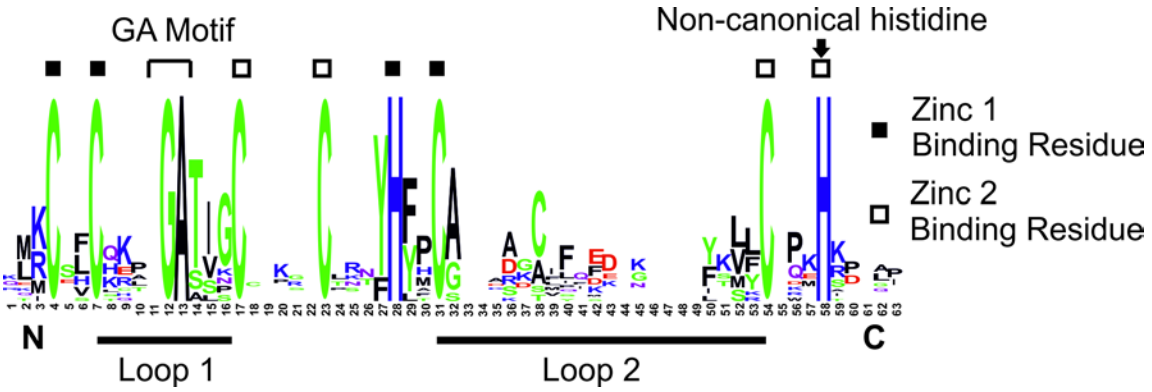


Figure 28. Sequence logo of sub-family 1. The GA motif in loop 1 is highly conserved throughout this sub-family. The eighth Zn(II) binding residue, which is histidine in place of the usual cysteine, is indicated.

The close grouping of MLL(4) and MLL4(4) is expected due to the similarity in domain architecture between these two proteins. Similarly, the close grouping of MLL2(1) and MLL3(1), as well as MLL2(7) and MLL3(8) is expected based on the similar domain architecture of these two proteins.[d] This sub-family does not possess residues capable of forming an aromatic cage; therefore it is highly unlikely to bind H3K4me3.

Sub-Family 2

Sub-family 2 is a diverse group containing some anomalous PHDs. This includes Absent Small or Homeotic-like 2 (ASH2L), which adopts a PHD fold as part of a larger winged-helix domain, despite only having one Zn(II) binding site.[107] This group also contains the histone acetyl transferases CREBBP and EP300. These proteins contain PHDs interrupted by a Really Interesting New Gene Domain (RING).[94]

[d] Note that MLL2 is missing a PHD in the position where the fourth PHD of MLL3 is found. Therefore MLL2(7) and MLL3(8) are found in similar positions within the sequence of these two proteins.

Sub-Family 3

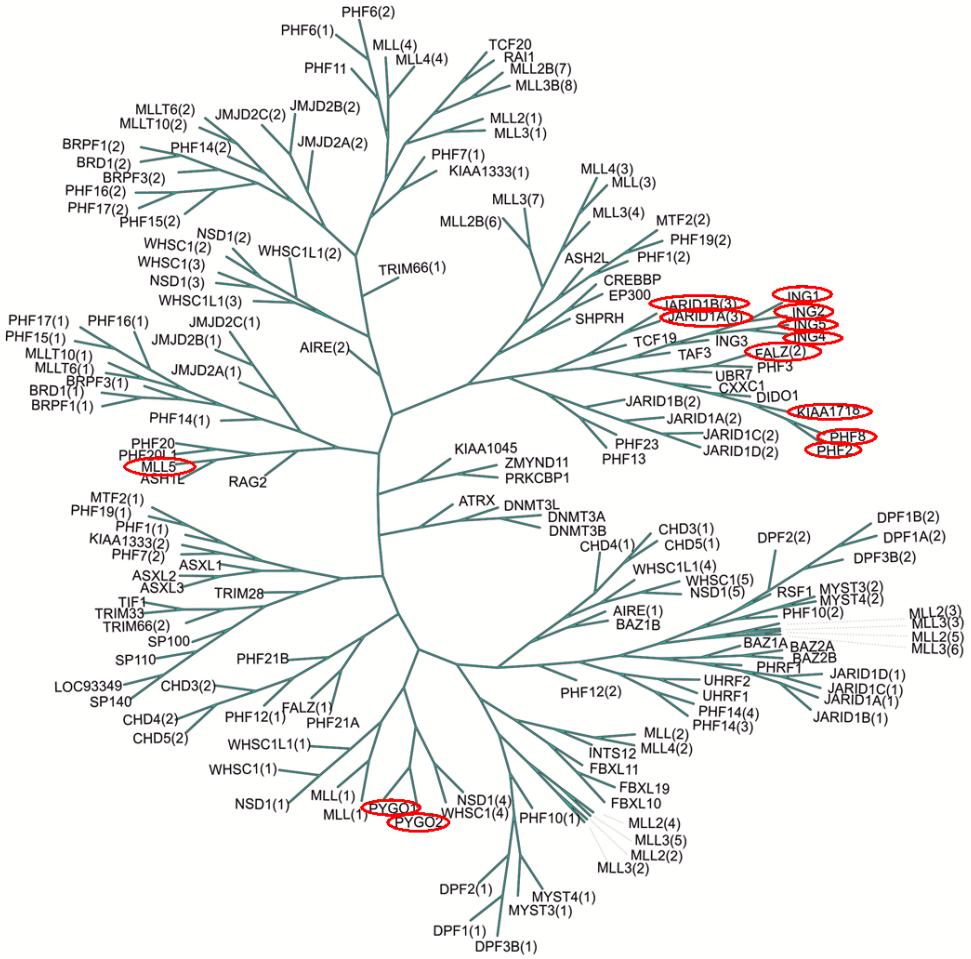


Figure 29. The positions of known H3K4me3 binders are highlighted on the tree with red circles. These H3K4me3 binders primarily cluster in sub-family 3. It is highly likely that sub-family 3 members with no reported function are H3K4me3 binders.

Sub-family 3 contains almost all reported H3K4me3 binders (Figure 29). All PHDs except JARID1B(2) contain a tryptophan residue two residues *N*-terminal of the fifth Zn(II) binding residue. This residue has been shown to form a cation-pi interaction with the quaternary amine during H3K4me3 recognition (Figure 30).[51,108–111] There is a strong conservation of either a methionine or an aromatic amino-acid 3 residues *N*-terminal of the third Zn(II) binding site. This residue forms that back of the aromatic cage and although methionine is not capable of forming an cation-pi interaction, it has been suggested that the polarisability of the thioether allows it to act as a suitable substitute for an aromatic residue.[112]

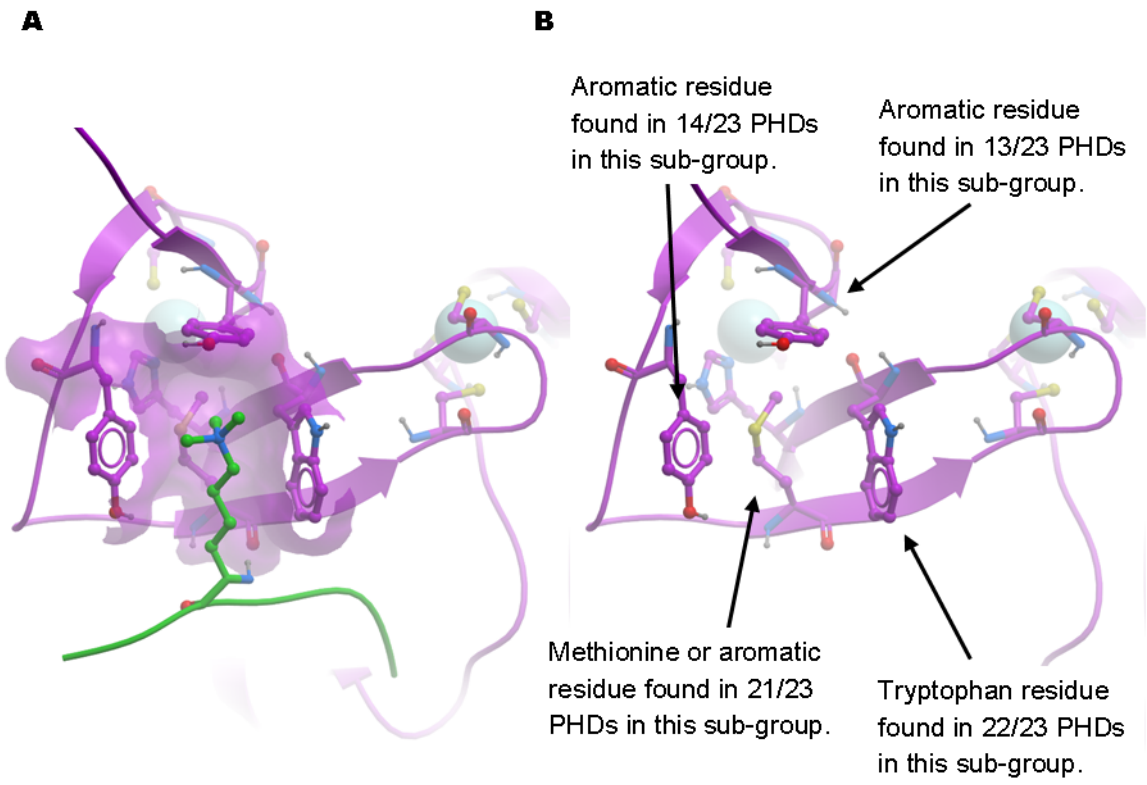


Figure 30. Depiction of aromatic cage residues found in sub-family 3. **A.** The trimethylated lysine of histone 3 (green) is shown in a hydrophobic pocket formed by three aromatic residues and a methionine. **B.** Depiction of the same aromatic cage with the histone 3 peptide removed. Each key cage forming residue is labelled with its prevalence within sub-family 3. The PHD shown is the PHD of PHD Finger Protein 8 (PHF8) PDB ID: 3KV4.

The prevalence of possible aromatic cage-forming residues within this sub-family would strongly suggest that members of this sub-family with no reported function are likely to preferentially bind methylated lysines.

Sub-Family 4

Figure 31. Sequence alignment of all three members of sub-family 4. There is no sequence conservation within loop 2, but good conservation in the central region of the PHDs.

Sub-family 4 is a small sub-family containing only three members. Zinc Finger, MYND-type containing 11 (ZMYND11) and Protein Kinase C Binding Protein (PRKCBP1 - also known as

ZMYND8) possess the same domain architecture in there *N*-terminal regions, namely a PHD-Bromodomain-PWWP region. It is therefore not unsurprising that their PHDs are closely related. KIAA1045 is a gene with no reported function, and with no other domains according to Pfam.[95] The lack of structural or functional data for this sub-family means that it is difficult to make any connections between sequence and function. It is also interesting to note that although there is good sequence conservation in the central region of the PHD, there is little sequence conservation within the loop 2 region (Figure 31).

Sub-Family 5

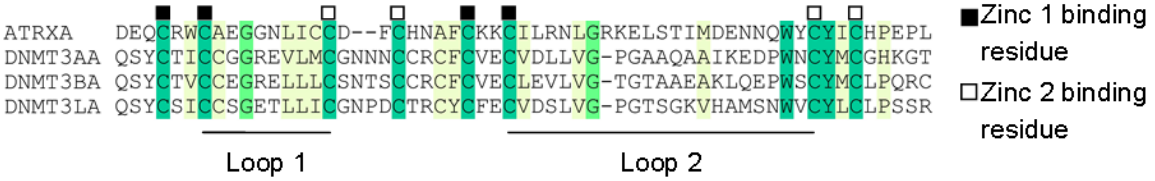


Figure 32. Sequence alignment of all four members of sub-family 5. This group contains the three closely related DNA methyl transferases DNMT3A, DNMT3B, and DNMT3L and more distantly related ATRX. Zinc 1 binding residues are indicated with black squares, zinc 2 binding residues indicated with white squares.

The PHDs found in sub-family 5 are all part of a larger ATRX-DNMT3-DNMT3L (ADD) domain. This is an enlarged domain containing an extra Zn(II) binding site *N*-terminal of the PHD (Figure 33), and is only found in the four proteins of sub-family 5.

The PHDs of this sub-family do not contain aromatic cage residues, and are known to preferentially bind histone 3 peptides unmodified at H3K4.[45,59,113] The ADD domain of alpha Thalassemia/Mental Retardation Syndrome X-linked (ATRX) contains a YY motif which provides an aromatic cage where H3K9 binds to the surface of the domain.[114] Therefore ATRX preferentially binds a histone 3 peptide trimethylated at H3K9. This recognition of H3K9me3 happens simultaneously to recognition of H3K4me0 at different binding sites.

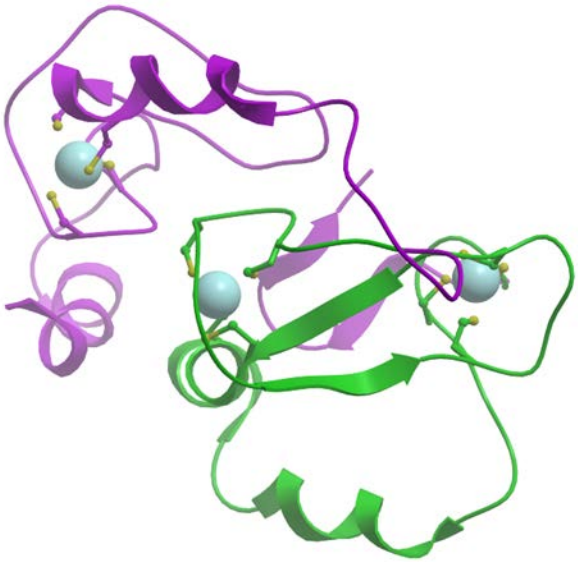


Figure 33. The ADD domain of DNMT3L. The PHD within the ADD domain is highlighted in green, with the other sections of the ADD domain shown in magenta. The ADD domain contains a third Zn(II) binding site as well as the two canonical Zn(II) binding sites found within the PHD. The ADD domain binding to unmodified histone 3 peptide via the PHD section in a canonical manner (not shown). PDB ID: 2PVC.

Sub-Family 6

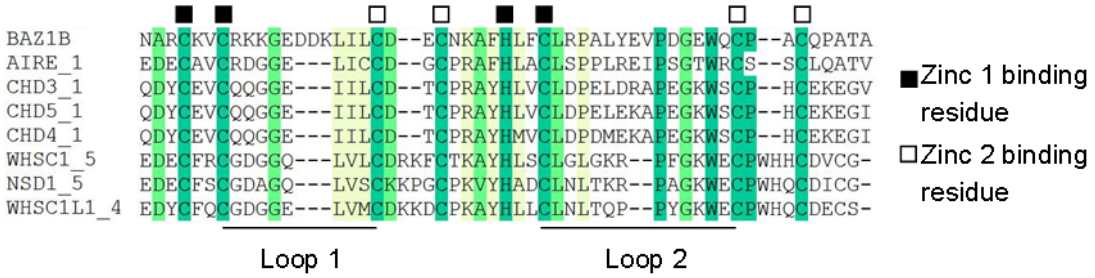


Figure 34. Sequence alignment of all members of sub-family 6. Zinc 1 binding residues are indicated with black squares, zinc 2 binding residues indicated with white squares. This sub-family contains a well conserved PxGxW motif in loop 2.There is also a strong conservation of acidic residues prior to the first Zn(II) binding residue. These acidic residues are likely to play a role in the recognition of unmodified H3K4.

Sub-family 6 contains 8 PHDs, although the positioning of Bromodomain Adjacent to Zinc Finger Domain 1B (BAZ1B) in this group is questionable, as BAZ1A, BAZ2A, and BAZ2B are found in sub-family 7. A key feature of this group is the conserved PxGxW motif found in loop 2. The presence of a conserved proline at the start of this motif prevents the formation of an alpha-helix in the loop 2. There is also strong conservation of acidic residues *N*-terminal of the first Zn(II) binding

residue. These residues have been shown to be involved in the recognition of unmodified H3K4 in Autoimmune Regulator (AIRE)[115] and WHSC1L1 (NSD3).[93]

Sub-Family 7

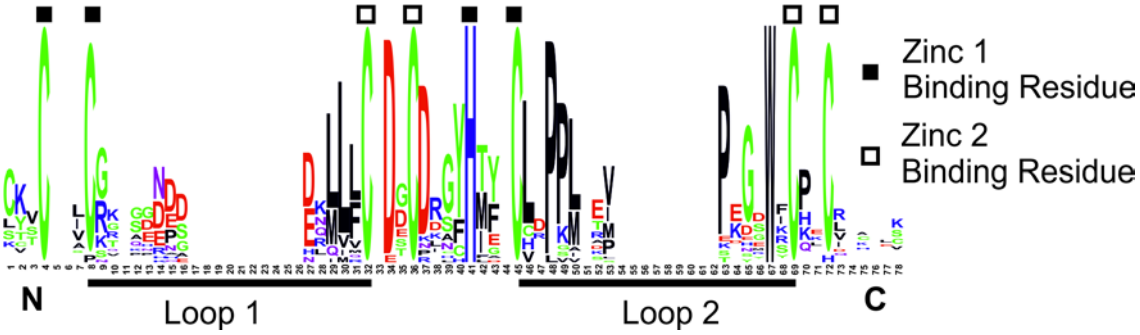


Figure 35. Sequence logo of sub-family 7. The Zn(II) binding residues as well as loop 1 and loop 2 are indicated.

Sub-family 7 is a large group containing 25 PHDs. As in sub-family 6, this group contains a well conserve proline in loop 2, preventing the formation of an alpha-helix in this region. This group does not containing residues capable of forming an aromatic cage, and would therefore not be expected to bind histone 3 peptides trimethylated at lysine 4; all available experimental data for this group backs up this hypothesis.[116–118]

This sub-family contains the second PHD of the set of tandem PHDs discussed in Chapter 4, as well as the second PHDs of the triple PHDs of Mixed-Lineage Leukaemia 2 (MLL2) and Mixed-Lineage Leukaemia 3 (MLL3). The distinction between tandem and triple PHDs is discussed in Chapter 4**.**

Sub-Family 8

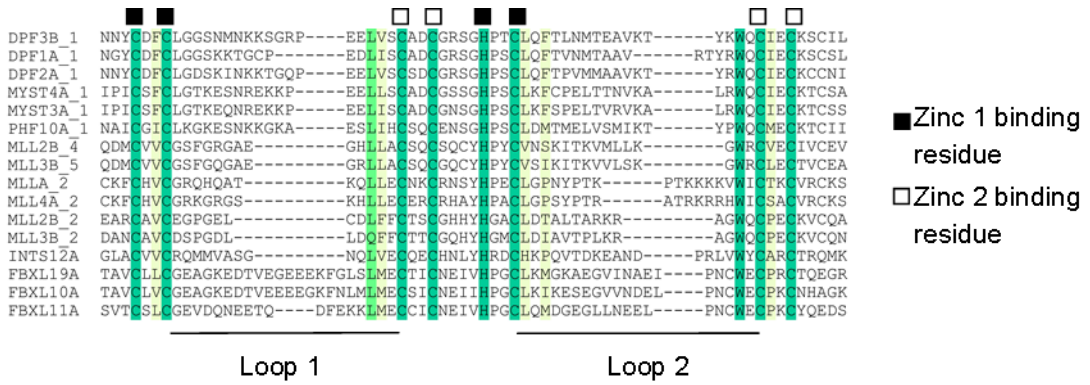


Figure 36. Sequence alignment of all members of sub-family 8. Zinc 1 binding residues are indicated with black squares, zinc 2 binding residues indicated with white squares.

This sub-family contains the first PHDs of the tandem and triple PHDs discussed above. This sub-family also contains the PHDs of FBXL lysine demethylase family (also known as the KDM2 family). The only conserved residue in this sub-family is a tryptophan residue two residues *N*-terminal of the seventh Zn(II) binding residue (Figure 35). This tryptophan is seen in PHDs across many different sub-families.

Sub-Family 9

This sub-family contains 9 PHDs, including the first PHD from each of the multiple PHD containing methyl transferases Mixed Lineage Leukaemia 1 (MLL1), Mixed Lineage Leukaemia 4 (MLL4), NSD1, WHSC1 (NSD2), and WSHC1L (NSD3). Both the MLL family and the NSD family contain a triple PHD (Chapter 4)**,** therefore it would be expected that the first PHDs of these triples would be closely related. However, there are no residues conserved across all of these PHDs (Figure 37).

This sub-family also contains the fourth PHD of NSD1 and WHSC1 (note that the putative fourth PHD of WHSC1L (NSD3) is missing its crucial Zn(II) binding residue, and is therefore not expected to be able to fold into a PHD structure) and the PHDs of Pygopus Family PHD Finger 1 (PYGO1) and Pygopus Family PHD Finger 2 (PYGO2). The PHD of PYGO1 is known to bind methylated H3K4, and therefore the PYGO family PHDs may be more correctly situated in sub-family 3. The

reason for their placement in subgroup 9 may be the preference for binding H3K4me2 over H3k4me3,[119] driven by the presence of an acidic residue on one side of the aromatic cage. This differentiates the PYGOs from other methylated H3K4 binders and may explain why they appear in a different sub-family.

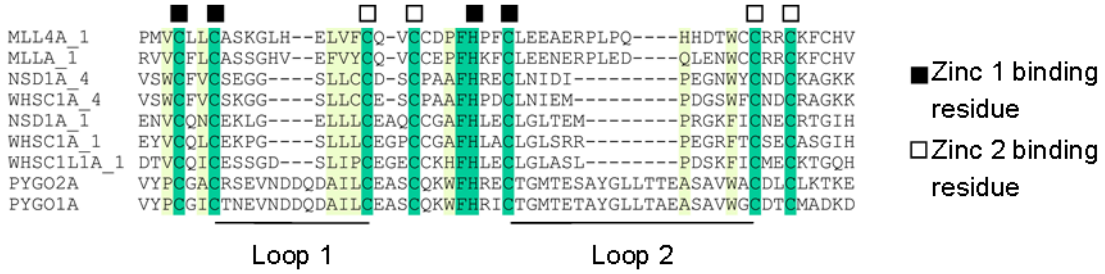


Figure 37. Sequence alignment of all members of sub-family 9. Zinc 1 binding residues are indicated with black squares, zinc 2 binding residues indicated with white squares.

Sub-Family 10

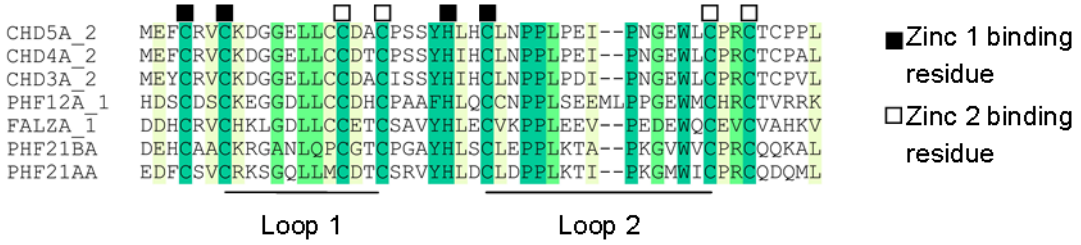


Figure 38. Sequence alignment of all members of sub-family 10. This sub-family contains strong conservation in loop 1, and a conserved PxGxW motif in loop 2.

Sub-family 10 contains seven PHDs with strong sequence conservation. Chromodomain Helicase DNA Binding Protein 4 (CHD4) is known to preferentially bind histone 3 peptides unmodified at H3K4 and trimethylated at H3K9. This interaction with H3K4 is mediated by an acidic residue two residues *N*-terminal of the first Zn(II) binding residue. This acidic residue is conserved throughout this sub-family, and it is therefore likely that all members of this sub-family will preferentially bind histone 3 peptides modified at H3K4. The interaction with H3K9 is mediated by a cation-pi interaction with a phenylalanine residue one residue *N*-terminal of the first Zn(II)

binding residue. An aromatic residue is found in this position in all PHDs of this sub-family except PHF12. Therefore it is possible that all members of this sub-family recognise H3K9me3.

Sub-Family 11

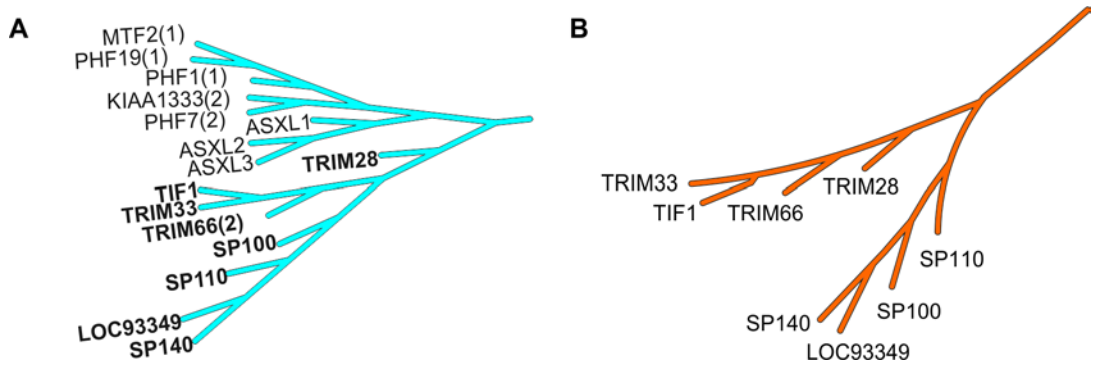


Figure 39. A comparison of the PHD and bromodomain phylogenetic trees for members of sub-family 11. **A.** Expanded view of PHD sub-family 11. The sub-family contains two distinct branches. The lower branch contains PHDs that are part of a PHD-bromodomain tandem (labels emboldened). **B.** Bromodomain sub-family V shares a similar topology to PHD sub-family 11. This indicates that these two domains evolved together as a single histone recognition module.

Sub-family 11 contains two distinct branches (Figure 39). The lower branch contains PHDs that belong to a PHD-bromodomain tandem motif. It is interesting to note that the topology of this branch closely matches the topology of the branch containing these proteins within the bromodomain family tree.[27] The PHDs and bromodomains of these proteins act as a single histone recognition module (Figure 40).[44] It is therefore likely that they evolved together, and it would therefore be expected that a phylogenetic analysis of the PHDs or bromodomains alone would produce a tree with the same topology.

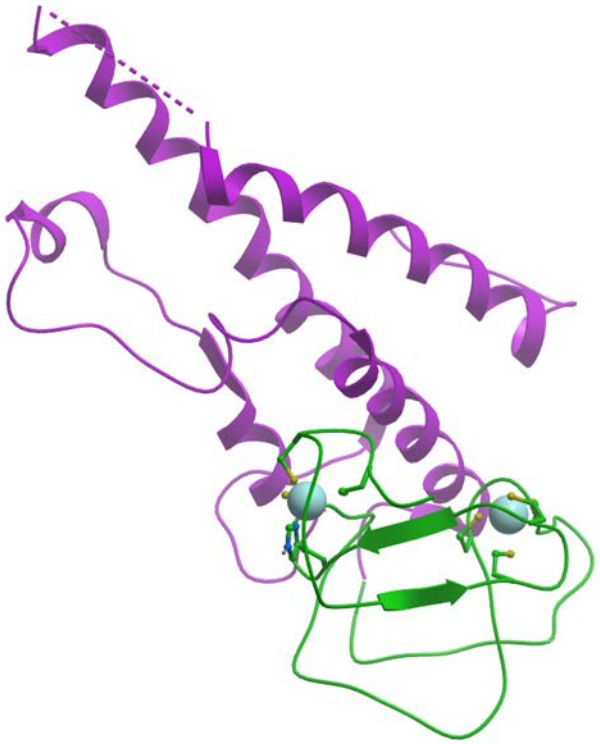


Figure 40. A PHD-Bromodomain tandem, with the *N*-terminal PHD shown in green and the *C*-terminal bromodomain shown in magenta. The PHD-Bromodomain shown is Tripartite Motif Containing 33 (TRIM33) PDB ID: 3U5M.

The upper branch of sub-family 11 contains the second PHDs of the closely related PHD Finger Protein 1 (PHF1), related PHD Finger Protein 19 (PHF19), and Metal Response Element Binding Transcription Factor 2 (MTF2). These three proteins are sometimes known as Polycomb Like 1 (PCL1), PCL2, and PCL3 and have a similar domain architecture containing two PHDs.

Sub-Family 12

Sub-family 12 is itself divided into two distinct branches. The smaller branch, containing PHD Finger Protein 20 (PHF20), PHD Finger Protein 20-like (PHF20L), Mixed Lineage Leukaemia 5 (MLL5), Absent Small or Homeotic-like 1 (ASH1L), and Recombination Activating Gene 2 (RAG2) contains the conserved tryptophan and methionine residues seen in sub-family 3. There is data available that show that MLL5 and RAG2 are H3K4me3 binders,[50,120] therefore based on this evidence and the presence of residues known to be involved in methylated lysine recognition, it

is likely that all five of these PHDs bind to H3K4me3. This predicted H3K4me3 binding specificity would suggest that this branch is closely related to members of sub-family 3.

The other branch of sub-family 12 contains PHDs known to form part of a PHD-Zn Knuckle-PHD motif. This motif has no known structure or function, but is well conserved within these 12 proteins (Figure 41).[121,122] This motif consists of two PHDs separated by two pairs of Zn(II) binding residues, with each pair separated by two amino acids. Within this group it is interesting to note that the Jumonji Domain 2 (JMJD2) family have a serine in place of a cysteine as the second Zn(II) binding residue of the PHD.

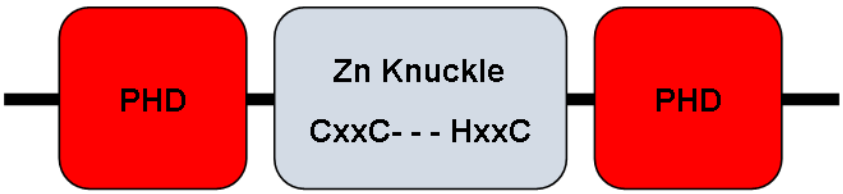


Figure 41. The PHD- Zn Knuckle-PHD motif. Members of sub-family 12 form the first PHD of this motif. The Zn knuckle section consist of four Zn(II) binding residues arranged in two pairs. Within each pair, the Zn(II) binding residues are separated by two amino acids.

Sub-Family 13

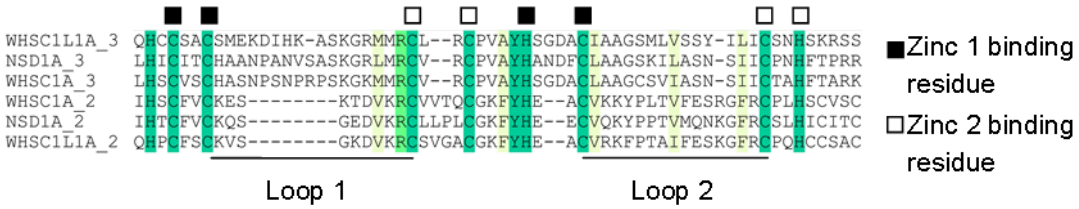


Figure 42. Sequence alignment of all members of sub-family 13. Zinc 1 binding residues are indicated with black squares, zinc 2 binding residues indicated with white squares.

Sub-family 13 contains six PHDs, all of which come from two closely related methyl-transferases NSD1, WHSC1, and WSHC1L (Figure 42). They have a conserved domain architecture, although

WSHC1L is missing the fourth PHD. No structural or functional information is available for this sub-family.

Sub-Family 14

Sub-family 14 contains twelve PHDs, all of which are the second PHD of a PHD-Zn Knuckle-PHD discussed above. This sub-family has a highly conserved hydrophobic core, with all twelve members of the group having a FHVTCA motif encompassing the fifth and sixth Zn(II) binding residues (Figure 43).

It has been suggested that the second PHD of Bromodomain Protein 1 (BRD1) does not have histone 3 binding activity but does bind DNA non-specifically.[122] Positively charged residues have been identified as being responsible for this functionality. There is partial conservation of these residues, but there is not enough evidence to predict with any certainty if this functionality is conserved within the sub-family.

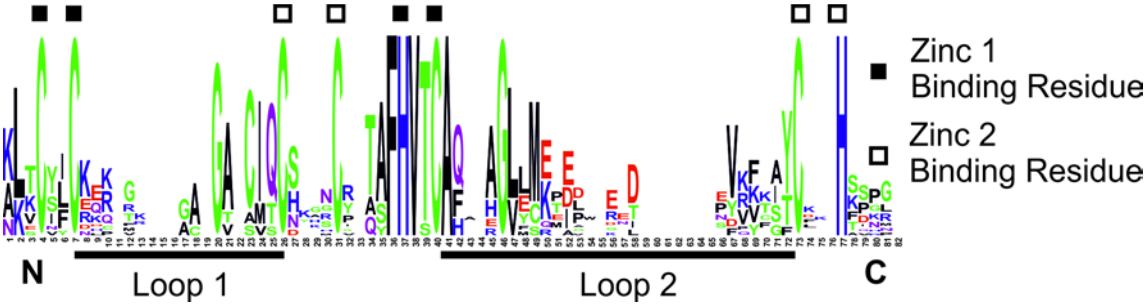


Figure 43. Sequence logo of sub-family 14. There is a conserved FHVTCA motif in the core of the PHDs in this sub-family.

Sub-Family Comparison

The above analysis shows that although the PHD family share a motif of Zn(II) binding residues, they have many differences in other parts of the sequence that leads to a diverse range of functions (Table 4). One of the key differences in determining histone binding specificity is the presence of either aromatic or acidic residues around the H3K4 binding site.

An example of a sequential feature that can be used to predict structure is the presence of a proline residue in loop 2. The presence of this a proline in this position plays an important role in determining the secondary structure of this region. Sub-families 6, 7, and 10 all share a conserved proline in loop 2 and it is therefore highly unlikely that PHDs from these sub-families will contain an alpha-helix in this region.

Sub-Family Name	Number of Members	Key Features
Sub-family 1	13	A highly conserved GA motif in loop 1. Not expected to be H3K4me3 binders.
Sub-family 2	12	A diverse group containing anomalous PHDs such as CREBBP and EP300 (large loop 1 insert) and ASH2L (only binds one Zn(II) but still has PHD fold.
Sub-family 3	23	H3K4me3 binding PHDs.
Sub-family 4	3	Conserved RV motif in central beta-sheet. Not expected to be H3K4me3 binders.
Sub-family 5	4	PHDs which form part of a larger ADD domain. Not H3K4me3 binders.
Sub-family 6	8	Contains conserved proline in loop 2 preventing alpha-helix formation. Not H3K4me3 binders
Sub-family 7	25	Contains conserved proline in loop 2 preventing alpha-helix formation. Not expected to be H3K4me3 binders.
Sub-family 8	16	Contains some PHDs that are the first of a tandem PHD (Chapter 4). Not H3K4me3 binders
Sub-family 9	9	Contains some PHDs that are the first in a triple PHD. With the exception of PYGO1 and PYGO2 not expected to be H3K4me3 binders.
Sub-family 10	7	Possible H3K9me3 binders.
Sub-family 11	16	Contains PHDs that are part of a PHD-bromodomain tandem.
Sub-family 12	17	Contains some H3K4me3 binders and some PHDs which are the first PHD in a PHD-Zn Knuckle-PHD motif.
Sub-family 13	6	All members come from a related family of methyl transferases. Not expected to bind H3K4me3.
Sub-family 14	12	All PHDs are the second PHD of a PHD-Zn Knuckle-PHD motif.

Table 4. Summary of 14 identified PHD sub-families showing the number of members of each sub-family and describing key features of each sub-family.

This analysis shows that it is possible to define a PHD family with well conserved similarities such as the Zn(II) binding motif. However, it is also possible to break this overall family down into sub-families which contain sequential features which make the groups distinct from each other. These features allow for a diverse range of functions within a family that still maintains fundamental similarities.

Disease Associated PHD Mutations

Disease associated mutations of PHDs are discussed in (Chapter 1). The majority of these disease associated mutations occur in the Zn(II) binding residues. For examples that do not occur in Zn(II) binding residues it is interesting to studying the conservation of these residues within closely related PHDs.

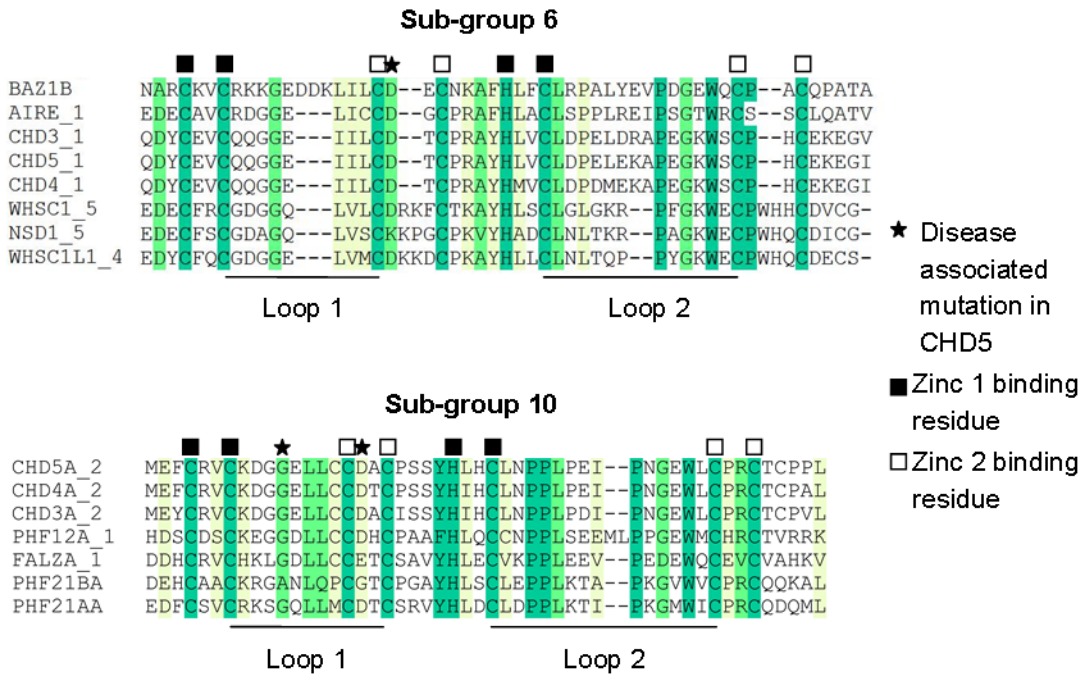


Figure 44. Sequence alignment of sub-family 6 and sub-family 10. Positions where disease associated mutants of CHD5 are found are highlighted with stars.

The only example that has been identified is of CHD5, which has disease associated mutations in both its first (G355A, D361A) and second (D434A) PHDs. These mutations have been shown to abrogate binding to unmodified histone 3.[63] There is strong conservation of these residues

within the sub-family (Figure 44). This gives further evolutionary evidence to the importance of these residues for the correct function of the PHDs within this sub-family.

Summary

This chapter addresses the lack of a family wide analysis of human PHDs. This work provides the first definitive list of human PHDs, with a full description of the methodology used to construct the list. The full set of human PHD sequences were aligned using structural information, and a phylogenetic tree constructed using a robust maximum likelihood method.

This analysis revealed that PHDs can be grouped in to fourteen sub-families. Each sub-family has been analysed and features that can be used to predict histone binding selectivity have been identified. This grouping of PHDs into sub-families is also likely to be of great importance to future medicinal chemistry efforts to design selective PHD inhibitors. As it will allow the creation of selectivity panels that effectively sample the PHD family. It will also guide structure based drug discovery, as medicinal chemists will be able to identify conserved residues from non-conserved residues when analysing the binding site of a small molecule inhibitor. Designing molecules to interact with non-conserved residues is likely to deliver more selective inhibitors.

A key conclusion of this work is that the majority of PHDs are not predicted to be H3K4me3 binders. Almost all PHDs with the required residues to form an aromatic cage to bind H3K4me3 are found in a single sub-family. This challenges that common assumption that all PHDs are methyl-lysine reader domains.

It is hoped that this work will motivate further research into the function and biological roles of PHDs. In particular, this work discusses the similarities between PHDs found in proteins with multiple PHDs and similar domain architecture. This evolutionary conservation of these multiple PHDs points to an important biological role that demands further study.

Chapter 3 - Computational Assessment of the Ligandability of PHDs and Other Epigenetic Reader Domains

The human PHD family is large, with 173 PHDs identified in the human proteome (Chapter 2). The size of the PHD family precludes using purely experimental means in order to compare the ligandability of PHDs. Therefore a higher throughput, computational method is required. This chapter describes the use of SiteMap to analyse all available PHD structures to identify which PHDs are more likely to be amenable to inhibition by small molecule ligands. A similar study has been conducted on methyl-lysine binding domains which included some PHD structures.[78] The work described in this chapter will expand on this previous research by studying all available PHD structures. The results of this study will be used to prioritise ligandable PHDs for screening and inhibitor development. This chapter will then go on to use similar analysis techniques to assess the ligandability of Tudor domains.

A further set of structures that contain multiple domains involved in reading, writing, and erasing post-translational histone modifications were also analysed by SiteMap in order to identify potential small-molecule binding sites at domain-domain interfaces. A subsequent crystal soaking assay for the PHD-bromodomain of SP100 developed and performed by colleagues within the Structural Genomics Consortium identified a series of ligands that bind at such an inter-domain site identified by SiteMap. This provides evidence of the potential to develop small molecule ligands that bind at domain-domain interfaces in multi-domain proteins.

Computational Methods for Assessing Ligandability

There are various software applications for identifying ligand binding sites. These fall into two broad categories; those that use a purely geometric based approach such as FPocket,[123] and those that also take into account physiochemical properties of the binding site. Prominent

examples of programmes that identify sites using the physiochemical properties of the binding site include FTMap[124], ICM PocketFinder,[125] and Schrödinger SiteMap.[126]

FTMap uses a series of 16 small, organic molecule probes which are given energetic binding scores for each of 500 rotations at a wide ranging number of positions on the protein surface. The 2000 lowest energy conformations for each probe are retained and clustered. Regions containing several probe clusters are reported as potential ligand binding sites. Pocket Finder identifies pockets by calculating the Lennard-Jones potential between an aliphatic carbon probe and the surface of the protein. SiteMap uses a similar approach to PocketFinder, in that it uses a water molecule probe to calculate van der Waal's interaction with the protein surface. SiteMap also includes a geometrical test when identifying binding sites. The full details of the SiteMap algorithm are discussed below.

SiteMap

Of the available software, SiteMap was chosen for this work. A comparison between SiteMap and FPocket has shown that they perform very similarly.[127] SiteMap has the advantage over PocketFinder and FTMap in that it would facilitate direct comparison with similar work undertaken with SiteMap to study bromodomains,[128] methyl transferases,[129] and methyl-lysine binding domains.[78]

A brief description of the SiteMap algorithm is provided below, and summarised in Figure 45. SiteMap initially creates a grid of points separated by 1.0 Å. Points within the van der Waal's radius of nearby protein atoms are deemed to be 'inside' the protein and are discarded. For each remaining point the van der Waal's interaction with the protein surface and the 'enclosure' score are calculated. The van der Waal's interaction is calculated using a water molecule probe. The enclosure score is calculated by determining what fraction of 'rays' emitted in all possible directions from the site point strike the surface within a given distance, which by default is 8 Å.

By default only points that have a van der Waal's interaction stronger than -1.1 kcal/mol and an enclosure score greater than 0.5 are retained. These retained points are known as site points.

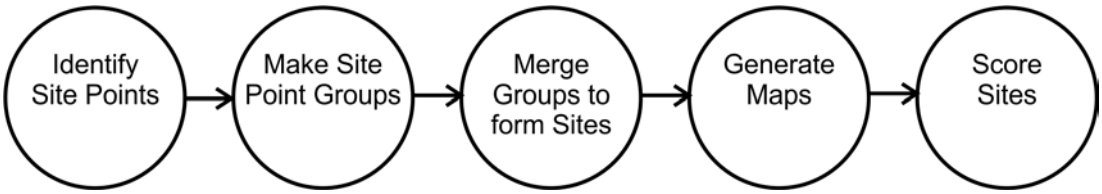


Figure 45. A schematic representation of the SiteMap algorithm used for identifying and scoring potential small molecule ligand binding sites on a protein surface.

The next stage is to combine site points into groups. A site point is added to a group if is within a given distance of a minimum number of group members. By default a site point is added if there are 3 group members within a squared distance of less than 3.1 $Å^2$. Site points that do not meet the criteria for joining a site point group are discarded. Groups within a certain distance (default: 6.5 Å) are merged, the resultant merged group constitute the identified potential ligand binding site.

SiteMap then generates a set of maps that define the site, these five maps are: hydrophilic, hydrophobic, H-bond donor, H-bond acceptor, and surface. SiteMap then calculates the SiteScore and DScore for the sites.[130] SiteScore scores how likely a site is to tightly bind a small molecule ligand; whereas DScore scores how likely a site is to tightly bind a drug-like small molecule ligand. For example, a site which tightly binds a highly charge ligand would have a high SiteScore, but low DScore. These scores, along with the maps, allow sites to be evaluated and compared. The SiteScore and DScore values are calculated using three of the parameters that define the site: number of site points, enclosure, and hydrophilicity (Equation 1 and Equation 2). Note that the hydrophilicity term is negative to penalise sites for being hydrophilic. DScore includes a greater penalty for a site being hydrophilic, as such a hydrophilic site may be ligandable, but not druggable, due to the likely polar nature of any ligand that would bind such a site.

$$SiteScore = 0.0733\sqrt{n} + 0.6688e - 0.2000p$$

Equation 1. The SiteScore of a given site is calculated from the number of site points, *n* (capped at 100); the enclosure score, *e*; and the hydrophilicity score, *p* (capped at 1.0).

$$DScore = 0.094\sqrt{n} + 0.60e - 0.324p$$

Equation 2. The DScore of a given site is calculated from the number of site points, *n* (capped at 100); the enclosure score, *e*; and the hydrophilicity score, *p*.

SiteMap Analysis of PHDs

A key aim of this work was to identifying which PHDs are most amenable to inhibition by small molecules in order to prioritise future screening and chemical probe discovery efforts. In order to achieve this, SiteMap analysis was performed on all available PHD structures. In cases where a PHD structure was solved by NMR, SiteMap analysis was performed on all models within the PDB deposition. This is in keeping with the precedent set by Vidler et al. on their analysis of the bromodomain family.[128]

	NMR	X-ray	Total
Apo	762 (24)	17 (13)	779 (32)
Holo	187 (9)	43 (19)	230 (24)
Total	949 (25)	60 (22)	1009 (37)

Table 5. Breakdown of PHD structure set used in SiteMap analysis. Structures defined as *holo* were solved in the presence of a histone peptide ligand; structures defined as *apo* were solved without a bound ligand. The number of unique genes in each set is shown in parentheses. No member of the set was solved with a small molecule ligand.

A structure set was compiled consisting of 103 PDB depositions, representing 37 unique PHDs (Appendix 3.1). PHD structures solved by NMR were separated into individual models and structures containing multiple PHDs had been split into individual PHDs, giving a final set of 1009 structures.

Protein Preparation

Prior to analysis by SiteMap, all structures were put through the same structure preparation procedure. Structures were downloaded and superimposed using the ICM superimposition tool using the sequence alignment described in Chapter 2 as a guide. This created a PDB format file for every structure saved in the same frame of reference. The structures were then re-downloaded using PyMol; extra chains, extra domains, and water molecules removed; and superimposed with the identical structure created using ICM. This allowed use of ICM's superior superimposition tool, while negating an intrinsic incompatibility between PDB files created by ICM and Schrödinger.

Schrödinger's Protein Preparation Wizard[131] was then used to add hydrogens to the structure where appropriate; adjust the protonation state of acidic and basic residues for pH 7.4; and refine the structure to relieve strain and optimise the hydrogen bonding network. Finally the bound peptide ligand was deleted from all structures with a bound ligand (*holo* structures).

Running Site Map

SiteMap was run on all structures using variations on the default parameters. The SiteMap manual recommends that for the detection of shallow sites, the thresholds for the initial identification of site points should be modified. Specifically, the threshold for the van der Waal's interaction required for a point to be included to be reduced from -1.1 kcal/mol, and threshold enclosure score reduced from 0.5. As no guidance is given on the degree to which these two parameters should be reduced, two different reduced values were tested for each parameter alongside the default parameters. To this end, SiteMap analysis was performed on each structure five times, each with a variation in the thresholds for the initial identification of site points (Table 6).

Parameter Set	van der Waal's Interaction Threshold (kcal/mol)	Enclosure Score Threshold	Discovery Rate
1 (default)	-1.1	0.5	40%
2	-1.0	0.45	73%
3	-1.0	0.40	84%
4	-0.90	0.45	81%
5	-0.90	0.40	90%

Table 6. The SiteMap analysis was performed using five parameter sets for the initial identification of site points. These parameters were varied in order to increase the chances of identifying a potential small-molecule binding pocket at the shallow histone binding site of PHDs. A discovery was defined as any structure where SiteMap identified a potential small molecule binding site on the histone 3 binding surface of the PHD.

Identifying Potential Small Molecule Binding Sites at the Histone Binding Face

Running five SiteMap analyses on each of 1009 structures created a large volume of data. Therefore a method to automate the procedure for identifying whether SiteMap had found a potential small molecule binding site or not was required. It was also necessary to automatically separate sites found at the histone binding face from those identified on other surfaces of the PHD.

To accomplish this the centroid of each identified site was calculated by taking the mean of the *x*, *y*, and *z* coordinates of each site point within the site to give a single (*x, y*, *z*) coordinate for each site identified. As all structures had been superimposed prior to SiteMap analysis it was possible to overlay the centroids of the identified sites onto a representative structure, and quickly separate sites on the histone binding face from other identified sites. Structures where SiteMap found a potential small molecule binding site on the histone binding surface were described as 'discovered sites', allowing the calculation of discovery rates for each parameter set used (Table 6).

As expected, the discovery rate increases as the thresholds for van der Waal's interaction and enclosure are reduced. Parameter set 5 has the lowest thresholds and hence has the highest discovery rate (90%). The low discovery rate for Parameter Set 1 suggests that the PHD family is likely to be a difficult target class for small molecule inhibition.

Comparing Results of Different Parameter Sets

As the aim of this work was to compare and contrast the potential for small molecule binding at the histone 3 binding surface across the PHD family, it was necessary to determine which parameter set provided the most suitable data in order to make this comparison.

As expected the discovery rate was highest for Parameter Set 5, as this had the least stringent criteria for the initially classification of site points. Parameter Set 5 therefore subsequently characterised a greater number of potential binding sites at the histone 3 binding surface than the other parameter sets. It was therefore decided to use Parameter Set 5 when making comparisons within the PHD family; Parameter Set 1 (the default parameters) was used for comparing PHDs to other domains.

In the SiteMap analysis of bromodomains conducted by Vidler et al, the default parameters used identify sites in twenty-three of the twenty-four bromodomains studied (a 96% discovery rate on a per gene basis, this differs from the per structure discovery rates reported in this work)[128] Comparing the low discovery rate for Parameter Set 1 with that of the bromodomain study suggests that the PHD family is likely to be a difficult target class for small molecule inhibition.

Analysing Results

As described in Table 5, the PHD structure set contained a mixture of structures solved by NMR spectroscopy and by X-ray crystallography. The structure set also contained both *apo* and *holo* structures. Before analysing the results of SiteMap study, it is important to rule out any bias introduced by the method used to solve the structures.

Comparison of NMR and X-ray Structures

Of the thirty-seven PHDs of which there are structures, there are only nine for which both NMR and X-ray structures are available (Table 7). The results for these nine PHDs can be used to test whether there is any intrinsic difference in ligandability between PHDs that have had their structure solved by NMR or X-ray.

Gene Name	Median DScore	Median NMR DScore	Median X-ray DScore	More Druggable Structure
FALZ(2)	0.69	0.68	0.91	X-ray
ING2	0.56	0.58	0.86	X-ray
ING4	0.66	0.66	0.90	X-ray
MLL(3)	0.97	0.96	1.03	X-ray
ATRX	0.89	0.89	0.67	NMR
PHF21A	0.99	0.99	0.74	NMR
RAG2	1.00	1.01	1.01	NMR
TRIM28	0.53	0.56	0.46	NMR
UHRF1	0.89	0.90	0.86	NMR

Table 7. PHDs for which there is both an NMR and X-ray structure available. These structures were used to investigate whether the method used to solve the structure affects the calculated DScore. As multiple structures are present for each PHD, median DScores are used for comparison.

Of the nine PHDs where both an NMR and X-ray structure were available, five appeared more ligandable in the NMR structure, and four more ligandable in the X-ray structure. This suggests that there is unlikely to be any intrinsic difference between PHD structures solved by NMR and X-ray, and therefore these NMR and X-ray structures can be directly compared in terms of ligandability.

Comparison of Apo and Holo Structures

Gene Name	Median DScore	Median *Apo* DScore	Median *Holo* DScore	More Druggable Structure
CHD4(2)	0.60	0.73	0.54	*Apo*
JARID1A(3)	0.93	0.95	0.61	*Apo*
PHF21A	0.99	0.99	0.74	*Apo*
PYGO1	0.97	0.97	0.96	*Apo*
RAG2	1.01	1.01	1.01	*Apo*
TIF1A	0.51	0.53	0.49	*Apo*
TRIM33	0.51	0.59	0.48	*Apo*
AIRE	0.60	0.60	0.60	*Holo*
ATRX	0.89	0.82	0.93	*Holo*
DNMT3A	0.74	0.59	0.90	*Holo*
DNMT3L03	0.83	0.78	0.87	*Holo*
FALZ(2)	0.69	0.54	0.89	*Holo*
ING2	0.56	0.56	0.86	*Holo*
ING4	0.66	0.66	0.90	*Holo*
MLL(3)	0.97	0.97	1.03	*Holo*
MYST3(1)	0.80	0.79	0.94	*Holo*
MYST3(2)	0.91	0.90	0.95	*Holo*
PHF13	0.67	0.64	0.71	*Holo*
TAF3	0.93	0.87	0.98	*Holo*
UHRF1	0.89	0.89	0.89	*Holo*

Table 8. PHDs for which there is both an *apo* and *holo* structure available. These structures were used to investigate whether the presence of a bound peptide ligand affects the calculated DScore.

There are twenty PHDs whose structures have been solved with and without a bound peptide ligand (Table 8). It would be expected that *holo* structures would appear ligandable, as the bound peptide would force the PHD into a conformation with a well-defined binding interface. Comparing the results of the SiteMap analysis of the twenty PHDs for which both *apo* and *holo* structures are available would appear to give some weight to this argument. Of the twenty PHDs

thirteen were more ligandable in their *holo* form. By looking at the SiteMap parameters that are used to calculate DScore and SiteScore, it can be seen that the difference between *apo* and *holo* structures is driven by the differing sizes of the binding site (Table 9). This observation is in line with the expectation that the presence of a bound peptide ligand would force the PHD to adopt a conformation with a more well defined binding site.

Table 9. The median values for the SiteMap parameters used to calculate DScore and SiteScore are shown for both the *apo* and *holo* versions of the PHDs listed in Table 8.

Identifying Ligandable PHDs

The above comparisons of NMR and X-ray structures as well as *apo* and *holo* structures suggests that it is valid to compare NMR and X-ray structures, but it is less valid to compare *apo* and *holo* structures. Therefore while attempting to identify the most ligandable PHDs, two rankings have been complied; one ranking for *holo* PHDs (Figure 46), and another for *apo* PHDs (Figure 47).

The PHD of Recombination Activating Gene 2 (RAG2) is ranked as the most ligandable in the *apo* ranking, and second most ligandable in the *holo* ranking. This would indicate that RAG2 is one of the most ligandable PHDs. However, when analysed using Parameter Set 1 (the default parameters) the median DScore for *apo* RAG2 structures (20 NMR models, PDB ID: 2JWO) is 0.69, which would previously be considered as unligandable.

Another PHD that scores highly on both rankings is that of Pygopus 1 (PYGO1), ranked as third on the *apo* ranking, and fourth on the *holo* ranking. Despite this high ranking, Parameter Set 1 results would suggest that this is also a PHD with low ligandability, with a maximum DScore of 0.78 for the six *holo* structures analysed. However, since the completion of this work, the PHD of

PYGO1 has become the first PHD for which a strong evidence of a small molecule inhibitor has been publsished.[74] This shows that although isolated PHDs are likely to be difficult targets, it is possible to identify small molecule inhibitors. The ranking of PHDs with known structures established by this work can be used to prioritise screening efforts, and ensure that resources are directing at identifying ligands for the 'least-difficult' PHDs.

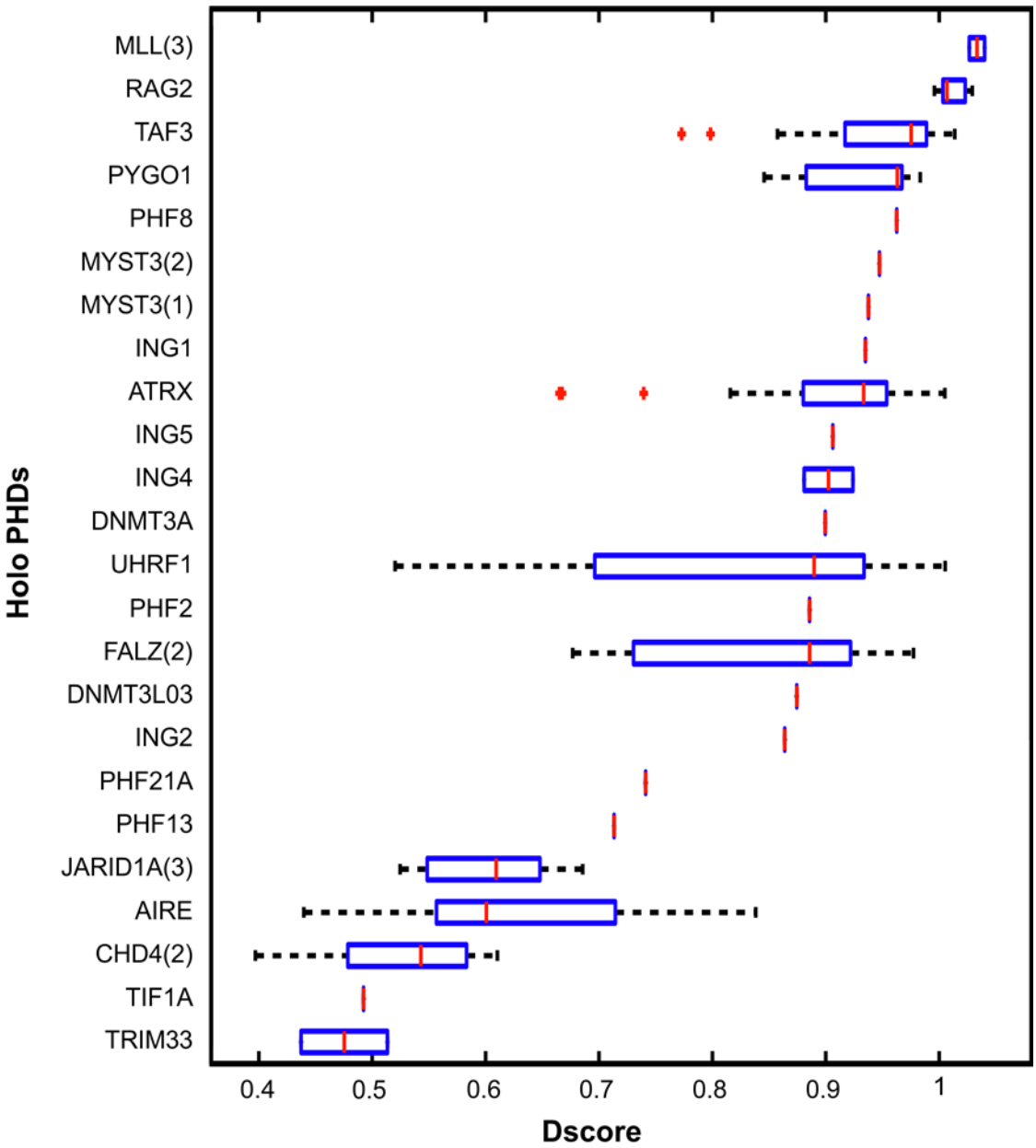


Figure 46. Box plots showing the range of DScores for each *holo* PHD, calculated using Parameter Set 5. The box plots are marked with the median value and the edges of the box represent the inter-quartile range. The whiskers extend to the most extreme data not considered an outlier, and outliers are plotted individually. An outlier is classed as any data point more than 1.5 inter-quartile ranges below the first quartile or above the third quartile.

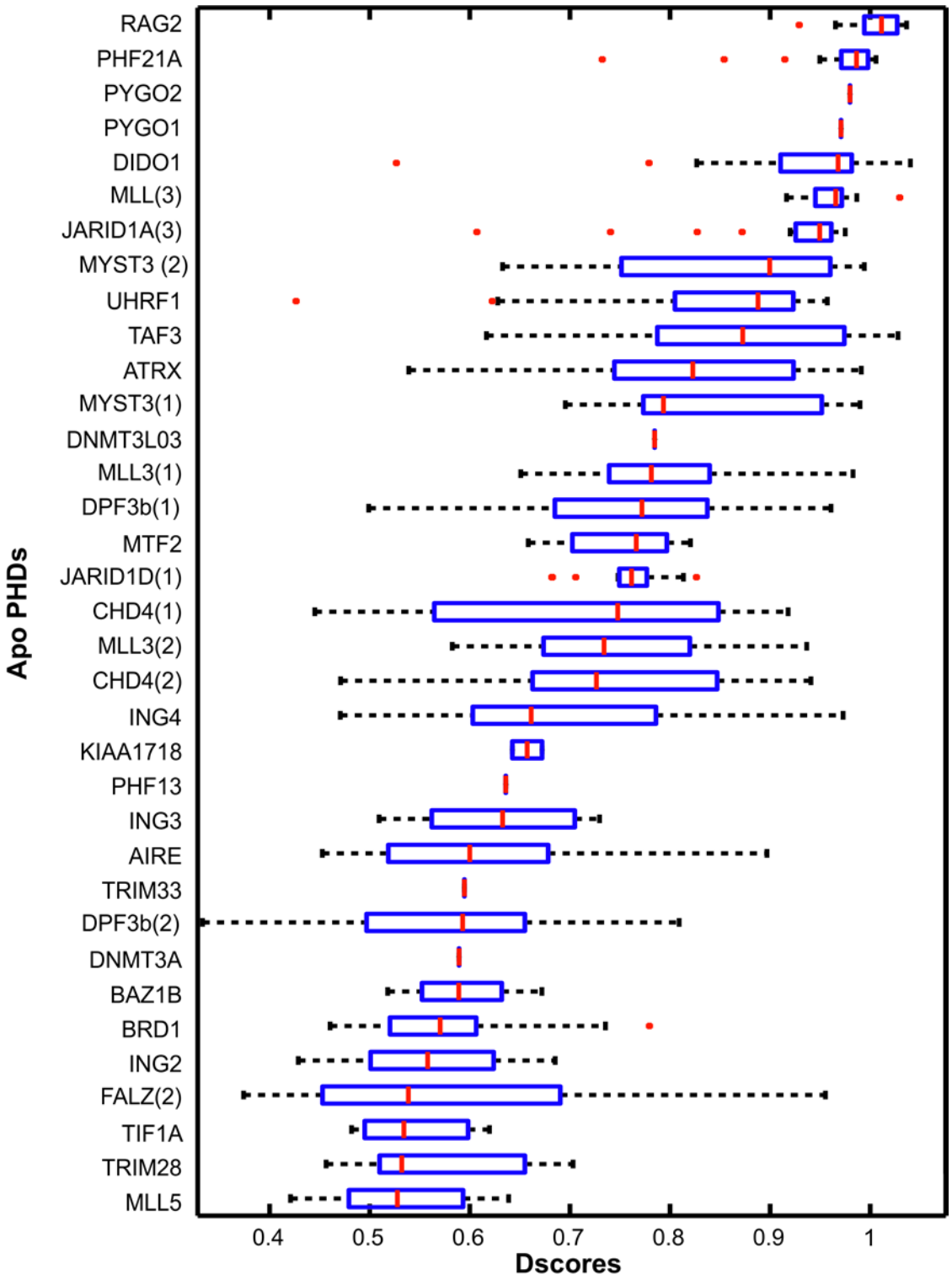


Figure 47. Box plots showing the range of DScores for each *apo* PHD, calculated using Parameter Set 5. The plots are sorted by the median value with the most ligandable number at the top. The box plots are marked with the median value and the edges of the box represent the inter-quartile range. The whiskers extend to the most extreme data not considered an outlier, and outliers are plotted individually. An outlier is classed as any data point more than 1.5 inter-quartile ranges below the first quartile or above the third quartile.

Comparison of Single PHDs with Tandem PHDs

The PHD structure set contained many examples of structures containing a PHD and other domains, including some examples of structures containing two adjacent PHDs. To investigate whether these tandem PHDs represented more ligandable targets that isolated PHDs, SiteMap was run on a set of 121 structures of tandem PHD-PHD domains. This structure set represents three genes covered by seven Protein Data Bank entries (Table 10).

Gene Name	PDB ID	Method Used to Solve Structure.
DPF3b	2KWJ	NMR
	2KWK	
	2KWN	
	2KWO	
MYST3 (KAT6A)	2LN0	NMR
	3V43	X-ray Crystallography
MLL3[e]	2YSM	NMR

Table 10. A list of tandem PHDs with known structures. These were analysed with SiteMap and compared to the se of single PHDs. These tandem PHDs were included in the study of isolated PHDs, with the domains separated prior to analysis by SiteMap.

Double PHD Finger Protein 3b (DPF3b) and MYST3 are known to engage histone tails via an extended interface involving both PHDs.[132,133] SiteMap identified this novel histone 3 binding site as a potential ligand binding site in all but four structures, a much lower failure rate than for single PHDs.

The enclosure, hydrophilicity, and number of site points as calculated by SiteMap of single and tandem PHDs were compared (Figure 48). These three parameters were chosen as they are the

[e] The second and third PHDs of MLL3 are considered as a tandem PHD for this analysis, as they share the same domain orientation as DPF3b and MYST3. However, these PHDs are not included as tandem PHDs as defined in Chapter 4, this is because these PHDs are part of a wider quadruple PHD including the first and fourth PHDs of MLL3.

three parameters used to define DScore and SiteScore, which are SiteMap's built in scoring functions.[130] Figure 48 shows the values for the three chosen parameters normalised against the mean values of the submicromolar sites used to train DScore and SiteScore. This data shows that both single and tandem domains have much smaller potential ligand binding sites than the submicromolar set. The median number of site points for single and tandem PHDs are twenty-five and fifty-two respectively; the mean number of site points of the submicromolar sites is 132.

	Mean Number of Site Points	Mean Enclosure Score	Mean Hydrophilicity Score
Single PHDs	25	0.56	1.06
Tandem PHDs	52	0.61	0.92
Submicromolar Set	134	0.76	1.0

Table 11. Comparison of single PHDs, tandem PHDs, and the set of structures with known submicromolar ligands used to train the Site Map scoring functions. The parameters used are those used to calculate SiteScore and DScore. Note that sites with lower hydrophilicity scores are considered more ligandable (Equation 1 and Equation 2).

Therefore it appears that tandem PHDs have a larger potential ligand binding site than single PHDs, with the highest scoring tandem PHDs approaching the mean score of a submicromolar site. Comparison of the enclosure score showed that both single and tandem PHDs (median enclosure of 0.56 and 0.61 respectively) have a less enclosed binding site than the submicromolar set (mean 0.76). DScore and SiteScore impose a penalty for a binding site being too hydrophilic, with the hydrophilic score being computed so that the average score of the submicromolar binding set is 1.0. The median hydrophilic scores of single and tandem PHDs are 1.06 and 0.92 respectively, showing that although their shape and size suggest they should be considered less ligandable targets, their hydrophilicity does not.

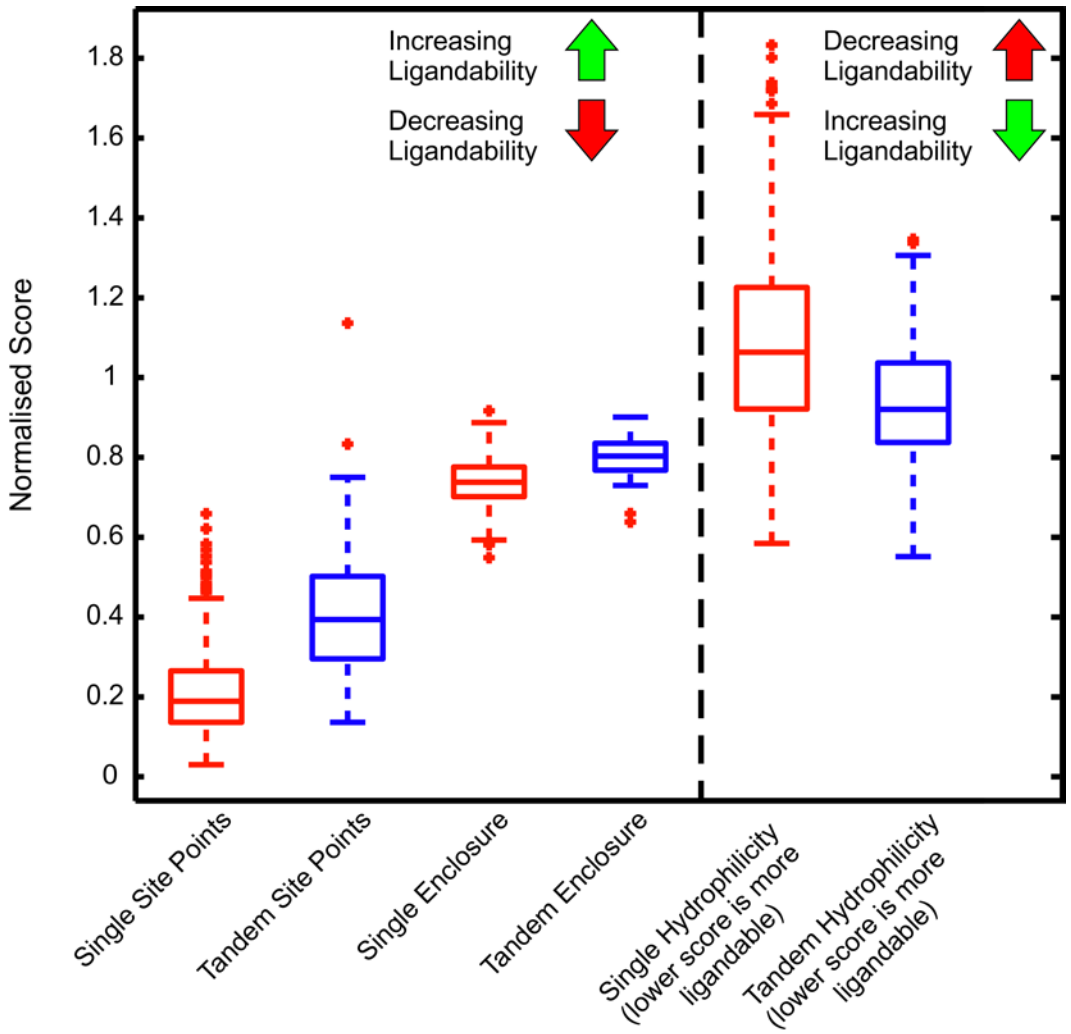


Figure 48. Box plots showing the number of site points, the enclosure, and the hydrophilicity score for tandem and single PHDs. All values have been normalised with respect to the mean values for these parameters derived from 342 sites with submicromolar inhibitors used to train DScore and SiteScore. The box plots are marked with the median value and the edges of the box represent the inter-quartile range. The whiskers extend to the most extreme data not considered an outlier, and outliers are plotted individually. An outlier is classed as any data point more than 1.5 inter quartile ranges below the first quartile or above the third quartile.

Based on these results we decided to focus our initial experimental investigation on tandem PHDs, as SiteMap suggests these are likely to be more ligandable than single PHDs, due to their larger potential ligand binding sites. There are six human tandem PHDs, found in DPF1, DPF2, DPF3, PHF10, MYST3, and MYST4. One of the tandem structures used in the study above was from MLL3; however, inspection of the protein sequence suggests that this is likely to form part of a quadruple PHD. The distinction between tandem PHDs and triple and quadruple PHDs, and

a detailed description of experimental efforts to identify a small molecule ligand of a tandem PHD are described in Chapter 4.

SiteMap Analysis of Tudor Domains

The above analysis of PHD domains suggests that tandem PHDs are likely to be more ligandable targets than isolated PHDs. Tudor domains are another family of epigenetic reader domains involved in histone tail recognition. Specifically they are known for binding to methylated lysine.[134] Similar to PHDs, Tudor domains are known to appear as both single and tandem domains. A structure set of Tudor domains was analysed with SiteMap in order to determine whether the more ligandable nature of tandem domain over single domains identified in PHD is seen in other epigenetic reader domain families.

As is the case for PHDs, a small set of Tudor domains had been analysed using SiteMap by Santiago et al.[78] Their analysis investigated five Tudor domains, and concluded that only the Tudor of Tumour Protein p53 Binding Protein 1 (TP53BP1) was ligandable. The work discussed in this section expands on this previous work by surveying all available Tudor domain structures, and provides a detailed discussion of the factors that make TP53BP1 a ligandable Tudor domain.

There are currently no known reported Tudor domain inhibitors, therefore it is hoped that this study will help prioritise more ligandable Tudors for screening efforts aimed at the discovery of small molecule ligands.

Single Tudor Domains

Forty-nine Protein Data Bank entries were identified that contained at least one Tudor domain (Appendix 3.2). In all of these structures at least one of the Tudor domains present had residues capable of forming an aromatic cage for methyl-lysine recognition.

In general these aromatic cages offered well enclosed, hydrophobic regions which could be used to anchor an inhibitor. However, these enclosed regions are typically small, with typically less

than 40 site points. Therefore these sites tend to have low SiteScores and are therefore likely to be difficult targets for small molecule inhibition when taken as single domains, but may be more ligandable when considered as tandem Tudor domains.

Multiple Tudor Domains

The structure set contained seven examples of multi-Tudor domain structures. One example, Ubiquitin-like with PHD and Ring Finger Domains 1 (UHRF1), containing a tandem Tudor and a PHD. In most of these cases, only one of the multiple tudor domains contained the aromatic cage that is typically found in tudor domains and plays a crucial role in binding to methylated lysines.[134] As discussed above, the aromatic cage found in many Tudor domains for methyl-lysine recognition provides a well enclosed hydrophobic region which appears to be a suitable small molecule binding site. However, analysis by SiteMap reveals these sites are too small to be considered ligandable. In the case of tandem Tudor domains, the second domain provides some expansion space around the aromatic cage. This creates a site where an anchoring head group could bind to the aromatic cage, with the rest of the molecule interacting with the expanded surface created by the second Tudor domain. It may be possible to develop methyl-lysine mimetics to exploit these aromatic cages, in a similar way to how acetyl lysine mimetics have been used to produce bromodomain inhibitors. Three examples of multi-Tudor domain containing proteins with potentially ligandable sites taking advantage of more than one domain are discussed below.

TP53BP1

The tandem Tudor domain of DNA repair factor Tumour Protein p53 Binding Protein 1 (TP53BP1) is an illustrative example of a case described above, taking advantage of the aromatic cage and an expanded surface created by the presence of a second Tudor domain. Analysis by SiteMap identifies a site with a SiteScore 0.98 containing the aromatic cage of N-terminal Tudor and extending to an area at the interface of the two Tudor domains (Figure 49). This site is large and

open, with size and enclosure scores of 98 and 0.61 respectively. Closer inspection suggests that 11 of these site points are situated in an opening that it is too far from the other points to be bridged by a single small molecule. However, based on a size score of 87 and assuming the enclosure and hydrophilicity scores remain unchanged, the site would still have a promising SiteScore of 0.92.

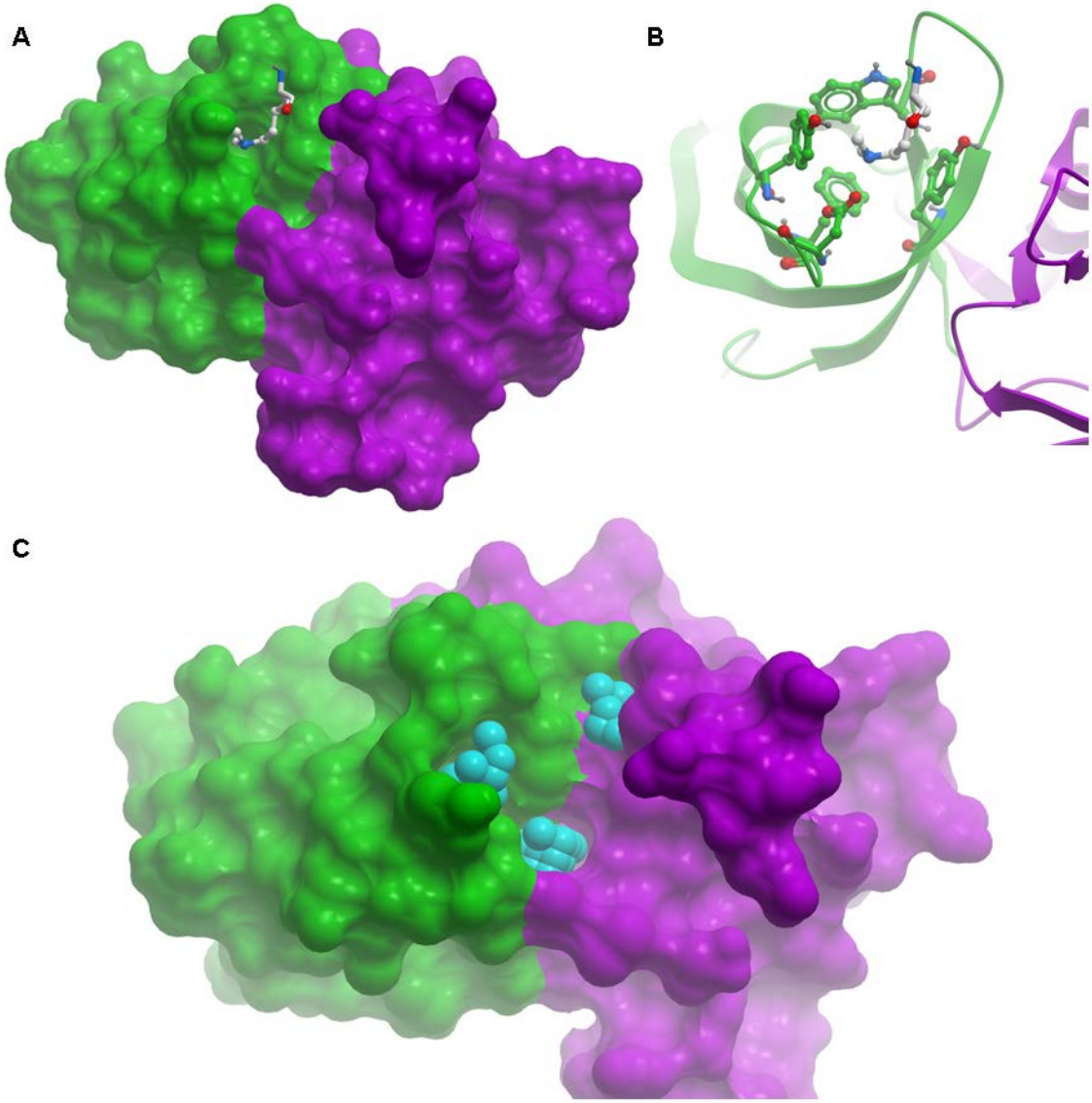


Figure 49. A. H3K4me2 in the binding site of the *N*-terminal Tudor (green) of TP53BP1, near the interface with the *C*-terminal Tudor domain (magenta). **B.** Dimethyl lysine binding in the aromatic cage of the *N*-terminal Tudor. **C.** The site points (cyan) extend from the methyl lysine binding site to include an area at the interface of the two domains. This site has a SiteScore of 0.98 and is therefore likely to be amenable to inhibition by a small molecule. PDB ID: 3LGL.

SETDB1

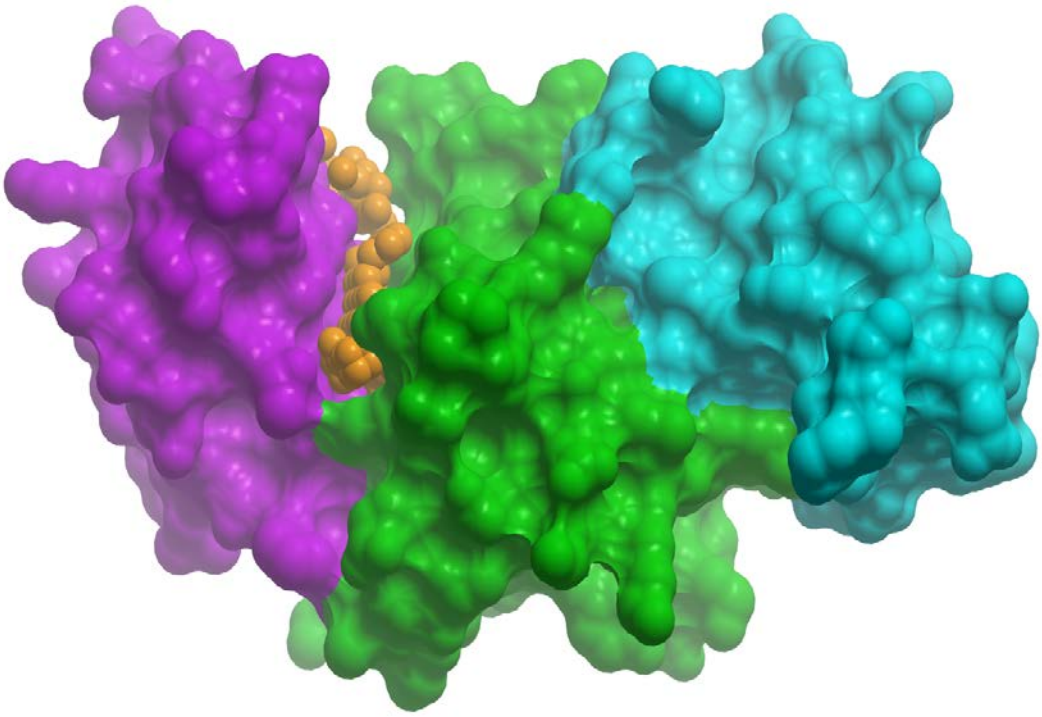


Figure 50. The ligandable cleft between the second (green) and third (magenta) Tudor domains of SETDB1. Site points are shown as orange balls. This site has a SiteScore of 1.00. PDB ID: 3DLM.

The histone methyltransferase SET Domain Bifurcated 1 (SETDB1) contains three contiguous Tudor domains. The available crystal structure (PDB ID: 3DLM) does not contain a well formed aromatic cage in any of the three Tudor domains. However, closer inspection reveals that the second and third Tudor domains contain suitable residues to form an aromatic cage, the formation of which could be induced by binding of a tri-methylated lysine. The first Tudor domain also contains two aromatic residues; however this potential aromatic cage is blocked by a lysine side chain. The triple Tudor domain contains a potential ligand binding site between the second and third Tudor domains (Figure 50). As discussed above, there are the two most likely to be involved in binding to methyl lysine, and therefore a ligand that binds between the second and third tudors is likely to impair peptide binding. This site has a SiteScore of 1.00 and this good score is primarily down to the site's large size (size score 160) and low hydrophilicity (0.88).

SPIN1

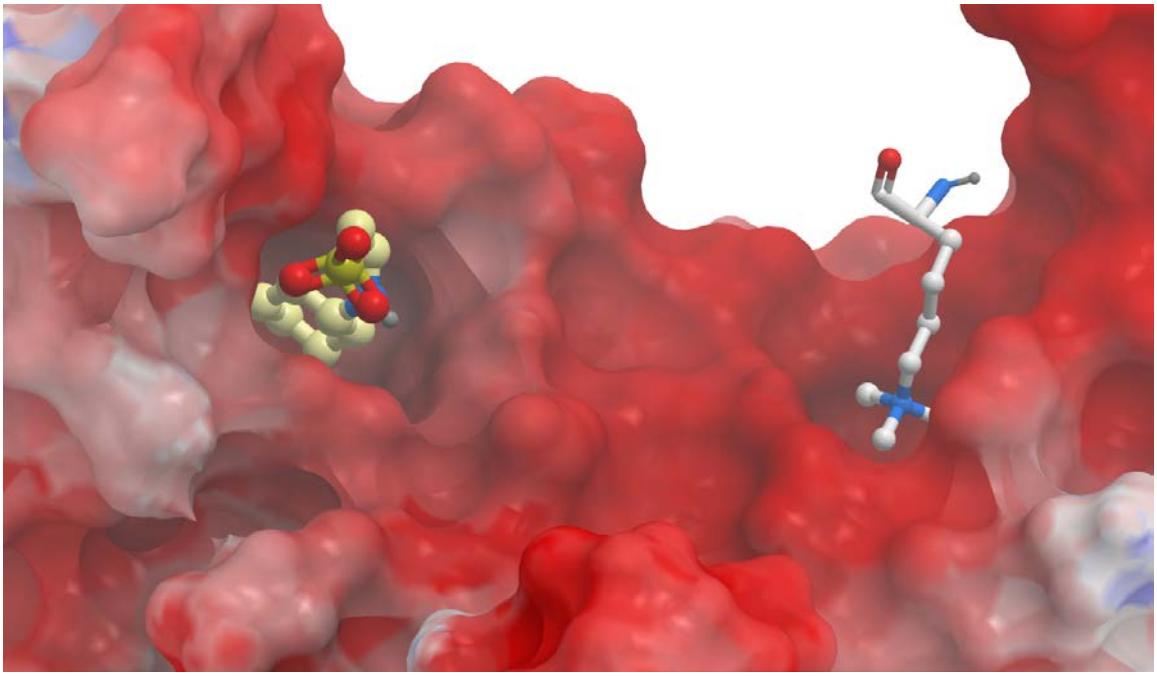


Figure 51. H3K4me3 binding to the second Tudor domain of SPIN1, and a molecule of the buffer CHES in binding site of the nearby first Tudor domain. It is possible a small molecule could bridge these two binding sites. PDB ID: 4H75.

Spindlin 1 (SPIN1) is a tudor domain containing protein involved in the regulation of rRNA expression.[135] It contains an unusual triple Tudor motif that is only seen in SPIN1, SPIN2, and SPIN3. The structure of the triple tudor of SPIN1 has been solved as a co-crystal with a histone 3 peptide ($H3_{1-8}$K4me3).[136] The tri-methylated lysine binds to an aromatic cage in the second Tudor domain, with the aromatic cage of the first Tudor occupied by a molecule of the buffer CHES (Figure 51). However, inspection of the protein-peptide complex suggests that this first aromatic cage is perfectly placed to bind a tri-methylated lysine 9. The H3K4me3 cage is a small hydrophobic site with a SiteScore of 0.87. Although the site is quite small (size score of 49), it achieves a good site score due to its highly hydrophobic nature. The hydrophobic cage of the N-terminal, potential H3K9me3 binding, Tudor domain has a lower SiteScore of 0.68. This site is similar to the H3K4me3 binding in that is small and hydrophobic. The openings of these pockets are 17 Å apart, which suggests that it would be possible to design a ligand that would take advantage of both of these potential sites.

SiteMap Analysis of Domain-Domain Interfaces

Having identified a pattern of tandem domains being more ligandable than single domains for PHDs and Tudor domains, it was decided that sites formed at domain-domain interfaces for other epigenetic proteins should be investigated. There are a growing number of X-ray crystal structures containing multiple domains involved in recognising and modifying post-translational marks on histone tails (Figure 52). These include examples of multiple reader domains from the same protein,[118,137] and also reader-writer[94] and reader-eraser combinations.[111] Structures containing multiple epigenetic domains provide an opportunity to study sites formed at the domain-domain interfaces, and investigate whether these sites may be suitable for ligand binding.

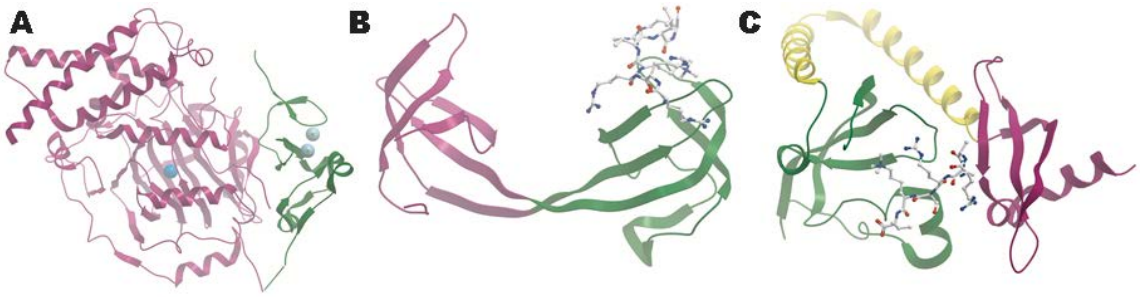


Figure 52. Examples of protein structures containing multiple epigenetic domains. **A.** The N-terminal PHD (green) and the JmjC domain (magenta) of PHF8. PDB ID: 3KV4. **B.** The tandem Tudor domains (N-terminal = magenta, C-terminal = green) of JMJD2A. A histone peptide bound to this tandem reader module is shown in stick form. PDB ID: 2QQS. **C.** Tandem chromodomains of CHD1 (N-terminal = green, C terminal = magenta). A histone peptide bound to this tandem reader module is shown in stick form, and the helix-turn-helix linker is shown in yellow PDB ID: 4NW2.

This section describes the use of SiteMap to analyse structures containing multiple histone binding/modifying domains in order to identify potential ligand binding sites at domain-domain interfaces. It is hoped that these sites may offer a potential solution for inhibitor design of histone binding/modifying proteins, where the individual domains do not possess ligandable sites.

Prevalence of Multi-Domains

The structures used in this study were identified using Chromohub.[138] Any structure of a histone reader, writer, or eraser which also contained another histone reader, writer, or eraser domain was included. A total of 108 structures were identified, representing 33 unique gene products. The most common domain combination was multiple-tudor of which seven examples were identified (Figure 53). A full list of structures identified is included in Appendix 3.3.

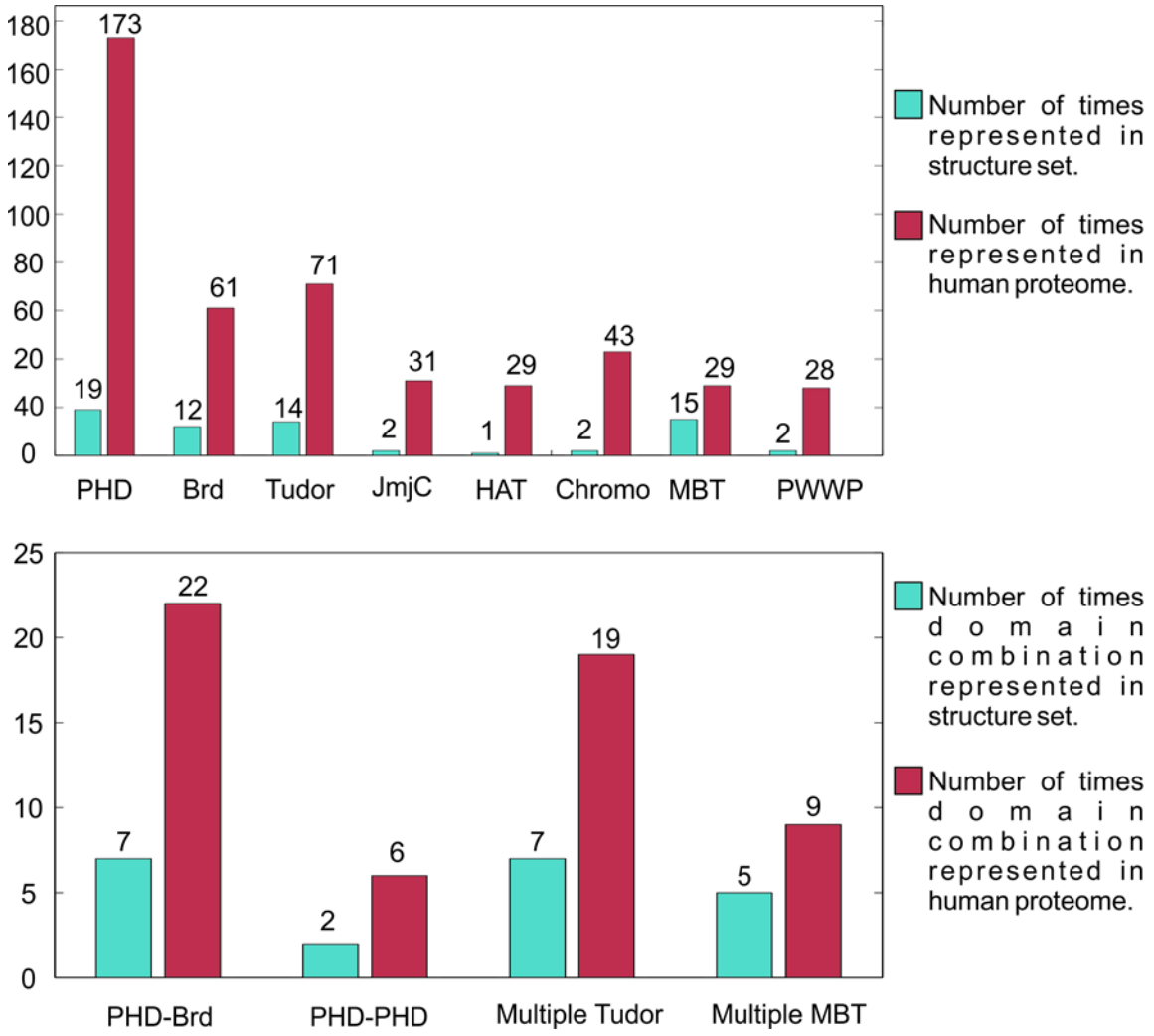


Figure 53. PHDs, bromodomains, Tudor domains, and MBT domains are the most common domains in our structure set. Structures containing multiple Tudor domains appear 7 times, as do structures containing multiple MBT domains. The most common hetero-domain combination in our structure set is PHD-bromo, of which 7 examples were identified. In our notation, a PHD-bromo has an N-terminal PHD and a C-terminal bromodomain; this is distinct from a bromo-PHD, in which the domain order is reversed.

The domain combinations in Figure 53 are all found adjacent to each other in the protein sequence, or have crystallographic evidence that they act together to engage histones. It is possible that domain combinations can have the histone binding potential greatly affected by the linker between the two domains. It is interesting to compare the JmjC-PHD domain combination found in PHD finger protein 8 (PHF8) with the JmjC-PHD domain combination found in Lysine-specific demethylase 7A (KDM7A, KIAA1718). The two JmjC domains, and the two PHDs are very close homologues to each other (63% and 77% sequence identity respectively), but differences in the linker effect the conformation of the two domains and hence the specificity of the demethylase activity of the JmjC domain.[111] In general it was possible to identify potential binding sites in most structures, and in some cases these were found at domain-domain interfaces. Specific examples are discussed below, and the key findings summarised in Figure 54.

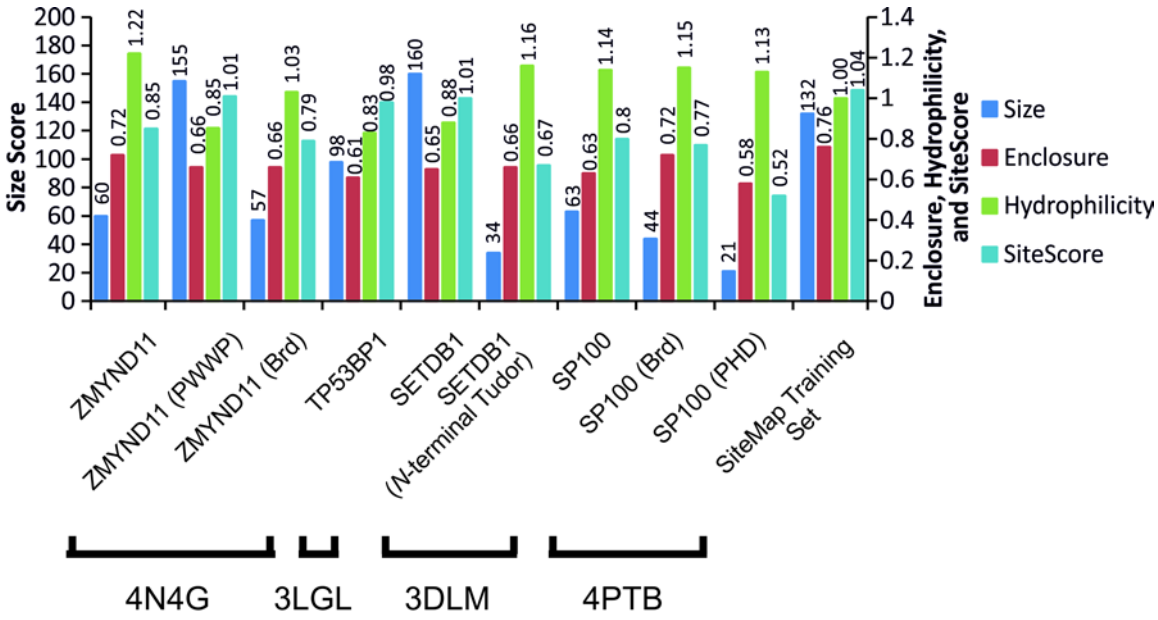


Figure 54. The size, enclosure, hydrophilicity, and SiteScore of all the examples discussed below are shown. Scores are also shown for those cases where SiteMap identified a site in a single domain of the multidomain complex. The mean values for the 326 binding sites with known submicromolar ligands that were used as a training set during the development of SiteMap are show for comparison. Brd = Bromodomain.

Bromodomain-PWWPs

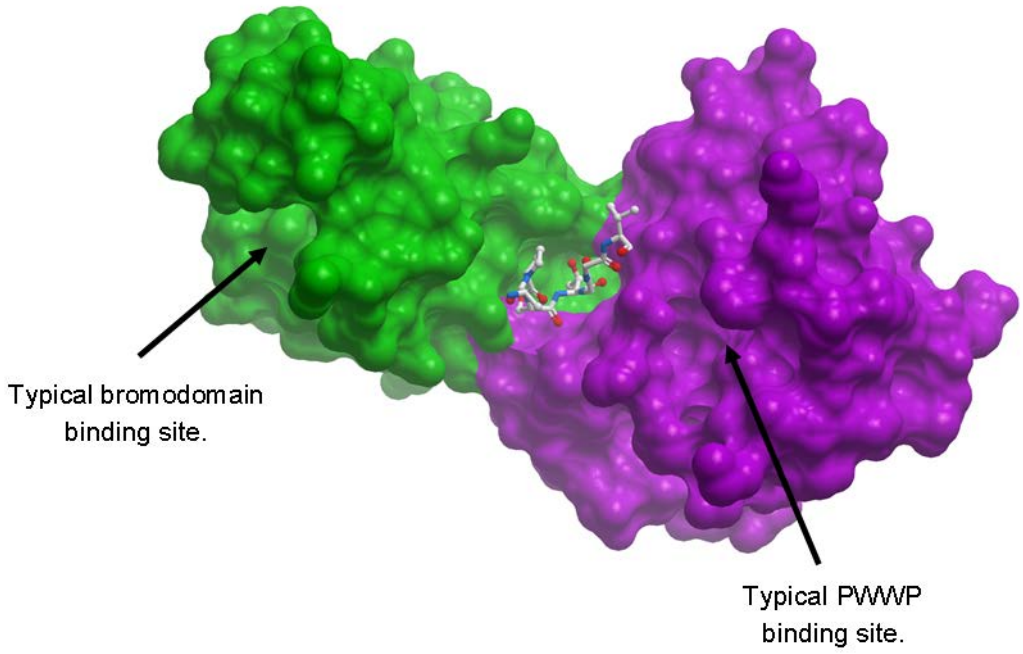


Figure 55. The structure of the bromodomain (left, green) and the PWWP domain (right, magenta) of ZMYND11. Residues 29-39 of H3.3 are shown in stick form at the binding site identified between the two domains PDB ID: 4N4I.

Bromodomains are a family of acetyl-lysine reader domains.[27] The acetyl-lysine binding sites are well characterised as potential small molecule ligand binding sites,[128] and many bromodomain inhibitors have been discovered.[29] PWWP domains are a part of the Royal family of methyl-lysine binding domains. The only previous study of PWWP ligandability considers the PWWPs of only Bromodomain and PHD Finger containing Protein 1 (BRPF1) and Hepatoma-Derived Growth Factor-Related Protein 2 (HDGF2). This study found these PWWPs to have SiteScores of approximately 0.90, which suggests they are at the more difficult end of proteins that would be considered druggable.[78] However, this study only takes into account two of the 28 known human PWWPs, and therefore may not be represent the entire family.

The structure of the PHD-bromodomain-PWWP domain of tumour suppressor Zinc Finger MYND-type containing 11 (ZMYND11) has been solved with a bound H3.3 peptide (Figure 55).[139]

Wen et al. identify an interaction between a serine residue (S31) found in H3.3 and the bromodomain-PWWP domain-domain interface. This recognition allows ZMYND11 to differentiate between H3.3 and the more common H3 which has an alanine at position 31. Isothermal calorimetry (ITC) showing a 7-fold difference in binding to ZMYND11 between a H3.3 peptide and a H3 peptide (for PHD-bromodomain-PWWP of ZMYND11 and $H3.3_{19\text{-}42}$K36me3 K_D = 6 µM, for H3K36me3 K_D = 431 µM). Analysis of the structure of ZMYND11 (PDB ID: 4N4G) with SiteMap identified the bromodomain as the most ligandable site (SiteScore = 1.01); however, this is unlikely to be biologically relevant as this bromodomain has not yet been shown to bind to acetyl-lysine.[139] A binding site at the bromodomain-PWWP interface where S31 binds is identified as having a SiteScore of 0.85. This is at the lower end of what would be considered ligandable.[130] Further analysis of this binding site shows it is a well enclosed site, with a SiteMap enclosure score of 0.72, this compares favourably to the mean value 0.76 for the sub-micromolar sites in SiteMap's training set. The site is relatively small; with a SiteMap size score of 60, whereas the mean size score of the sub-micromolar sites in the training set is 132. However, this smaller size should not prove inhibitive to ligand design.

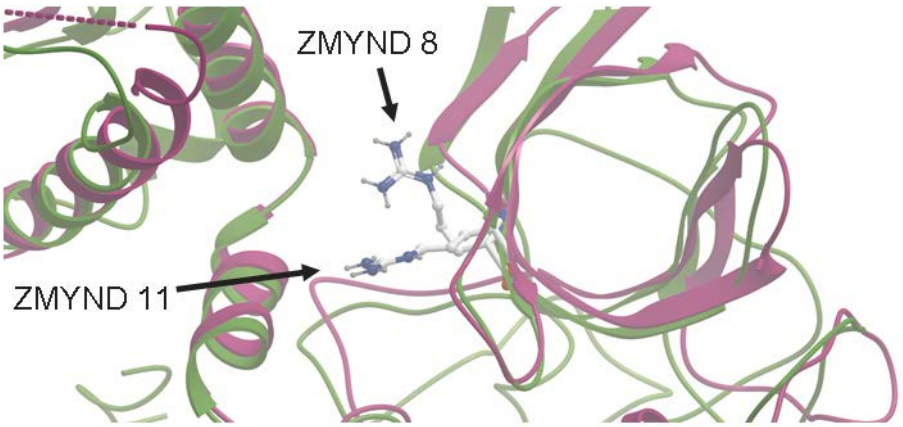


Figure 56. The differing positions of R309 of ZMYND11 (magenta) and the equivalent R306 of ZMYND8 (green). The difference in conformation partially explains the difference in predicted ligandability between the two homologues. PDB IDs: 4COS (ZMYND8), 4N4H (ZMYND11).

It is clearly apparent that a chemical probe that bound at this site would prove of great use to the study of the in vivo function of the interaction between ZMYND11 and H3.3.

It is interesting to compare this binding site to the equivalent site on homologue Zinc Finger MYND-type containing 8 (ZMYND8, PRKCBP1); this contains a binding site at the bromo-PWWP interface with a SiteScore of 0.92. This site is the same size as the one on ZMYND11, but slightly more enclosed (enclosure score 0.81) and less hydrophilic. The bromo-PWWP of ZMYND8 has 36% sequence identity with ZMYND11; however, differing conformations of an arginine residue found in both ZMYND11 and ZMYND8 results in a difference in ligandability (Figure 56). R309 of ZMYND11 forms part of the bottom of the pocket, whereas R306 of ZMYND8 sits at the top of the pocket making it more enclosed. However, as the structure of ZMYND8 was solved without a peptide ligand, it may be the case that this arginine adopts the position seen in ZMYND11 on peptide binding

PHD-Bromodomains

The structure set used included 20 structures containing a PHD and another domain, the most common partner domain being a C-terminal bromodomain (8 examples). PHD-bromodomains have been shown to act together in substrate recognition.[44] The PHD-bromodomain structures investigated in this study fell into two categories, those where the PHD and bromodomain form a compact globular structure, and those where the PHD and bromodomain are separated by a rigid linker. An example of a case where the PHD-bromodomain forms a globular structure is the putative transcriptional corepressor Tripartite Motif Containing 33 (TRIM33) (Figure 57A). The case with a rigid linker is exemplified by Bromodomain PHD Finger Transcription Factor (BPTF) which forms part of the NURF nucleosome remodelling factor (Figure 57B).

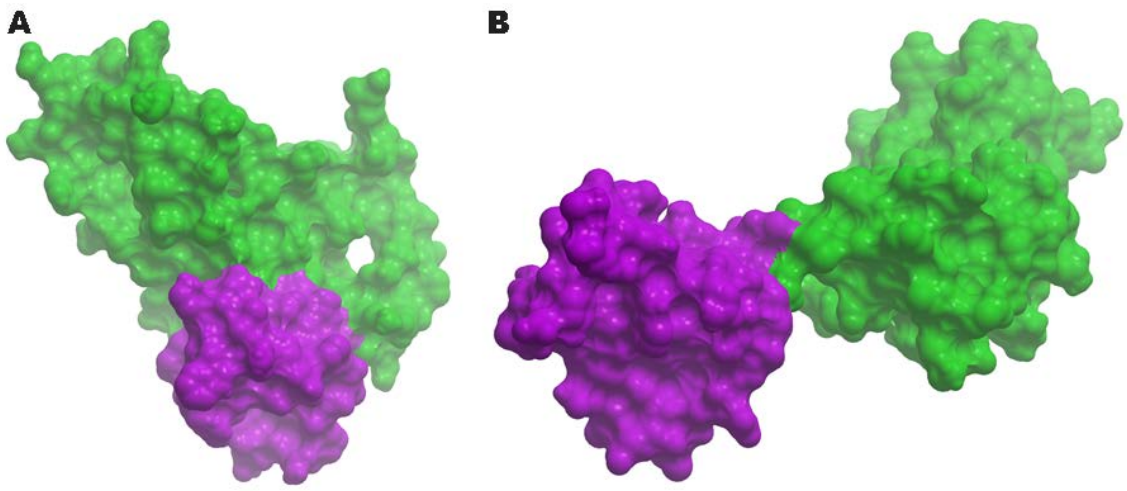


Figure 57. Examples of two types of PHD-bromo. **A.** The PHD (magenta) and bromodomain (green) of SP100 form a compact, globular structure. PDB ID: 4PTB. **B. T**he PHD (magenta) and bromodomain (green) of BPTF are separated by a rigid linker. PDB ID: 2FSA.

Amongst the group of structures with a globular arrangement of the PHD and bromodomain was SP100. Analysis of this structure by SiteMap revealed the presence of a novel potential ligand binding site at the interface of the PHD and bromodomain. This site has a SiteScore of 0.81, which would place it the range of difficult but ligandable binding sites. Further analysis of the site revealed that it is large and well enclosed, but suffers a penalty to its SiteScore due to its hydrophilic nature (Figure 58). This domain-domain interface site has a higher SiteScore than the typical bromodomain binding site (SiteScore = 0.77) or the histone binding face of the PHD (SiteScore = 0.53).

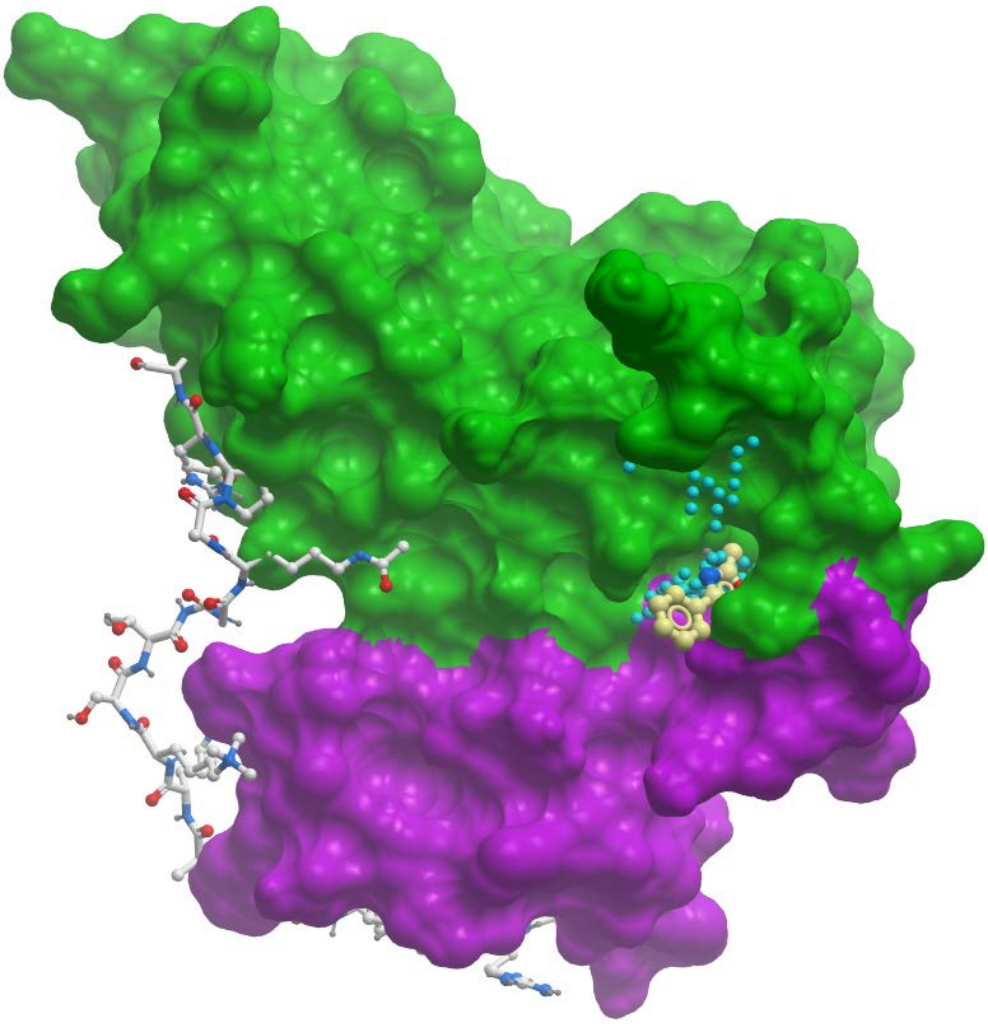
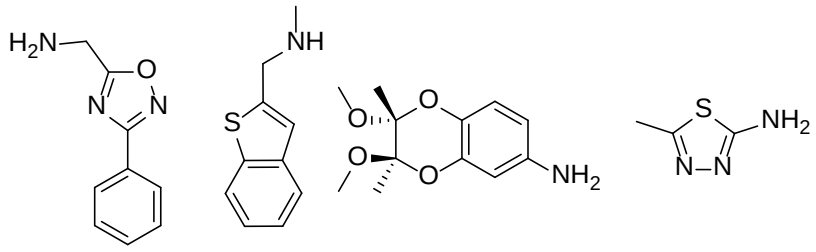


Figure 58. The PHD-Bromodomain of SP100 is shown with the site points as identified by SiteMap, and overlaid ligand **5** identified by a crystal soaking experiment. The novel potential ligand binding site at the domain-domain interface of the PHD-bromodomain of SP100 is indicated by cyan dots, the PHD is shown in magenta, and the bromodomain is shown in green. The peptide ligand shown is taken from the structure of the related PHD-bromodomain of TRIM33. PDB ID: 3U5M.

In order to further investigate this newly identified binding site, a fragment soaking experiment was performed using a 527 member fragment library.[f] This soaking experiment revealed three ligands that bind in the novel pocket at the interface of the PHD and bromodomain (Figure 59). Based on comparisons to the related PHD-bromodomain of TRIM33,[44] this pocket is not

[f] This work was performed by Romain Talon and other colleagues at the Structural Genomics Consortium.

expected to be involved in histone peptide recognition; however, it is possible that a ligand binding at this position could allosterically modify the PHD-bromodomain.



5 **6** **7** **8**

Figure 59. A fragment soaking experiment identified three compounds (**5** – **7**) that bind at the novel binding site at the PHD-bromodomain interface. Compound **8** was initially characterised as a hit, but a review of the crystallographic evidence revealed there was no strong evidence to support this characterisation.

Oxadiazole **5** and benzothiophene **6** bind deeper into the pocket than aniline **7**. They exploit hydrophobic interactions with surrounding residues and oxadiazole **5** forms a hydrogen bond with the backbone carbonyl of residue I871 via its primary amine. Ligands **5** and **6** also induce a conformational change in residue Q751 not observed for ligand **7**. The primary amide of Q751 is rotated through 120° relative to the *apo* structure (Figure 60).

Aniline **7** does not bind as deeply in the pocket as oxadiazole **5** and benzothiophene **6**, and therefore do not cause any movement of residue Q751 relative to the *apo* structure. Aniline **7** exploits hydrophobic interactions with surrounding residues, whereas thiadiazole also forms a hydrogen bond with K785 (Figure 61).

A comparison of the binding mode of the three identified fragments shows that oxadiazole **5** and benzothiophene **6** bind deeply into the pocket predicted by SiteMap, with aniline occupying a hydrophobic site at the mouth of the pocket. This arrangement of these hits suggests that a fragment linking strategy may lead to higher potency ligands (Figure 62).

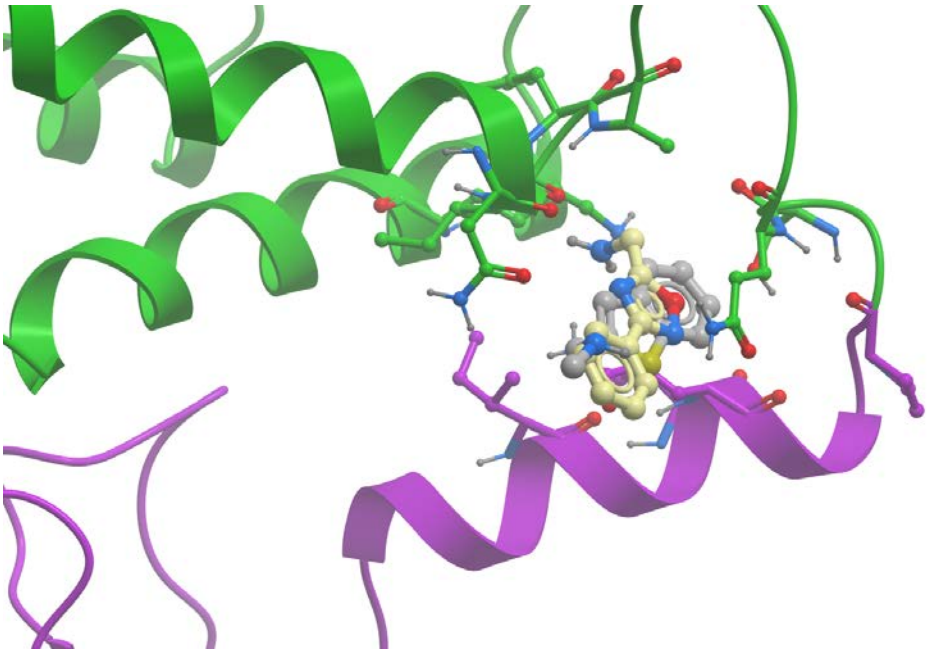


Figure 60. Binding mode of oxadiazole **5** (yellow) and benzothiophene **6** (grey) to SP100. The backbone ribbon and side chains of important residues from the PHD are coloured magenta, and the backbone ribbon and important residues from the bromodomain are coloured green.

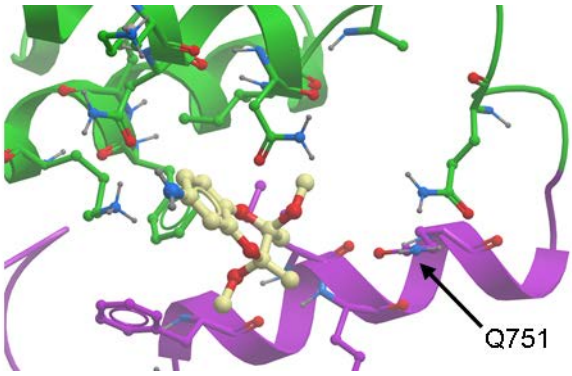


Figure 61. Binding pose of aniline **7**. This ligand does not bind as deeply in the pocket as oxadiazole **5** and therefore does not cause a conformational change in residue Q751.

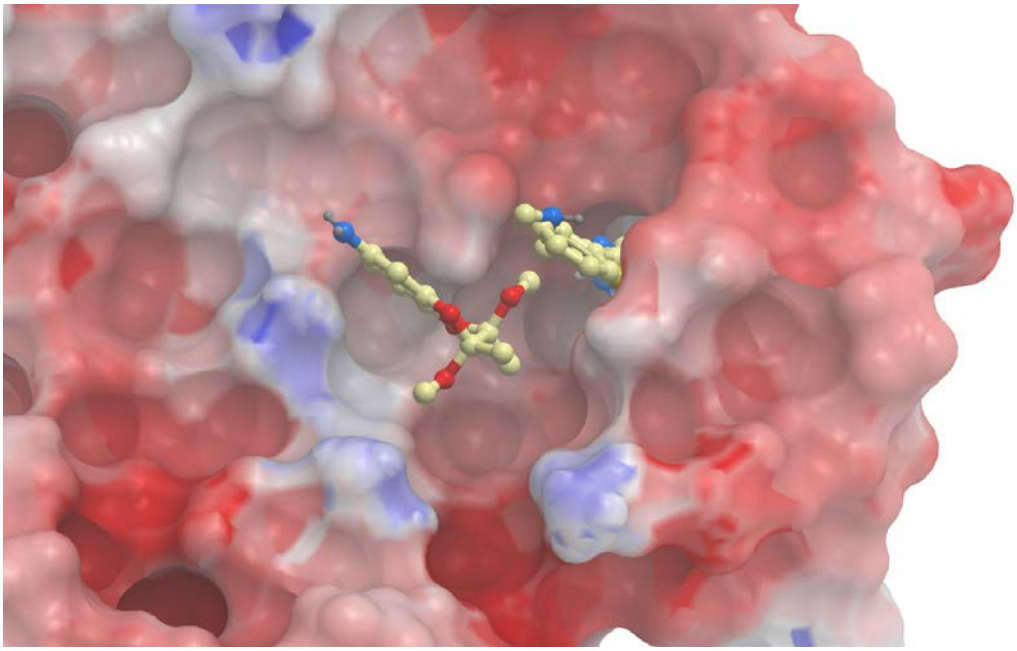


Figure 62. The three fragment hits identified for the inter-domain binding site of SP100 are shown overlaid with an electrostatic map of the surface of the *apo* structure of SP100.

Chromodomains

Chromodomains are another member of the royal family. The only chromodomain containing protein in the structure set was the double chromodomain of Chromodomain Helicase DNA Binding Protein 1 (CHD1). Although druggable chromodomains have been identified[78] only one chromodomain inhibitor has been reported.[140] The double chromodomain of CHD1 is known to utilise the domain-domain interface in the selective binding of methylated H3K4. Although the histone binding site itself was not ligandable, there is a potential allosteric site on the reverse face of the tandem chromodomains, at the domain-domain interface. This site is large (size score = 95) and open (enclosure score = 0.68) and quite hydrophilic. Although a ligand binding at this site would not compete directly with the histone, it may allosterically modify the histone binding ability of the tandem chromodomains of CHD1.

Conclusions

Based on these results we have decided to focus our initial experimental investigation on tandem PHDs, as SiteMap suggests these are likely to be more druggable than single PHDs, due

to their larger potential ligand binding sites. There are six human tandem PHDs, found in DPF1, DPF2, DPF3, PHF10, MYST3, and MYST4. One of the tandem structures used in the study above was from MLL3; however, inspection of the protein sequence suggests that this is likely to form part of a triple PHD. Triple PHDs are also found in MLL1, MLL2, MLL3, MLL4, NSD1, WHSC1, and WHSC1L1. For simplicity, tandems that are part of a larger triple have been excluded from further study at this stage.

Work described in this chapter also suggests that the observation that tandem domains are more ligandable than single domains may also be true for tandem Tudor domains, and for proteins containing multiple domains involved in epigenetic regulation. Experimental work by colleagues at the Structural Genomics Consortium has partially validated this prediction by identifying ligands that bind at the PHD-bromodomain interface of SP100.

Chapter 4 - Assessing the Ligandability of Tandem PHDs

As discussed in Chapter 3, PHDs that exist as part of multi-domain complexes are more likely to be amenable to inhibition by small molecules. The specific examples of the tandem PHDs of the type found in MYST3 and DPF3b were discussed. These two proteins share a similar tandem PHD domain, with the two PHDs arranged in the same front-to-back manner. In both cases the tandem PHD is thought to bind to residues 1-14 of histone 3, with a non-canonical pocket on the rear face of the *N*-terminal PHD which recognises acetylated lysine 14 (H3K14ac).[116,132,133] This chapter will provide a working definition of a tandem PHD domain, and describe assay development and screening efforts for the tandem PHDs of DPF2.

Definition of a Tandem PHD

PHDs are often found multiple times within a single protein (Chapter 2), with one example, Mixed Lineage Leukaemia 3 (MLL3), containing eight PHDs. This chapter will only cover PHDs that are found as a pair, which are not part of a larger triple or quadruple PHD, and where the final zinc binding residue of the *N*-terminal PHD is separated from the first zinc binding residue of the *C*-terminal PHD by exactly two residues (Figure 63). This definition has been chosen as it matches the domain architecture found in the structures of MYST3[133] and DPF3b.[132]

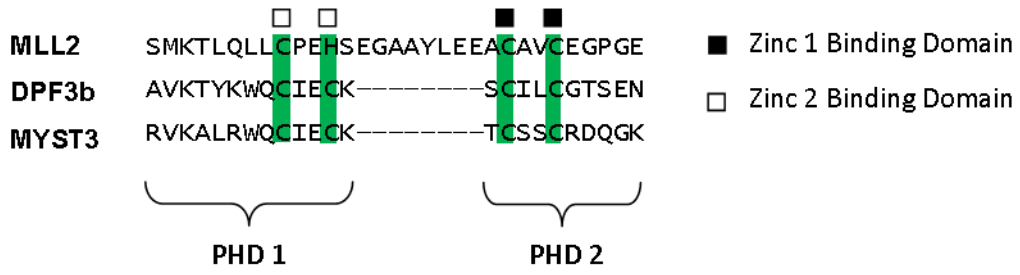


Figure 63. Sequence alignment of the linker region between PHD 1 and PHD 2 of Mixed Lineage Leukaemia 2 (MLL2) and Double PHD Finger Protein 3 (DPF3). MLL2 contains ten residues between the final Zn(II) binding residue of PHD 1 and the first Zn(II) binding residue of PHD 2, whereas DPF3 only has two residues. Therefore DPF3 fits within the definition of tandem PHDs that will be discussed in this chapter.

Often PHDs are found near to each other in a sequence, and in some cases are thought to interact in a biologically relevant way, such as in the case of Chromodomain Helicase, DNA-binding 5 (CHD5).[141] There are also many examples of two PHDs separated by a 'zinc knuckle' (Figure 64).[121,122] The evolutionary conserved nature of these zinc knuckle separated PHDs, suggest that they may act as one distinct unit (Chapter 2). Although these PHDs exist in evolutionary conserved pairs they will not be classed as tandems in this chapter as they do not meet the linker requirement described above.

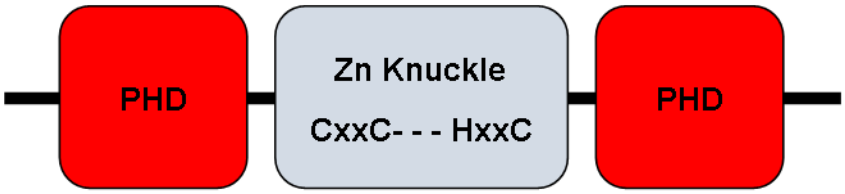


Figure 64. The PHD- Zn Knuckle-PHD motif. Members of sub-group 12 form the first PHD of this motif. The Zn knuckle section consist of four Zn(II) binding residues arranged in two pairs. Within each pair, the Zn(II) binding residues are separated by two amino acids.

This definition excludes tandem PHDs that appear to be part of a larger triple or quadruple PHD arrangement, such as those found in the MLL and NSD families. This is because there is a lack of structural and functional data available for triple and quadruple PHDs. Therefore considering a pair of PHDs from within these larger multiple PHDs may not be biologically relevant.

Phylogeny of Tandem PHDs

Examining the phylogeny of tandem PHDs shows that there is a clear difference between tandem PHDs that are part of a larger triple/quadruple PHD and those that are not (Figure 65). The tandems PHDs that fit the definition described above fall into three groups. DPF1 (where the tandem PHD differs slightly between the two isoforms), DPF2, and DPF3 form one group; MYST3 and MYST4 the second; with PHF10 as an outlier. The outlying nature of PHF10 is unexpected, as it has functional similarity to the DPF family, playing a role in the BAF chromatin remodelling complex.[142]

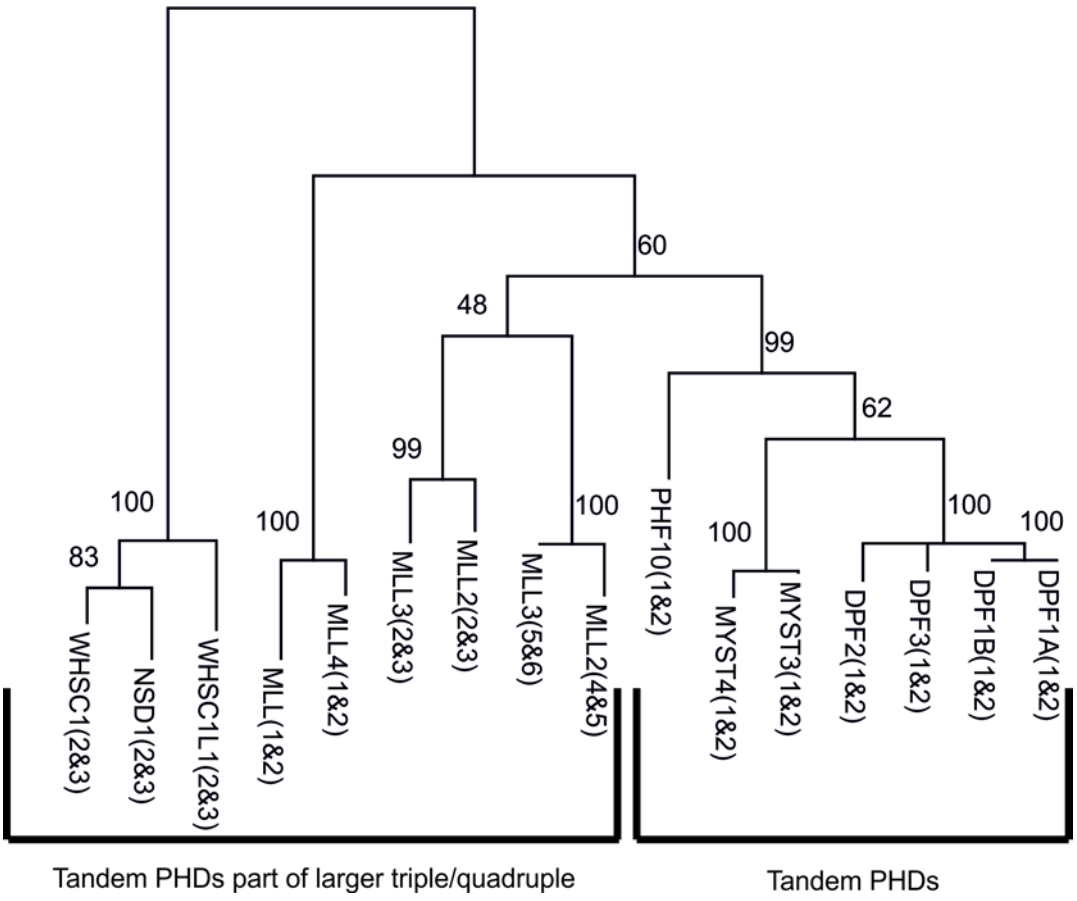


Figure 65. Phylogenetic tree showing the six tandem PHDs that fit the definition described above and those that don't as they are part of a larger triple/quadruple PHD. Bootstrap values are shown on branch points.

Structural Features of Tandem PHDs

There are six tandem PHDs in the human proteome that meet the definitions described above. These are found in the histone acetyl transferases MYST3 and MYST4 and the BAF chromatin remodelling associated proteins PHD Finger Protein 10 (PHF10), DPF3b, DPF1, and DPF2.

Four members of this set of tandem PHDs have been shown to bind acetylated lysines, specifically at H3K14.[132,133,143,144] The available structures of DPF3b and MYST3 show that the two domains are arranged in a front to back arrangement (Figure 66). With the face typically involved in histone recognition (the 'front') of the *C*-terminal PHD in contact with the back face

of the *N*-terminal PHD. The *C*-terminal domain is twisted by 90° along the front-back axis compared to the *N*-terminal domain.

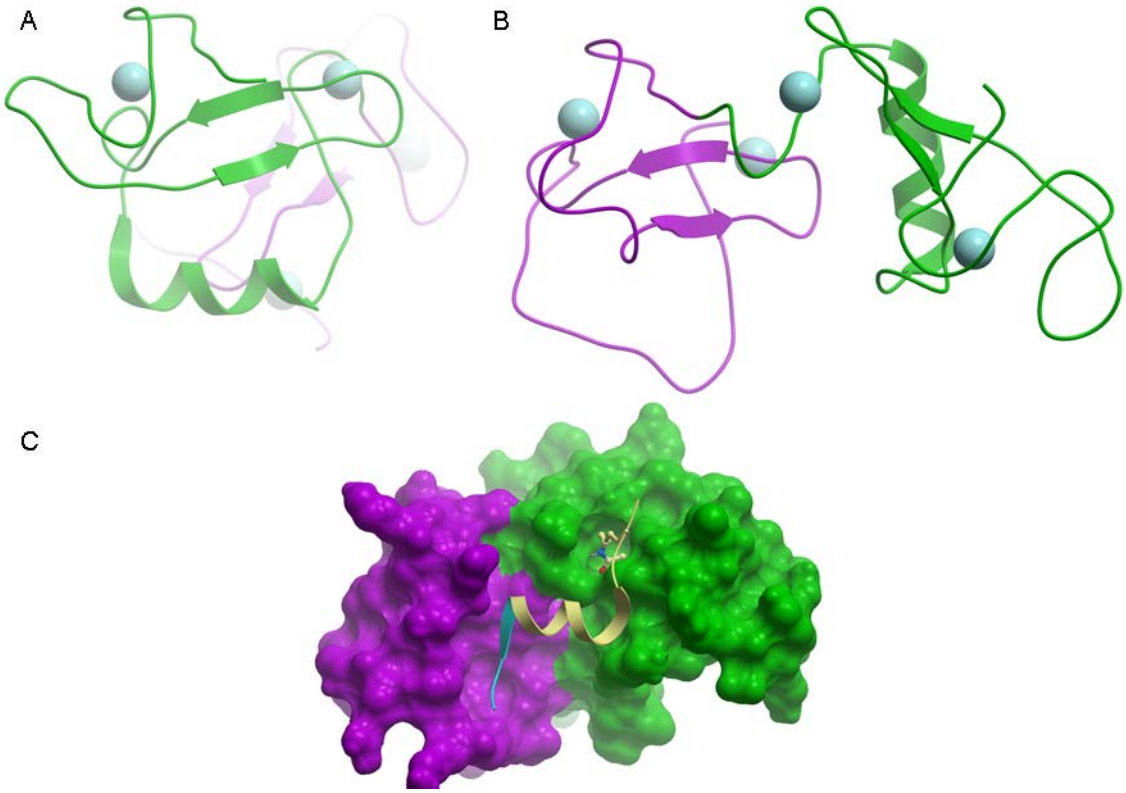


Figure 66. All structures of tandem PHDs show the same front to back arrangement. **A.** The front of the *N*-terminal PHD (green) of MYST3 with the *C*-terminal domain (magenta) shown on its rear face. **B.** The front of the *C*-terminal domain of MYST3 (magenta) which is making contact with the *N*-terminal PHD (green). **C.** The tandem PHD of MYST3 has a novel pocket on the back of the *N*-terminal PHD (green) which binds H3K14ac (yellow). The *N*-terminus of H3 binds to the *C*-terminal PHD of MYST3 in a canonical manner (cyan). PDB ID: 4LLB.

The available solution NMR structures of DPF3b and the crystal structure of MYST3 show that the H3K14ac binds in a novel pocket on the back of the back of the *N*-terminal PHD, with residues 1-4 engaging the *C*-terminal PHD in a canonical fashion (Figure 66C).

Biological Function of Tandem PHD Containing Proteins

DPF1-3 and PHF10

PHF10 and DPF1-3 are sometimes grouped together as the d4 gene family and are known to play a role in neural development.[145] Proteomic analysis of BAF chromatin remodelling complexes purified from neural stem cells shows that PHF10, DPF1, and DPF3 are present as sub-units of this complex. Further analysis shows that PHF10 is present in proliferating stem cells, but is replaced by DPF1 and DPF3 in differentiated neural cells.

MYSTs

MYST3 and MYST4 are histone acetyl transferases (HATs) and are known to form part of a HAT complex with Inhibitor of Growth 5 (ING5) and Bromodomain PHD Finger Protein 1 (BRPF1). MYST3 plays an important role in controlling the renewal and differentiation of haematopoietic stem cell (HSCs) and is a common target for mutations and translocations in leukaemia.[146] MYST4 is involved in the regulation of neurogenesis, controlling the regulation of renewal and differentiation of neural progenitor cells.[147]

Tandem PHDs in Disease

It is interesting to note the potential role of tandem PHDs in human disease. PHF10, which contains no other domains recognised by Pfam[95] or SMART,[96] has been shown to inhibit the expression of caspase-3 in gastric cancer cells, and knock down of PHF10 by miRNA has been shown to induce apoptosis.[148,149] A chemical probe that inhibits the tandem PHDs of PHF10 would facilitate study of the role the tandem PHD plays in this process. MYST3 and MYST4 can form fusion proteins, containing their tandem PHDs, with the histone acetyl transferases CREBBP and p300, forming an aberrant acetylation complex that plays a role in acute myeloid leukaemia.[147] Selective inhibitor of the interaction of the tandem PHDs of MYST3 and MYST4 with chromatin would allow study of the role these domains play in the fusion protein and disease progression.

Assay Development for Tandem PHDs

Having identified tandem PHDs as potentially being ligandable than single PHDs, and in the case of MYST3/4 and PHF10 identifying links to disease, a selection of screening methods were used in an attempt to discover a small molecule inhibitor of a tandem PHD. This section describes the choice of assays and screening libraries used, and the optimisation of these assays. The results of these screening efforts will be discussed in a subsequent section alongside secondary assays used for validation of primary hits.

AlphaScreen Development

An Amplified Luminescent Proximity Homogeneous Assay (AlphaScreen) was chosen as a suitable primary screening platform for tandem PHD domains. The assay uses a tagged protein and an orthogonally tagged ligand: a His6-tagged tandem PHD of DPF2 and a biotinylated histone peptide. The protein and peptide are incubated with streptavidin coated donor beads and Nickel-nitrilotriacetic acid (Ni-NTA) coated acceptor beads. When illuminated with red light (680 nm) the photosensitiser phthalocyanine immobilised on the donor beads converts ambient oxygen to its excited singlet state. This excited species can diffuse approximately 200 nm in solution before being quenched; however, if it comes into contact with a rubrene dye immobilised on the acceptor bead it will generate an emission between 520 nm and 620 nm. Therefore if the protein is bound to the peptide ligand, the donor and acceptor beads will be close enough together for singlet oxygen to diffuse between them and generate a signal (Figure 67). If a small molecule is inhibiting the interaction between protein and peptide, the donor and acceptor beads will no longer be in close proximity and a loss of signal will be observed.

One of the main strengths of AlphaScreen is the low amount of protein and peptide required. Each bead has multiple binding sites, therefore an avidity effect caused by multiple binding events between protein and peptide means that the affinity between beads is much greater than the affinity between protein and peptide in solution. This means that a good signal can be achieved with low nM concentrations of protein and peptide.

Compounds that absorb light in the range of the excitation or emission wavelengths have the potential to cause a loss of signal in AlphaScreen and appear as a false positive. Similarly singlet oxygen quenchers can also cause a loss of signal. It is therefore important to compare any primary hit identified by AlphaScreen with the many documented functional groups which are known to be problematic for use with AlphaScreen.[150]

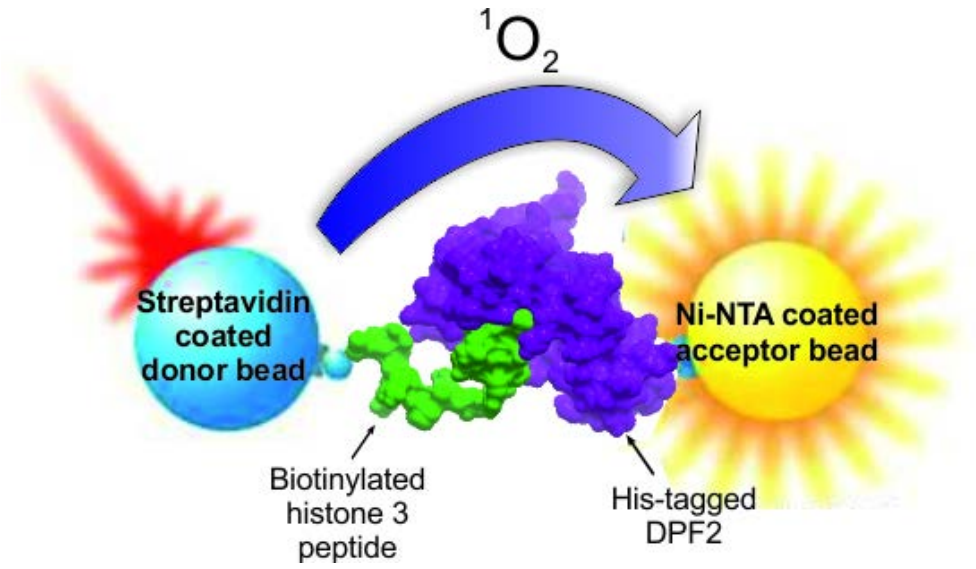


Figure 67. An AlphaScreen assay consisting of a biotinylated peptide and His6-tagged protein. When the protein and peptide are in contact they bring the donor and acceptor beads close together. When the donor bead is excited with red light single oxygen is generated, which diffuses to the acceptor bead causing an emission between 520 nm and 620 nm.[151]

All AlphaScreen hits should also be triaged using a suitable counter-screen. A peptide containing both a His6-tag and a biotinylated lysine can be used for such a counter-screen. This dually tagged peptide can be used in place of the protein and peptide used in the initial screen and will link donor and acceptor beads. Therefore any loss of signal caused by a compound in the counter-screen must be due to interference with the assay. This counter-screen procedure is useful for identifying false positives that act by interfering with the assay, but it will not identify false positives than act through other mechanisms such as aggregation or protein reactivity.

In summary AlphaScreen is a sensitive assay that is well suited for medium throughput screening. However, it is important to triage all hits with appropriate counter-screens as the assay offers numerous possible mechanisms for assay interference.

Protein Production

The tandem PHDs of DPF2 and PHF10 were chosen for AlphaScreen assay development due to the availability of Rosetta *Escherichia coli* (*E. coli*) cells transformed with a suitable plasmid vector. The vectors used encoded for a glutathione *S*-transferase (GST) tandem PHDs fusion protein with an *N*-terminal His6 tag and a Tobacco Etch Virus (TEV) protease cleavage site between the GST and tandem PHDs. The His6 facilitates use in an AlphaScreen assay, the GST tag help maintain protein solubility both during bacterial expression and purification, and the TEV protease cleavage site allows for removal of these tags if untagged protein is required for further experiments.

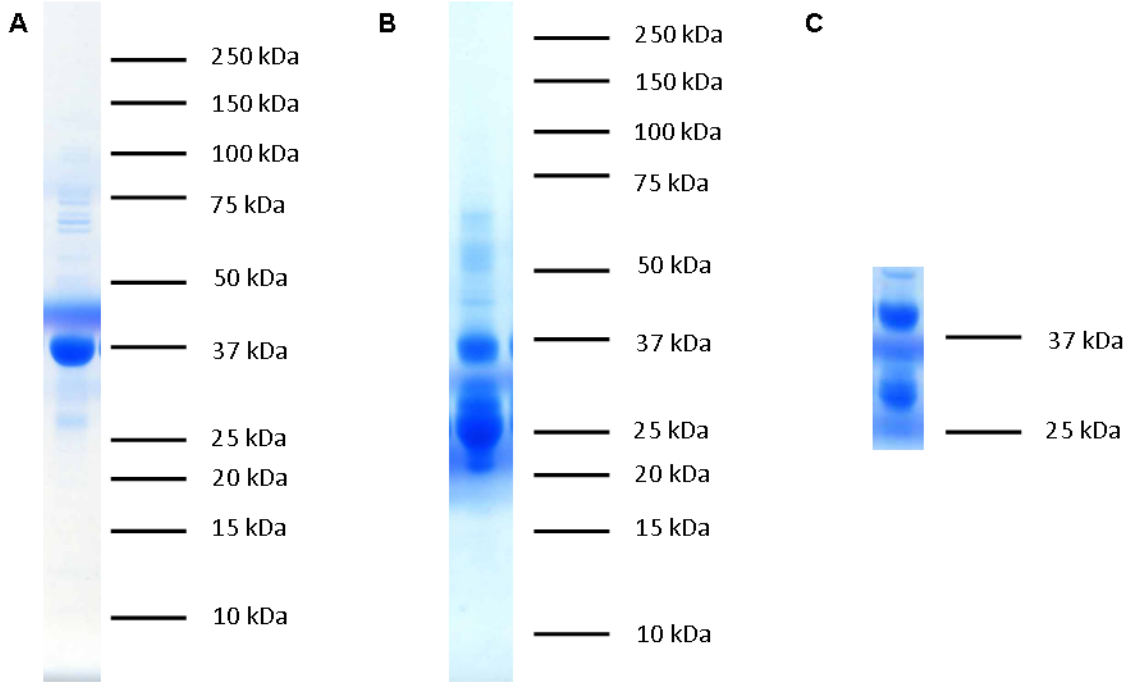


Figure 68. SDS-PAGE stained with Coomassie Brilliant Blue showing purification of DPF2 and PHF10. **A.** Tandem PHDs of DPF2 after size exclusion chromatography. **B.** Tandem PHDs of PHF10 after size exclusion chromatography. The band present at approximately 25 kDa was shown to be GST by tryptic digest MSMS. **C.** Tandem PHDs of PHF10 after anion exchange chromatography. The band present at approximately 25 kDa is still present.

The required proteins were purified from cell lysate using a Ni-NTA column followed by size exclusion chromatography. It was possible to obtain suitably pure DPF2 in this manner (Figure 68A), but it was not possible to separate PHF10 from a truncated protein containing just the GST tag (Figure 68B). A further attempt to remove this impurity with anion exchange chromatography proved unsuccessful (Figure 68C). Therefore only DPF2 was carried forward for further assay development.

Identification of Peptide Ligand

An AlphaScreen assay for DPF2 required the identification of a suitable peptide ligand. The tandem PHDs of the closely related DPF3b, and the less closely related MYST3 have been shown to bind H3 peptide acetylated at H3K14 (H3K14ac).[132,133] Dimethylation of H3R2 has also been shown to be deleterious to binding to the tandem PHDs of MYST3. The solution NMR structure of DPF3b shows that the first sixteen residues of H3 are involved in binding to the tandem PHDs. Therefore it was hypothesised that the DPF2 AlphaScreen peptide ligand would be required to be at least sixteen residues from the *N*-terminus with acetylation at H3K14 with no modification at H3R2.

A library of histone peptides were screened against DPF2 using a BioLayer Interferometry (BLI) assay. This assay involves optical fibres with streptavidin coated tips. Biotinylated histone peptides are immobilised on these tips. This creates two surfaces: one at the interface between the glass fibre and the streptavidin tip, and a second one between the immobilised peptide and bulk solution. White light is sent down the optical fibre and is reflected from both surfaces, creating an interference pattern where some wavelengths undergo constructive interference, and some destructive interference. The interference pattern is a function of the distance between the two interfaces. When the optical fibre is immersed in a solution containing DPF2, the tips coated with peptides which bind to DPF2 will undergo a change in interference pattern

as the distance between the two surfaces is increased when DPF2 binds. This allows rapid identification of which histone peptides bind DPF2 and which don't.

A total of 192 peptides were tested in two sets of ninety six (Appendix 4.1). The tandem PHDs of DPF2 had a strong preference for binding peptides which included the *N*-terminus of H3. Comparison of a H3 spanning 1-21 which was phosphorylated at S10 and T11, with a similar peptide containing the same modifications spanning residues 4-24 shows that H3 residues 1-3 play an important role in binding to DPF2 (Figure 69). These peptides were chosen for this comparison as these phosphorylations appear to have no effect on binding and the unmodified peptides of the corresponding lengths were not available.

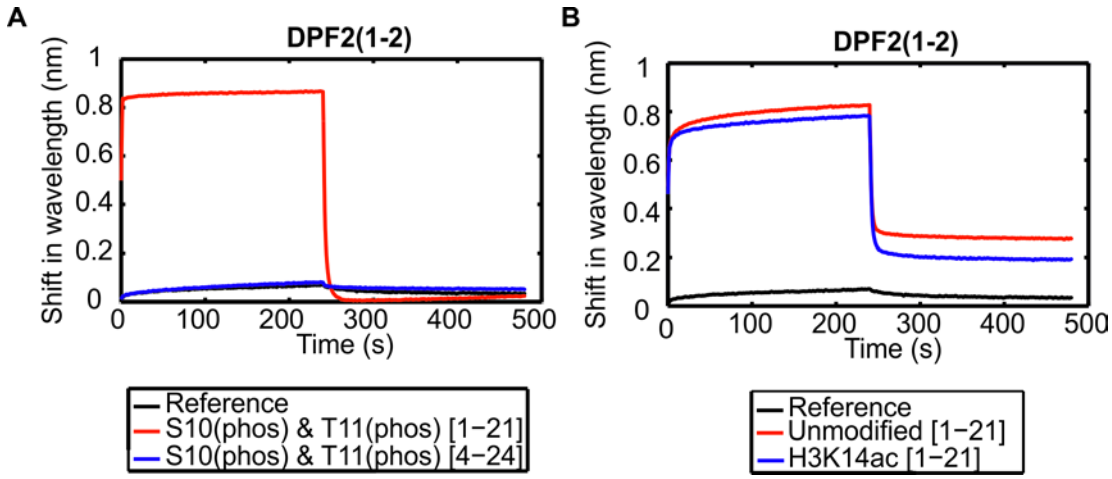


Figure 69. A. Comparison of the binding of peptides with and without H3 residues 1-3 as measured by BLI show that residues 1-3 are important for binding to DPF2. Peptides phosphorylated at H3S10 and H3T11 were used to assess the importance of residues 1-3 on binding as unmodified peptides of the required lengths were not available. Phosphorylation of peptides at H3S10 and H3T11 appears to have no adverse effects on binding. **B.** Comparison of the effect of acylation of H3K14 on binding as measured by BLI. For DPF2(1-2) acetylation of H3K14 appears to have no strong effect on binding.

Comparison of H3 peptides spanning residues 1-21 with and without an acetylation at K14 shows very little difference in DPF2 binding. This suggests that both of these peptide are suitable binding partners for AlphaScreen.

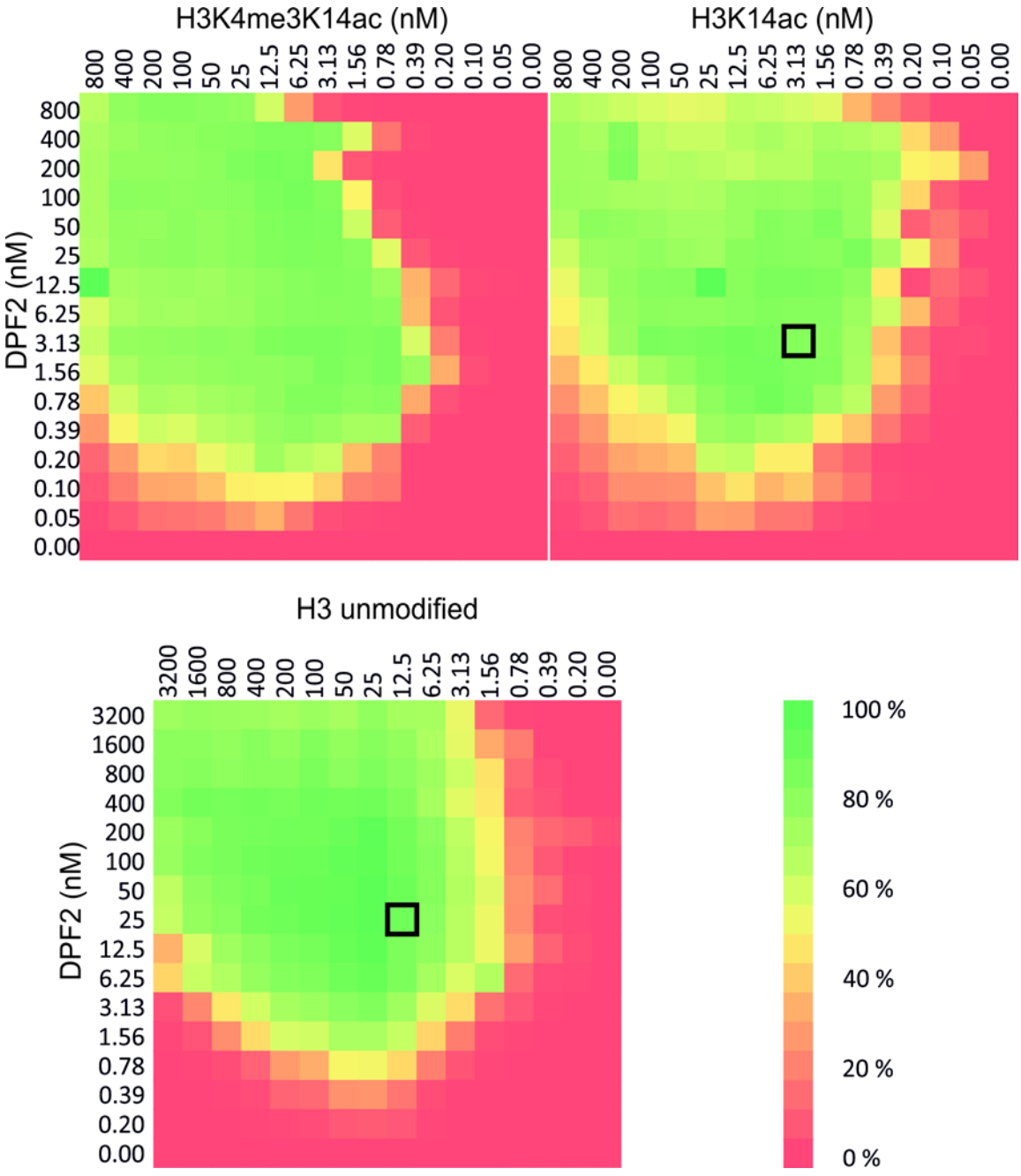


Figure 70. Heat maps showing AlphaScreen signal for varying concentrations of DPF2 and partner peptide. The concentrations of H3K14ac and H3 unmodified used in assays are highlighted with a black square.

It is necessary to identify optimal concentrations of protein and peptide for use in the AlphaScreen assay. Screening hits cause a loss of signal in an AlphaScreen assay. Therefore protein and peptide concentrations must be high enough to generate a background signal with a good signal to noise ratio. However, if the protein and peptide concentrations are too high the

AlphaScreen beads become saturated, leaving free protein in solution. Free protein may outcompete bead-bound protein for peptide binding and vice-versa in the case of excess peptide, reducing the assay signal. Therefore it is important to identify a suitable intermediate concentration for using in the assay.

A protein-peptide dose response test of multiple protein and peptide concentrations revealed that the for a H3K14ac peptide 3.13 nM of protein and peptide were suitable. For an unmodified H3 peptide, 12.5 nM peptide and 25 nM protein were optimum concentrations (Figure 70). This suggests that DPF2 has a slightly higher affinity for acetylated K14, in contrast to the initial BLI experiment. It was noted that in the BLI peptide screen peptides with a trimethylation of K4 (H3K4) had no effect on binding to DPF2. Therefore a H3K4me3K14ac peptide was also tested in the dose response test. This peptide showed similar behaviour to H3K14ac peptide, indication that H3 modification has little impact on binding to DPF2. This is unusual, as the interaction between a PHD and H3 is usually strongly dependent on the methylation state of K4.

DMSO Tolerance

It is necessary to identify the DMSO tolerance of the assay, as intolerance to DMSO can limit the concentrations of ligands in the assay. This is particularly true in the case of protein-protein interactions, are primary hits are likely to be weak, and may be missed if screened at too low a concentration. To this end, the AlphaScreen assay was ran with varying concentrations of DMSO in the buffer from 0.5% to 5.0% in increments of 0.5%. This revealed that there was no drop in signal as DMSO concentration increased. Therefore the assay can be used with up to 5% DMSO without loss of signal caused by DMSO induced protein denaturing.

Fragment Screening

Although the AlphaScreen assay described above would prove suitable to high throughput screening (HTS) of a large, highly diverse library, the high costs of the AlphaScreen reagents to screen more than 1000 compounds are prohibitive. Therefore careful library design is required

to maximise the chances of finding a primary hit suitable for optimisation and make best use of available resources. This section describes the design of libraries that were screened against DPF2 using the AlphaScreen assay described above.

Fragment Screening

Although tandem PHDs are hypothesised to be more ligandable than single PHDs, protein-protein interactions (PPIs) such as those observed between tandem PHDs and histone tails are regarded as being difficult targets for small molecule inhibition.[152,153]

Fragment based drug design is has been suggested as a method for tackling protein-protein interactions.[154–156] Although there is no firm definition of what constitutes a fragment, fragments are generally small, simple molecules which are likely to make weak but efficient interactions with protein targets. In 2003 Congreve et al. published an observation that fragments hits tend to obey a 'Rule of Three' where molecular weight is < 300, the number of hydrogen bond donors and acceptors are both ≤ 3, and clogP ≤ 3.[157] These rules are frequently challenged, and current opinion holds that a fragment contains less than seventeen heavy atoms.[158]

Fragment based ligand design is thought to be particularly applicable to PPIs due to their ability to search a larger area of chemical space.[159] This is based on the theory that a plot of number of heavy atoms versus chemical space (i.e. the total number of theoretically possible molecules) shows a massive increase in size for the addition of each heavy atom.[160] Therefore a 1000 member fragment library represents a far higher proportion of the available chemical space for molecules that weigh less than 300 Daltons, than a 1 million member library made up of compound weighing up to 500 Daltons.

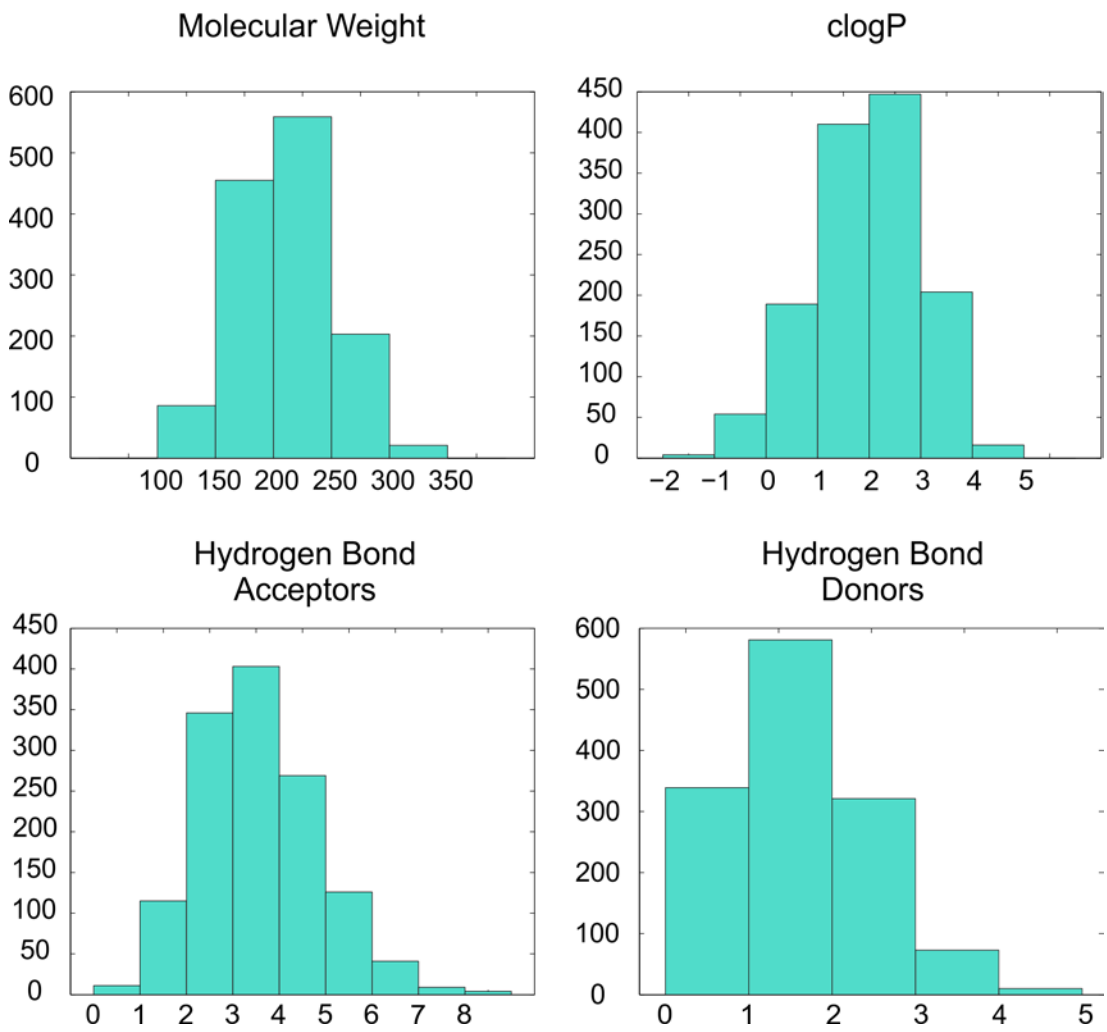


Figure 71. Histograms showing the distribution of molecular properties of the 1324 fragments that were screened against DPF2. These distributions show that the library is broadly in line with the 'Rule of Three'. All molecular properties were calculated using Molsoft ICM.

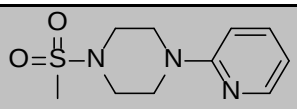
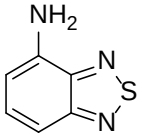
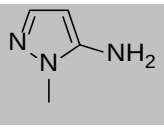
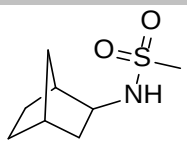
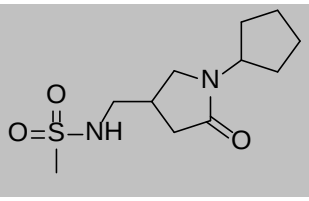
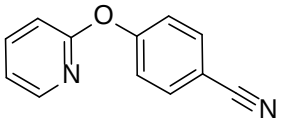
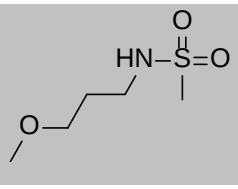
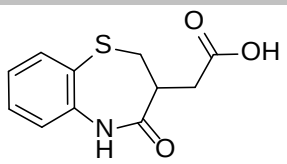
In total 1324 fragments were identified that could be screened against DPF2 at 2 mM using the AlphaScreen assay designed above. This consisted of 926 compounds from a Maybridge fragment library and 398 compounds from the 3D Fragment Consortium library.[161]

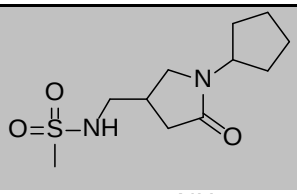
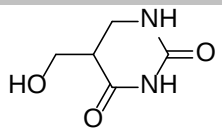
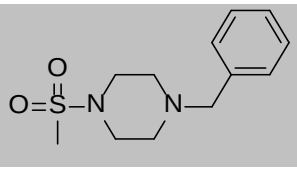
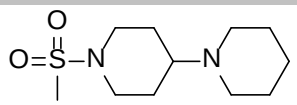
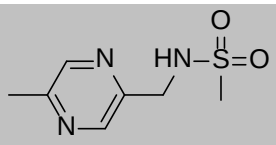
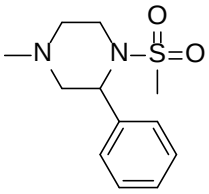
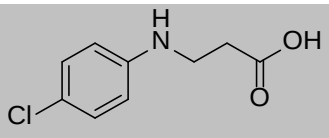
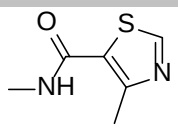
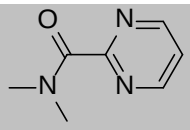
Analysis of this library shows that most have a molecular weight between 150 and 250 Daltons, and clogP in the range 0-3. The majority had between 2-4 hydrogen bond acceptors and < 3 hydrogen bond donors (Figure 71). This means that the fragment collection used broadly follows the 'Rule of Three'[157] but is not constrained by it.

Results of Fragment Screen

All fragments were screened in duplicate and results normalised against DMSO and water controls. Compounds selected for IC_{50} measurement showed activity in the DPF2 assay that was greater than 65 percentage-points higher than their activity in the counter-screen assay.

Fragment Hits

Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)
9		68 µM	79 µM	58 µM
10		87 µM	110 µM	69 µM
11		110 µM	160 µM	73 µM
12		110 µM	130 µM	93 µM
13		130 µM	150 µM	110 µM
14		130 µM	210 µM	78 µM
15		140 µM	170 µM	110 µM
16		190 µM	250 µM	140 µM

Compound Number	**Structure**	**IC_{50} (μM)**	**IC_{50} Upper 95% Confidence Limit (μM)**	**IC_{50} Lower 95% Confidence Limit (μM)**
17		190 μM	240 μM	150 μM
18		220 μM	280 μM	180 μM
19		260 μM	340 μM	200 μM
20		260 μM	310 μM	220 μM
21		290 μM	440 μM	190 μM
22		320 μM	400 μM	250 μM
23		440 μM	560 μM	350 μM
24		480 μM	670 μM	340 μM
25		510 μM	630 μM	410 μM

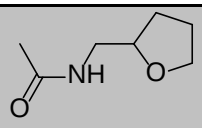
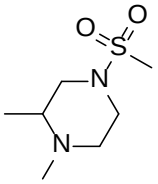
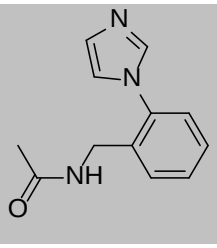
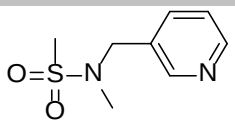
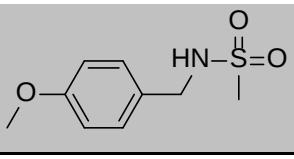
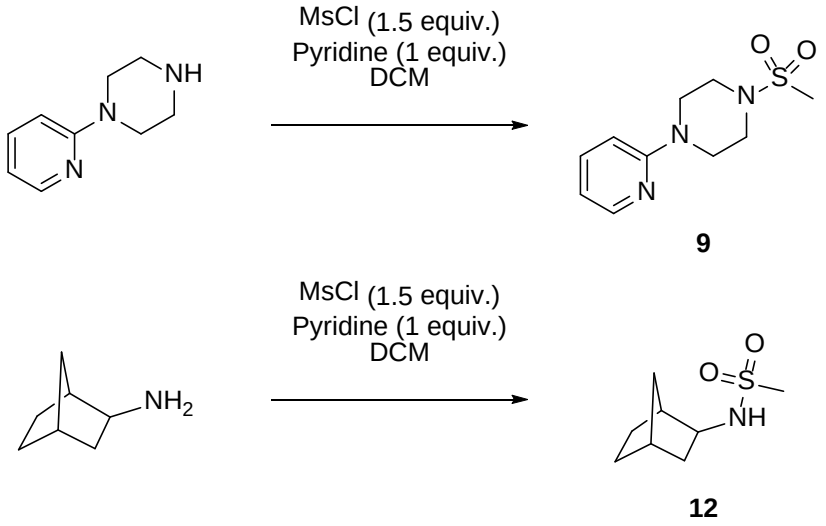
Compound Number	Structure	IC_{50} (μM)	IC_{50} Upper 95% Confidence Limit (μM)	IC_{50} Lower 95% Confidence Limit (μM)
26		540 μM	860 μM	340 μM
27		550 μM	710 μM	430 μM
28		760 μM	1200 μM	490 μM
29		790 μM	1100 μM	590 μM
30		920 μM	1300 μM	640 μM

Table 12. IC_{50} values were successfully measured for twenty-two fragments ranging from 68 μM to 920 μM.

IC_{50} values were successfully measured for twenty-two fragments, ranging from 68 μM to 920 μM (Table 12). This is < 2% of the fragments screened. Compounds **11** and **14** were not considered for further optimisation as they also appeared as hits in a fragment screen carried out against the PHD-JmjC of PHF8 (Chapter 5) and several other SGC targets and were assumed to be false positives.



Scheme 1. The resynthesis of *N*-mesyl compounds **9** and **12** was achieved from commercially available amines using methanesulfonyl chloride (MsCl) and pyridine.

Amongst these twenty-two hits were twelve compounds that contain a *N*-mesyl moiety. The diversity of the molecules to which the *N*-mesyl group was attached suggested that it was functional group that is making a key interaction with the protein surface and inhibiting peptide binding. In order to confirm this, the most potent *N*-mesyl compounds **9** and **12**, were resynthesised and retested (Scheme 1) for confirmation.

The resynthesised compounds did not show any activity when tested using the AlphaScreen assay, despite identical ^{1}H NMR, ^{13}C NMR, and LC/MS data. As the original hits were all contributed to the 3D Fragment Consortium library from the same laboratory, it is possible that they all contained a common contaminant that interfered with the protein or assay. For example, a protein denaturing contaminant would appear to be active in the AlphaScreen assay but not in the counter-screen assay where no protein is present.

Design of a Library Focused by Virtual Screening

In addition to the fragments, larger molecules were also selected for screening. A 10,000 member compound library was made available for use by a collaborator (Daniel Ebner; Target Discovery Institute, University of Oxford). Although the AlphaScreen described above would be suitable for screening such a library in its entirety, the cost of such a screen was prohibitive. Therefore a virtual screen was used to prioritise compounds from this library for experimental screening. This section provides a short overview of virtual screening and the software available, followed by a description of the virtual screen performed on the tandem PHDs of MYST3 as a surrogate for DPF2.

Overview of Virtual Screening

There are two main strategies used in virtual screening, ligand based screening and structure based screening. A ligand-based screen takes a known ligand of the target protein and attempts to identify other ligands similar to the known active ligand.[162] Due to the absence of known small molecule ligands of tandem PHDs this method could not be used for this work. Structure based approaches attempt to find ligands that best fit into a pocket on a protein structure. For this reason, structure-based virtual screening requires a high resolution structure of the target protein.

Structure based virtual screening can either treat the given protein structure as rigid, or allow a certain amount of protein flexibly allowing for the discovery of induced fit ligands. The latter method is computationally more expensive and therefore less suitable for high-throughput virtual screening. For this work the rigid structure based docking programme Glide was used.[163–165]

Validation of Virtual Screening Methods

The suitability of Glide for discovering inhibitors of protein-protein interaction binding sites was tested prior to the performance of the virtual screen.

To test Glide, a library of 346 compounds designed by Stephen Frye and co-workers to be methyl lysine mimetics was virtually screened against the triple Tudor domains of SPIN1. These compounds had previously been screened at the Structural Genomics Centre, University of Oxford; and active compounds confirmed. The aim of this work was to compare the results of the virtual screen performed on SPIN1 with previously performed experimental screen. If confirmed active compounds were identified as hits in the virtual screen, this would give confidence that the virtual screening methodology used to design libraries for DPF2 and PHF8 (Chapter 5) was valid.

Virtual Screening of SPIN1

Two crystal structures of SPIN1 are available in the Protein Data Bank: an *apo* structure (PDB ID: 2NS2)[166] and a *holo* structure with peptide ligand representing residues 1-8 of H3 with a H3K4me3 mark (PDB ID: 4H75).[136] The trimethylated lysine is bound to an aromatic cage in the second of the three Tudors of SPIN1. Comparison of the structures suggests that the aromatic cage that engages H3K4me3 showed slight differences between the *apo* and *holo* structures. Tryptophan 151 is moved by about 45° between the two structures (Figure 72). In the case of the *apo* structure, this residue is pointed in towards the centre of the aromatic cage, suggesting that this residue moves on H3K4me3 recognition to complete aromatic cage formation. For this reason the *holo* structure was chosen for virtual screening.

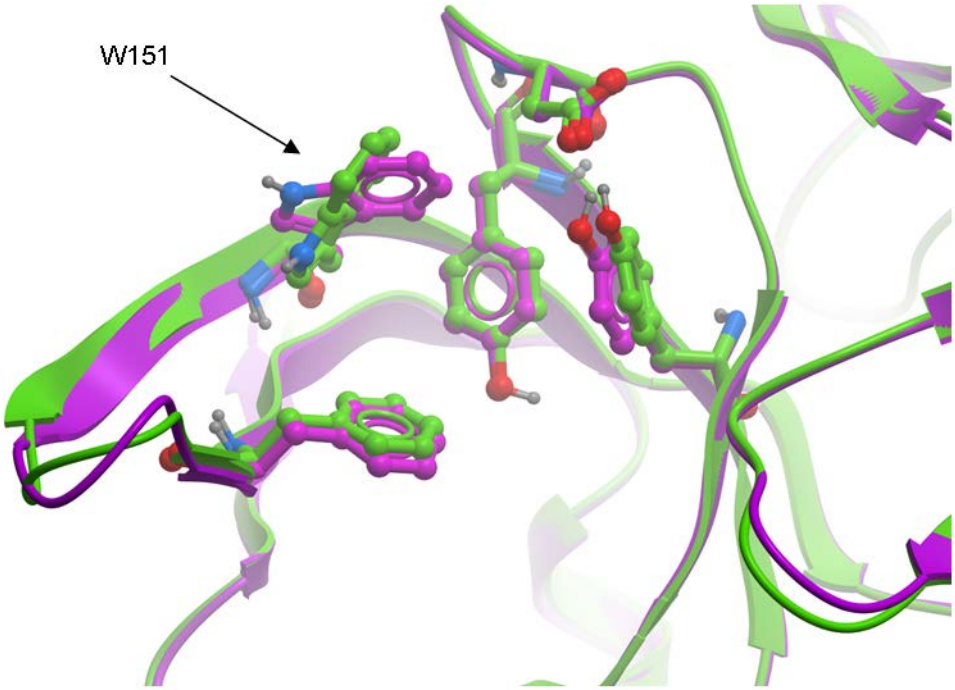


Figure 72. Comparison of the position of aromatic cage residues of the second Tudor domain of SPIN1 between *apo* and *holo* structures. Tryptophan 151 is rotated by 45° in the *apo* structure (magenta, PDB ID: 2NS2) compared to the *holo* structure (green, PDB ID: 4H75). This suggests that tryptophan 151 moves on binding to H3K4me3 to accommodate the trimethyl lysine in the aromatic cage.

The structure of SPIN1 was passed through the Schrödinger Protein Preparation Wizard prior to virtual screening. Protonation states were assigned at pH = 7 and *H*-bonds optimised. All water molecules were removed, as were buffer and salt molecules. A loop between the second and third Tudor domains consisting of sixteen residues was not resolved in the crystal structure, but this loop was not near the binding site, and no efforts were made to model this loop prior to virtual screening. The peptide ligand was used to define the boundary box of the binding site.

Comparison of the Results of Virtual and Experimental Screens

Of the 346 compounds, twenty-four were deemed to be hits in the experimental screen and had been progressed to IC_{50} measurement. These hits had IC_{50} values in the range of 2-10 µM.

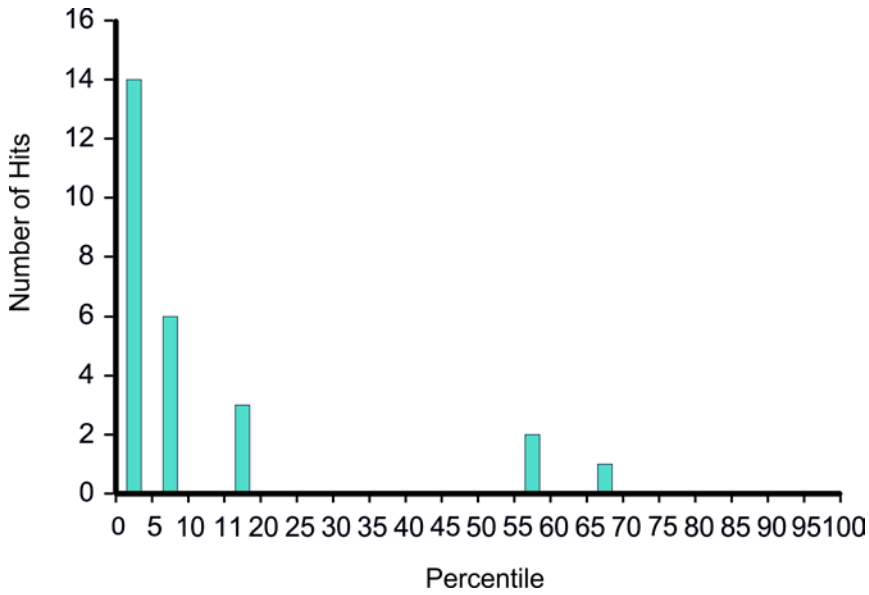


Figure 73. Histogram showing the clustering of confirmed hits within the ranking derived from virtual screening. Fourteen of the twenty-four confirm hits are in the top 5% as ranked by virtual screening. This provides partial validation of the virtual screening methods used in this work.

When the compounds were ranked by GlideScore results, fourteen of the confirmed hits were found in the top 5% of the ranking. This suggested that had this virtual screen been used to prioritise the library prior to experimental screening, the majority of confirmed hits would still have been identified. The results for indicate that Glide is suitable for prioritising compounds for experimental screening.

Virtual Screen Using MYST3

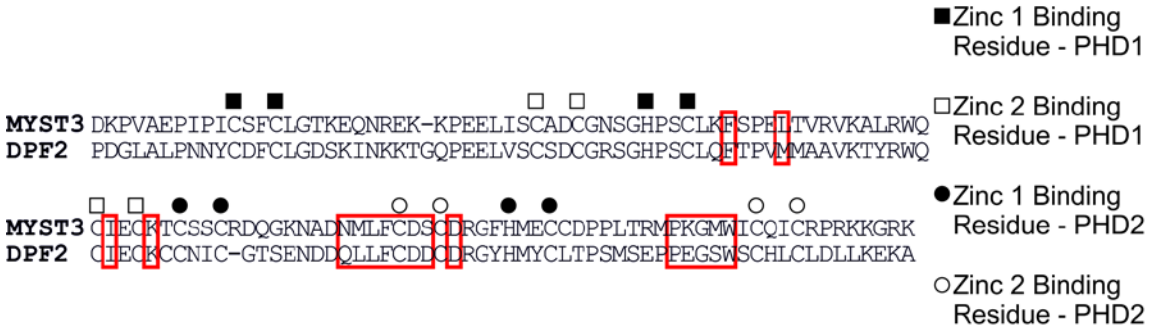


Figure 74. Sequence alignment of the tandem PHDs of MYST3 and DPF2. Residues that form part of the peptide binding site are highlighted with red boxes.

The aim was to prioritise the 10,000 member screening library into a smaller library of approximately 500 compounds that could be tested using the DPF2 AlphaScreen assay described

above. The tandem PHDs of MYST3 were used as the receptor for this structure based virtual screen, as this structure (PDB ID: 3V43) was the only available crystal structure of a tandem PHD at the time the work was performed (further structures of MYST3 have since been deposited in the Protein Data Bank). A solution NMR structure of the tandem PHDs of DPF3b was also available, but was not used for this work as it was felt that using a crystal structure – albeit of a more distantly related protein (Figure 65) – would produce higher quality results. An analysis of the binding site residues suggests that MYST3 is a suitable virtual screening surrogate for DPF2 due to their close homology (Figure 74).

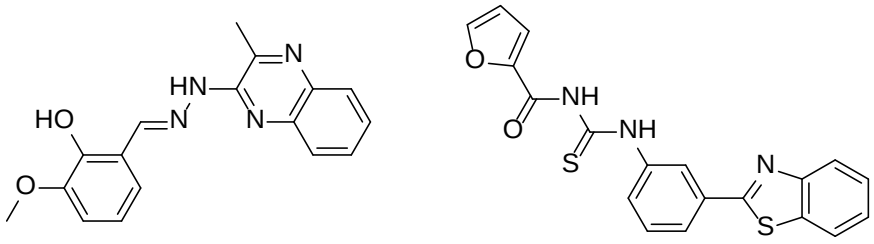
Prior to performance of the virtual screen the structure was passed through the Schrödinger Protein Preparation Wizard.[131] The bound peptide ligand and all water molecules were deleted, and missing side chains inspected to see if they were likely to affect the results. This inspection revealed that the missing side chains in the structure used were not near the peptide binding site, and therefore were not modelled prior to virtual screening.

The peptide ligand from the original crystal structure was used to define the boundary box for receptor generation. This box was expanded to a cube with sides of 25 Å to allow for a full exploration of the histone binding surface of the tandem PHD.

The 10,000 member compound library was filtered using the Schrödinger Rapid Elimination Of Swill (REOS)[167] filter which reduced the library size to 8,783. This filter removed compounds that contain reactive groups such as peroxides, and sulphonyl halides, as well as compounds with undesirable molecular properties, such as a clogP values less than -5. Three dimensional conformations and protonation states were assigned for the remaining ligand using Schrödinger's LigPrep.[168]

Results of Virtual Screening

The results of the virtual screen reflected the difficulty of discovering small molecules that bind to protein-protein interaction surfaces. Glide produce a GlideScore for each docked pose, with a more negative score corresponding to a more potent ligand. Schrödinger's online Knowledge Base states that "scores of -10 or lower usually represent good binding. For some targets, (e.g. with shallow active sites or predominantly hydrophobic interactions), scores of -8 or -9 might be very good." In this case, the best scoring molecule had a GlideScore of -7.5.



31 **32**

Figure 75. Compounds **31** and **32** were removed from the screening library following manual inspection owing to the presence of functional groups that are susceptible to hydrolysis under assay conditions. Compound **31** contains a potentially hydrolysable hydrazone moiety, whereas the acyl-thiourea of compound **32** is potentially hydrolysable under assay conditions.

The top 500 poses were filtered to remove duplicate compounds, leaving a remaining 463 compounds. These compounds were manually inspected, and a further twenty-six compounds with undesirable functional groups such as hydrazone **31** and acyl-thiourea **32** were removed (Figure 75). This left 437 unique compounds to be screened against DPF2.

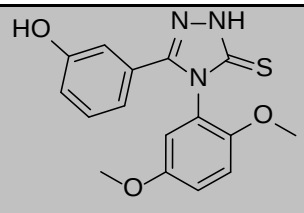
Virtual Screening Using Molecular Dynamics

A further set of virtual screens were carried out with a collaborator (Jan Domanski; Department of Biochemistry, University of Oxford). This work used the same crystal structure of MYST3 in the virtual screen described above; however, in this instance the structure was subjected to molecular dynamics (MD) simulations. The MD simulation revealed dynamic pockets on the

histone binding surface that were not seen in the static crystal structure. These structures were extracted from the MD simulation trajectory and used for a virtual screen similar to the one described above. As above, a REOS filtered version of the 10,000 member library was used for the virtual screen. As a result of this work, the author was provided with two libraries: a 398 member library as a result of using Glide in standard precision (SP) mode, and a 383 member library as a result of using Glide in extra precision (XP) mode.[165] These libraries did not overlap with each other, or with the virtual screening library designed by the author described above. This was due to the removal of overlapping compounds prior to the start of the virtual screen.

Experimental Results for Virtual Screening Library

1244 molecules from the three docking libraries were screened in duplicate at 100 µM against DPF2. Compounds that showed greater than 80% inhibition in the DPF2 assay and less than 20% in the counter-screen assay were selected for IC_{50} measurement. Compounds that showed greater than 60% inhibition in the DPF2 assay with less than 5% inhibition in the counter-screen assay were also selected. IC_{50} measurement was successful for six compounds, three from the virtual screening derived library designed by the author, and three from the libraries designed by Jan Domanski (Table 13). All six compounds showed no impurities when tested by LC/MS.

Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)	Library
33		11 µM	13 µM	8.6 µM	Author's Library

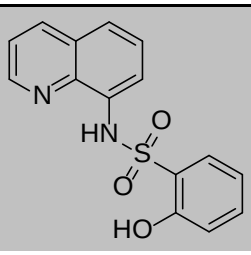
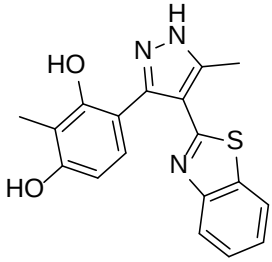
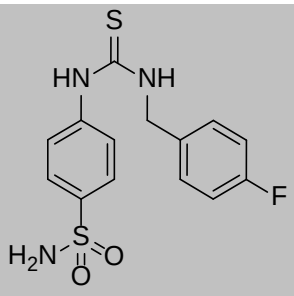
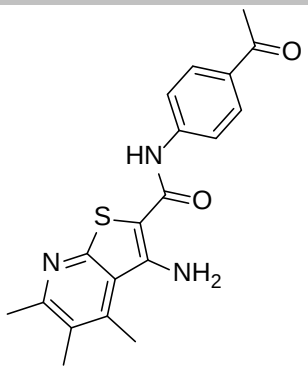
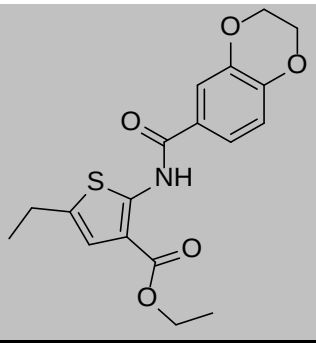
Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)	Library
34		10 µM	13 µM	8.3 µM	Author's Library
35		11 µM	16 µM	8.0 µM	Author's Library
36		32 µM	43 µM	23 µM	Domanski SP Library
37		28 µM	39 µM	20 µM	Domanski SP Library
38		56 µM	70 µM	46 µM	Domanski XP Library

Table 13. IC_{50} values were measured for six compounds which originated from virtual screening derived libraries.

Before investigating these compounds further, an informatics based promiscuity filter was applied to the hits. A thorough literature search was carried out to see if these scaffolds have been published as hits from other screening campaigns. This would indicate that they were likely to be pan-assay interference compounds (PAINs) rather than genuine hits.[150]

A search for compounds containing the core scaffold of 5-methylpyrazole **35** revealed that compounds featuring a 2-(5-methyl-3-phenyl-1*H*-pyrazol-4-yl)benzo[d]thiazole core appear in several patents. These patents result from screening efforts performed by academic groups for several unrelated targets. It is therefore likely that compound **35** is a PAIN compound rather than a genuine hit. Compounds **37** and **38** were also prevalent in ChEMBL records suggesting that they hit multiple targets.[169]

8-Aminoquinolines

8-aminoquiniline **34** shares the same structural core as known Zn(II) chelators 6-Methoxy-(8-*p*-toluenesulfonamido)quinolone (TSQ) and Zinquin. According to Fahrni and O'Halloran "these probes are suggested to remove Zn(II) from tightly bound sites in proteins."[170] It is therefore possible that 8-aminoquiniline **34** sequesters the Zn(II) from DPF2 leaving the domain structurally unstable and inactive.

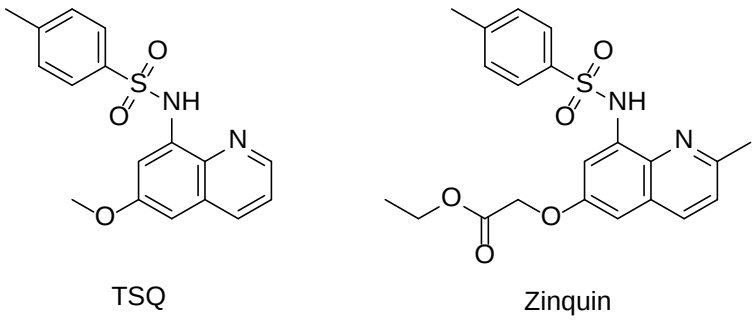
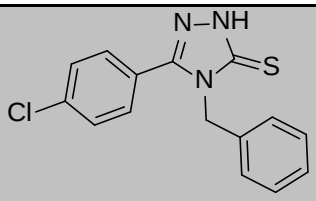
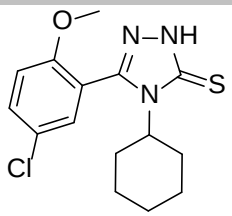
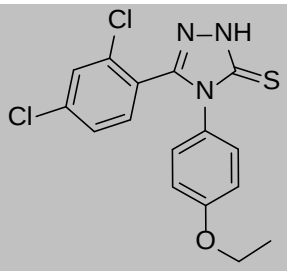
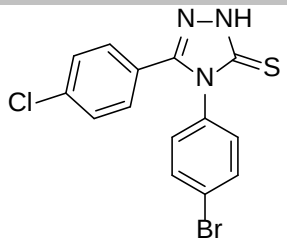
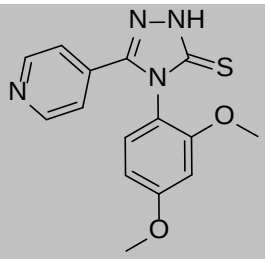
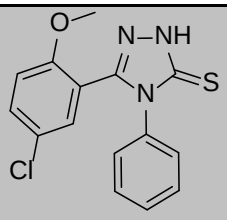
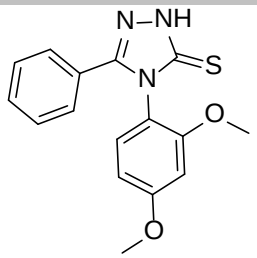
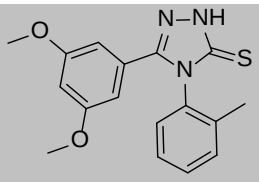
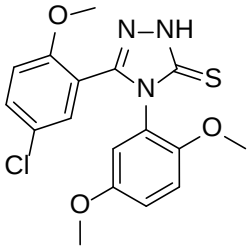
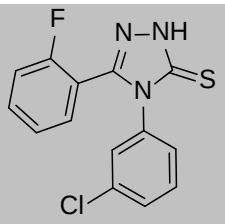
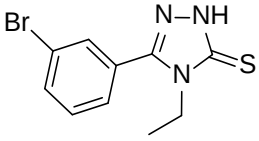


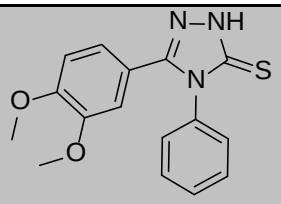
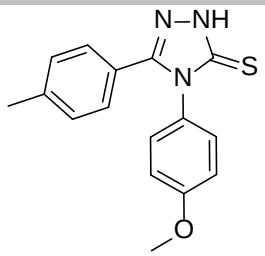
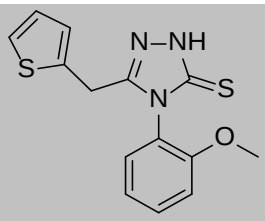
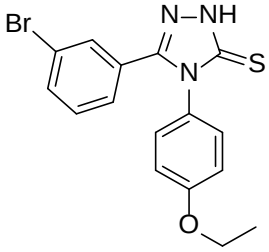
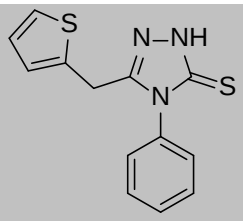
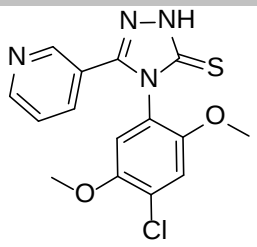
Figure 76. Known Zn(II) chelators 6-Methoxy-(8-*p*-toluenesulfonamido)quinolone (TSQ) and Zinquin share the same 8-aminoquinilone core as compound **34.**

3-Mercapto-1,2,4-Triazole Series

Compound **33** was progressed to analogue screening. A search of the 10,000-member library from which it originated found nineteen analogues for follow up screening (Table 14).

Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)
39		13 µM	17 µM	10 µM
40		20 µM	23 µM	17 µM
41		24 µM	49 µM	11 µM
42		26 µM	34 µM	20 µM
43		26 µM	36 µM	19 µM

Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)
44		27 µM	34 µM	22 µM
45		30 µM	43 µM	21 µM
46		30 µM	42 µM	22 µM
47		41 µM	56 µM	30 µM
48		46 µM	97 µM	21 µM
49		75 µM	230 µM	24 µM

Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)
50		92 µM	140 µM	62 µM
51		97 µM	550 µM	17 µM
52		120 µM	180 µM	70 µM
53		> 200 µM		
54		> 200 µM		
55		> 200 µM		

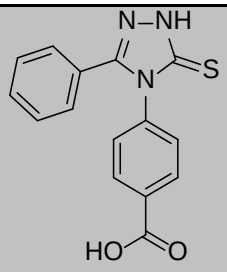
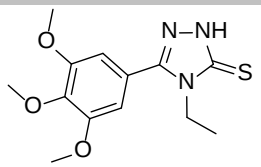
Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)
56		> 200 µM		
57		> 200 µM		

Table 14. Analogues of compound **33** that were screened in an attempt to discover more potent DPF2 inhibitors and establish structure activity relationship (SAR).

All analogues screened are less potent than the original hit, compound **33**. The series also shows no strong structure activity relationships (SAR). Compound **33** was tested by isothermal calorimetry, but no binding to DPF2 could be detected. Work on this series was therefore halted, as no direct binding could be established and no SAR identified, leading to the assumption that they are assay interference compounds.

Thiourea Series

Compound **36** was considered a promising candidate as it contained no structural alerts and no records of it being a hit for other targets. In order to probe SAR three analogues were purchased and tested using the AlphaScreen assay (Table 15). The substitution of the fluorine of compound **36** with a chlorine results in a three-fold increase in activity. The removal of the primary sulphonamide results in a loss of inhibition.

On the basis of these preliminary results, a further set of analogues were synthesised to establish SAR. The synthesis involved the condensation of a benzylamine with an isothiocyanate

(Scheme 2). The simplicity of this method facilitated the production of twenty-five analogues for screening. However, none of the synthesised analogues showed any activity, including resynthesised versions of compounds **36**, **58**, and **59**.

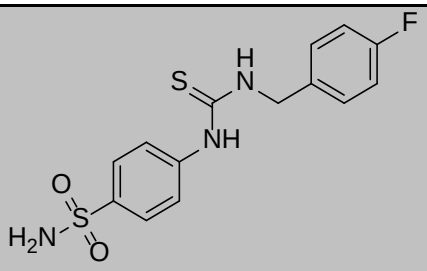
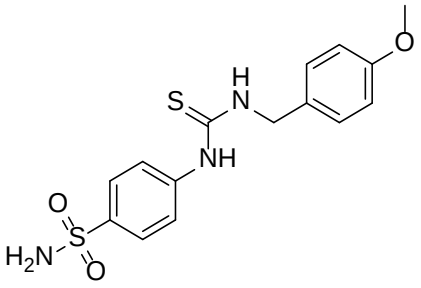
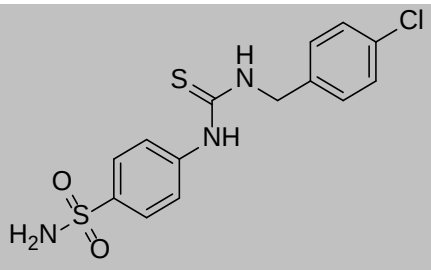
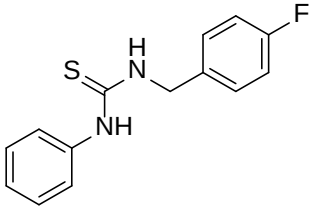
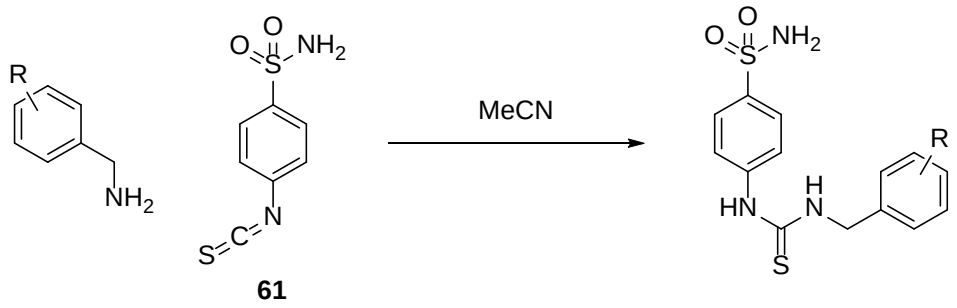
Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit (µM)	IC_{50} Lower 95% Confidence Limit (µM)
36		32 µM	43 µM	23 µM
58		39 µM	46 µM	33 µM
59		9.3 µM	11 µM	7.6 µM
60		No inhibition		

Table 15. Analogues of compound **36** that were purchased for testing. The inactivity of compound **60** suggests that the primary sulphonamide is important for activity. The substitution of fluorine with chlorine gives a three-fold increase in activity in the AlphaScreen assay.



Scheme 2. The synthesis of analogues of thiourea **36** was achieved with a range of commercially available amines and isothiocyanate **61**.

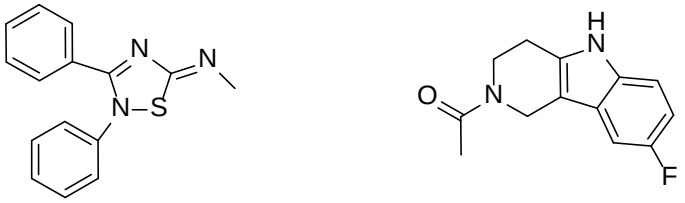
Original batches of compounds **36**, **58**, and **59** were submitted to co-crystallisation trials with a construct of DPF3b that was known to be crystallisable (below). Although crystals could be grown, analysis of the diffraction pattern showed no presence of a bound ligand. Following these negative results, work on this series was halted.

Acetyl Lysine Mimetic Library

As DPF2 is a putative acetyl lysine binding domain, a library of compounds designed to mimic acetyl lysine was identified for screening. The library consists of 967 compounds previously selected for screening against members of the bromodomain family. It was hypothesised that this acetyl lysine mimetics may act as DPF2 inhibitors, as its close homologue DPF3b has been shown to be an acetyl lysine binder.[132]

Acetyl Lysine Mimetic Library Hits

These compounds were screened at 500 µM. Compounds that showed greater than 50% inhibition in the DPF2 assay, less than 50% inhibition in the counter-screen assay, and with DPF2 inhibition greater than twice the counter screen inhibition were progressed to IC50 measurement. In total IC_{50} values were obtained for twenty-three compounds



62
0.5 µM

63
20.2 µM

Figure 77. The two most potent compounds identified from the acetyl lysine mimetic library. Compound **62** is known assay interference molecules having been identified as false positives in screens carried out against members of the bromodomain family.

Compounds **62** was the most potent inhibitor (Figure 77), but was immediately discarded as it has been known to appear as false positive in numerous screens against members of the bromodomain family. This false activity is thought to be due to protein reactivity caused by the labile N-S bond present in the molecule. Compound **63** was tested via isothermal calorimetry, but showed no evidence of binding.

Design of Secondary Assays to Confirm Primary Hits

A series of secondary assays were required to assure that compounds identified by AlphaScreen were genuine inhibitors and not assay interference compounds. The development of assays that show direct binding between protein and ligand were prioritised as they were less likely to give false positives than assays with indirect reporting mechanisms.

Biolayer Interferometry

A BLI assay for the identification of suitable DPF2 peptide substrates for an AlphaScreen assay was described above. The BLI experiment described involved peptides immobilised on the tip of the optical fibre with the protein in solution. For the detection of compound binding, an assay with protein immobilised on the tip of the optical fibre and compound in solution was required.

Chemical Biotinylation

The first requirement of such an assay was the production of biotinylated DPF2 which can be immobilised on streptavidin coated tips. For this purpose an ImmunoProbe Biotinylation Kit (Sigma-Aldrich) was used to chemically biotinylate DPF2. This used an *N*-hydroxy succinimide ester of biotin designed to react with lysine (Figure 78).

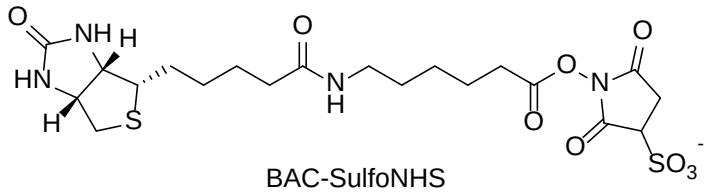


Figure 78. Activated *N*-hydroxy succinimide ester of biotin with an aminocaproic linker used for biotinylation of lysine residues.

Recombinant DPF2 which had the His6-GST tag removed was incubated with the biotinylating agent at 4 °C for 2.5 h. The mixture was purified by size exclusion chromatography to give a mixture of unmodified, mono-, di-, and tri-biotinylated species (Figure 79).

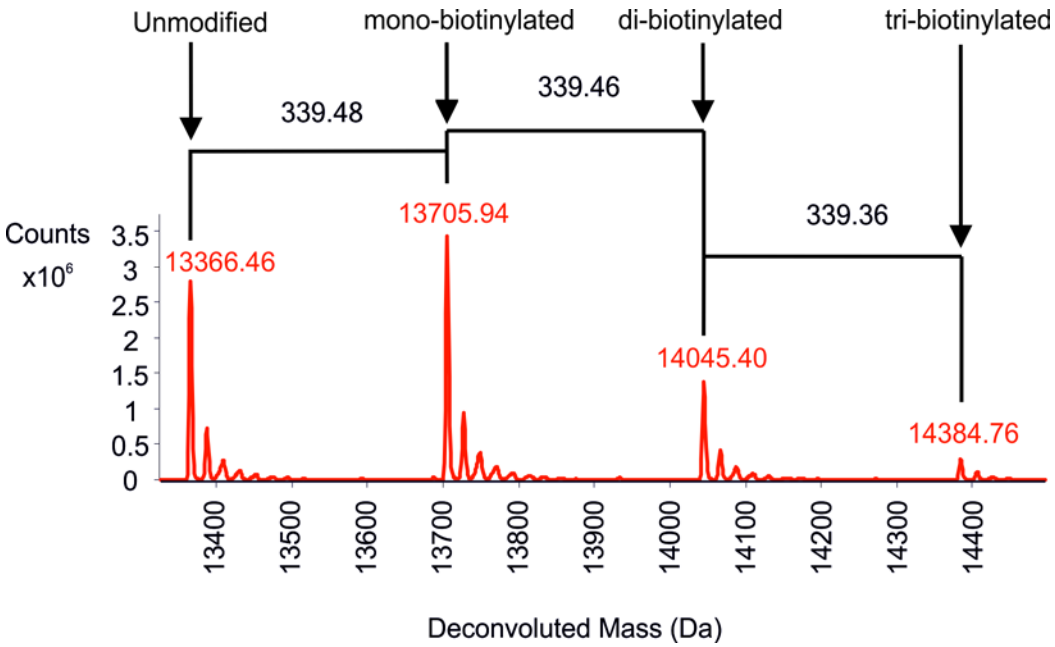


Figure 79. Deconvoluted mass spectrometry data for biotinylated DPF2. The DPF2 is present as an unmodified species, as well as the mono-, di-, and tri-biotinylated species, with the unmodified and mono-biotinylated species being dominant.

The K_D of biotinylated DPF2 with unmodified H3 was measured using BLI with the peptide immobilised. This was compared to that of unmodified DPF2 and unmodified H3. Both were in the range of 1-3 μM indicating that biotinylation had not affected the substrate binding site.

Biotinylated DPF2 was immobilised on the optical fibre tip and binding to an unmodified H3 peptide confirmed as a possible control; however, it was not possible to show any binding to any small molecule compound. This was because the high concentrations required to measure a full dose response curve for fragment hits caused non-specific binding to the streptavidin labelled tip, which meant specific binding to the immobilised DPF2 could not be identified.

Nuclear Magnetic Resonance

As the tandem PHD construct of DPF2 constitutes a small protein domain (< 14 kDa) it was a suitable candidate for the development of a nuclear magnetic resonance (NMR) binding assay. Work was done to develop a production method for ^{15}N-labelled DPF2 which could be used in a protein observed NMR binding assay. Such as assay would have provided direct evidence of binding interactions between a small molecule inhibitor and DPF2 and allowed the assignment of which residues are involved.

Labelled protein production expression trials were carried out in collaboration with Ivan Leung (then of the Department of Chemistry, University of Oxford; now of School of Chemical Sciences, University of Auckland). Three separate techniques were used in labelled protein test expressions. The first two test expressions were performed by the author, the third by Ivan Leung.

The first method used a minimal media consisting of $^{15}NH_4Cl$, glucose, and phosphate buffer. This media is supplemented with trace metals and vitamin B1. Although it was possible to grow *E. coli* cells which had previously been transformed with a plasmid vector encoding for DPF2 in these conditions, no expression of DPF2 was observed. The second method used a protocol described by Marley et al.[171] Cell mass was grown in unlabelled, rich media before being

harvested by centrifugation. The cell mass was then washed and resuspended in minimal media prior to the induction of protein expression by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG). This method produced a greater volume of cell mass than using minimal media alone, but no expression of DPF2 was observed.

The final method was performed by Ivan Leung. This method tested used the minimal media described above supplemented with ^{15}N-labelled rich media.[172] As with the previous methods, no expression of DPF2 was observed.

Work towards an NMR binding assay was paused following these unsuccessful attempts to express ^{15}N-labelled DPF2. It is possible that expression may be possible with use of alternative DPF2 constructs, plasmid vectors, or cell types.

Crystal Soaking

X-ray crystallography can offer a high standard of proof of binding of a ligand to a protein. A suitable crystal of a protein-ligand can be achieved either using a co-crystallisation method or crystal soaking. Co-crystallisation requires both protein and ligand to be present during crystallisation. This is a very low throughput method as the conditions (e.g. temperature, pH, buffer) required to induce crystallisation may vary for each ligand. Crystal soaking is a higher throughput method which can be used for testing multiple ligands. Multiple crystals of the protein in its *apo* form are produced and then soaked in solutions containing the ligands to be tested (Figure 80). As protein crystals often contain large solvent channels small organic molecules are free to diffuse through the crystal. Active compounds will bind the protein and can be detected by X-ray crystallography. This approach has been used within the SGC to discover fragment inhibitors of the bromodomain of ATPase family AAA domain containing 2 (ATAD2).[173]

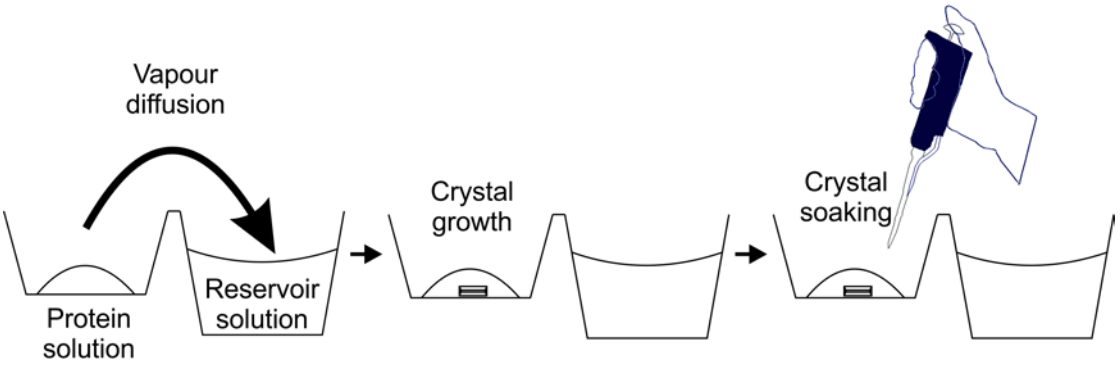


Figure 80. An overview of the procedure of crystal growth followed by crystal soaking. Crystals are grown using a sitting drop vapour diffusion method. Following crystal growth, the compound to be soaked is added directly to the crystallisation buffer.

A crystal soaking assay for DPF2 was sought to allow confirmation of primary hits and to provide structural data to inform analogue synthesis for optimisation of the primary hit. Such an assay could also be used to rescreen fragment libraries as an alternative primary screen.

The His6-GST tag used in the AlphaScreen assay was removed by incubation with TEV protease. The untagged tandem PHDs were submitted to a range crystallisation screens at both 4 °C and 20 °C. Despite the range buffers and precipitant tested across several pH's, no conditions could be identified that induced crystallisation of DPF2. These crystallisation trials were repeated with DPF2 with the GST tag intact. The use of a GST tag can improve protein solubility,[174] allowing crystallisation trials to be carried out at higher protein concentrations. Improved solubility may also prevent undesirable protein precipitation during crystallisation trials. However, despite the improved solubility of GST-tagged DPF2 compared to untagged DPF2, no crystals of GST-tagged DPF2 could be obtained.

Owing to the difficulty of crystallisation of DPF2, other members of the tandem PHD family were considered for use in a crystal soaking assay. Colleagues at the Structural Genomics Consortium, University of Toronto have obtained crystals of the tandem PHDs of DPF3b that allow the solution of their structure to be determined at a high resolution (1.57 Å).

Due to the close relationship of the tandem PHDs of DPF2 and DPF3b (79% sequence identity), it was decided that a crystal soaking assay using DPF3b would be a suitable follow up assay for primary hits identified for DPF2.

Protein production

A plasmid vector encoding for the crystallisable construct of DPF3b was donated by the group of Jinrong Min (Structural Genomics Consortium, University of Toronto). The plasmid encoded for *N*-terminally His6-tagged protein, with a TEV protease cleavage site between the tag and the tandem PHDs of DPF3b. This plasmid was transformed into Rosetta *E. coli* cells and grown in LB media, with expression of DPF3b induced by the addition of IPTG. The cells were harvested, lysed, and the recombinant DPF3b extracted using a Ni-NTA column. The resulting DPF3b was incubated with TEV protease and the cleaved His6 removed using a Ni-NTA column. The tandem PHDs of DPF3b were further purified using size exclusion chromatography prior to use in crystallisation trials (Figure 81).

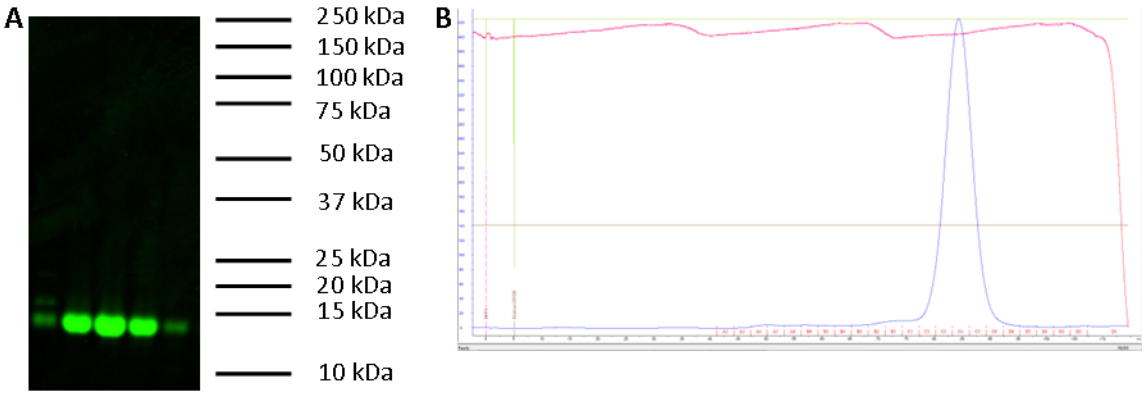


Figure 81. A. SDS-PAGE stained with Coomassie Brilliant Blue and imaged using a Li-Cor Odyssey CLx showing the purification of the tandem PHDs of DPF3b. **B.** Size exclusion chromatography trace of DPF3b.

Coarse Screens and Follow Ups on Toronto Conditions

Conditions for the crystallisation of DPF3b had previously been established at SGC Toronto. However, the protein produced for the purpose of this work was not expected to crystallise in

identical conditions owing to differences in expression and purification methods. Therefore a crystallisation screen was designed based upon the SGC Toronto conditions to identify similar conditions for the crystallisation of DPF3b. Crystallisation conditions were identified that gave DPF3b crystals X-ray diffraction data suitable to solve the structure of DPF3b to 1.75 Å (Figure 82, Appendix 4.2).

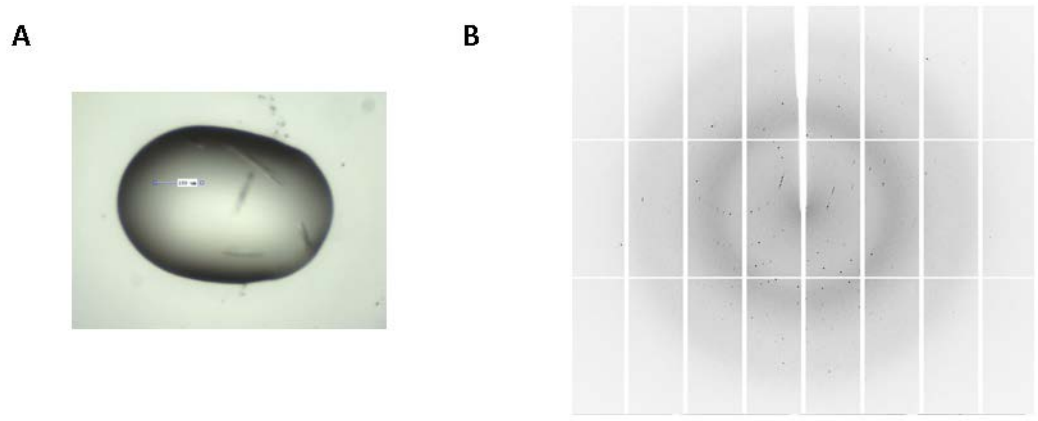


Figure 82. A. Crystals of DPF3b produced in crystallisation conditions based on those developed by SGC Toronto. **B.** Diffraction pattern of DPF3b crystals showing diffraction to 1.75 Å.

Crystal soaking assays require protein crystals to be produced reproducibly, so that multiple compounds can be tested. To test the reproducibility of DPF3b crystallisation a crystallisation plate was prepared with the conditions in all sub-wells matching those which had previously produced a diffracting crystal. Out of the 288 sub-wells 174 (60%) contained crystals, typically with a rod-like shape (Figure 83)

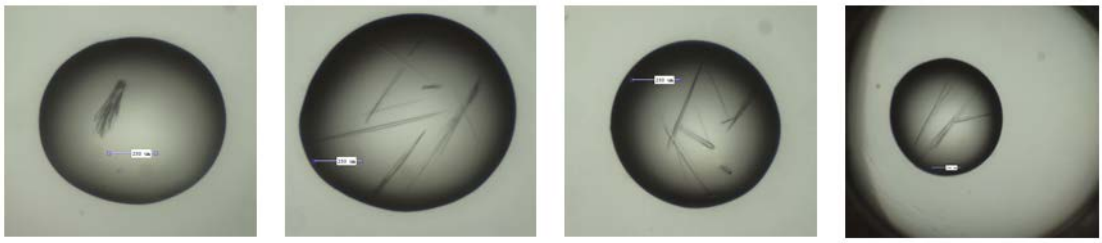


Figure 83. Crystallisation experiments showing crystals of DPF3b. All four of the crystals shown were grown in identical conditions (0.07 M HEPES pH 7.2; 1.4 M sodium citrate tribasic; 3% glycerol).

PEG Transfer

The crystallisation conditions used for the productions of DPF3b crystals described above use 1.4 M sodium citrate tribasic as the precipitant. Observations of previous crystal soaking experiments performed at the Structural Genomic Centre had suggested that crystal soaking is more successful when a polyethylene glycol (PEG) based precipitant is used (personal communication; Tobias Krojer). Therefore attempts were made to discover PEG based crystallisation conditions.

A crystallisation screen was performed using a sparse matrix of PEG based crystallisation conditions which failed to produce any crystals. Following this result, an experiment was designed to test whether crystals grown in 1.4 mM sodium citrate tribasic could be transferred to a PEG based condition. A screen of PEG containing crystallisation conditions was rationally designed based on the concept of relative humidity (RH).[175,176] Matching the relative humidity of the PEG based solution to that of the sodium citrate solution should prevent crystal dehydration on transfer to the PEG based solution. A web-based tool provided by European Synchrotron Radiation Facility (ESRF) was used to calculate the relative humidity of the sodium citrate solution and to design PEG based conditions with similar relative humidities. Despite the rational design of the PEG based solutions used, crystal transfer to the PEG solution rapidly dissolved, typically in under two minutes. The rapid dissolution suggested that DPF3b are incompatible with PEG based solutions; therefore further screening of PEG based solutions were not attempted.

Solvent Tolerance

Crystal soaking involves the addition of the compounds to be screened to the crystallisation wells after the formation of crystals. As organic molecules do not tend to be soluble in the aqueous crystallisation buffer at the high concentrations typically used in crystal soaking assays a suitable vehicle solvent must be used. Therefore it is important that the crystals tolerate the

addition of organic solvent to the crystallisation buffer. As the preparation of DPF3b crystals from PEG based solutions was not possible, crystals from the original sodium citrate tribasic conditions were used in solvent tolerance tests (Table 16).

Exposure time	Solvent percentage	DMSO	Glycerol	Acetonitrile
1 h	20% 30% 40%	Protein crystal dissolves and salt crystals form	Good diffraction	Good diffraction
4 h	20% 30% 40%	Protein crystal dissolves and salt crystals form	Good diffraction	Good diffraction
24 h	20% 30% 40%	Protein crystal dissolves and salt crystals form	Good diffraction	Good diffraction

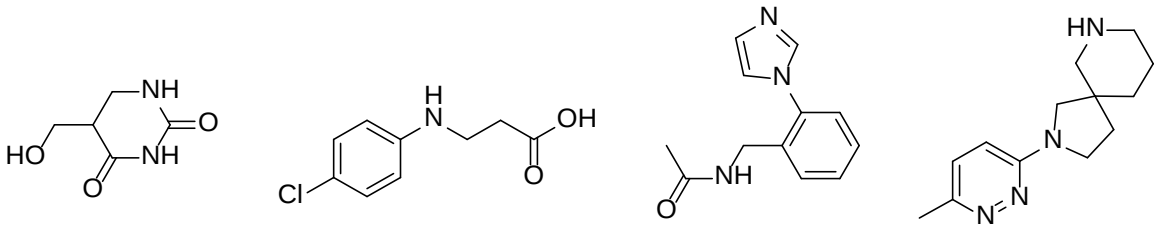
Table 16. Results of solvent tolerance tests carried out on crystals of DPF3b. Acetonitrile and glycerol appear well tolerated by DPF3b. DMSO initially causes the protein crystal to dissolve followed by the formation of salt crystals.

DPF3b crystals grown in the sodium citrate tribasic conditions were tested in terms of the tolerance to DMSO, glycerol, and acetonitrile. A mixture of well solution and organic solvent at the indicated percentages was prepared and added to the crystallisation buffer and left for a period of 1 h, 4 h, or 24 h. The crystal was then mounted and its diffraction measured using an in-house X-ray generator.

This solvent testing revealed that the crystals of DPF3b do not tolerate the presence of DMSO. DMSO is the most common solvent for compound libraries and the development of DPF3b crystals that tolerated DMSO would have been beneficial for a crystal soaking assay.

In order to stabilise DPF3b crystals to DMSO, a glutaraldehyde cross-linking strategy was employed.[177] This covalently links protein molecules within a crystal increasing the ability of the crystal to withstand changes to withstand changes to its environment. In the case of DPF3b, the addition of glutaraldehyde to either the protein solution after crystal formation, or to the reservoir solution after crystal formation, resulted in a loss of diffraction.

As DPF3b crystals could not be modified to tolerate DMSO, a small collection of compounds was transferred from DMSO to acetonitrile for a trial soaking experiment. The compounds chosen were weak fragments hits **18**, **23**, **28**, and a weak hit from the acetyl lysine mimetic library **64**. In all cases the soaked crystals showed diffraction but the structure could not be solved owing to crystal twinning.[178]



18 **23** **28** **64**

Figure 84. Fragments **18**, **23**, and **28** and a weak hits identified from the acetyl lysine mimetic library that was included as a test compound in crystal soaking trials.

Following the results of the experiments described above and consultation with members of the Structural Genomics Consortium protein crystallography group, work on developing a crystal soaking assay for DPF3b was paused. Although it was possible to achieve crystals of DPF3b that allowed for the structure to be solved at a resolution of 1.75 Å, these crystals were not reproducible enough or sufficiently solvent tolerant to enable their use in a crystal soaking assay.

Summary and Future Work

The SiteMap analysis described in Chapter 3 lead to the hypothesis that tandem PHDs are more ligandable that single PHDs. This chapter describes attempts to identify and characterise a small

molecule inhibitor of the tandem PHDs of DPF2. An AlphaScreen assay was developed and used to screen both fragments and molecules identified by virtual screening.

In total a fragment library consisting of 1324 unique molecules was collated from smaller fragment libraries. This library broadly follows the fragment 'Rule of Three'. Alongside this fragment collection three separate libraries were derived from a 10,000 member library by virtual screening, and an acetyl lysine mimetic library previously used in bromodomain screening.

Although primary hits were identified, none of these could be validated using a suitable secondary assay. The lack of active compounds is in line with the results of the virtual screen performed on the related tandem PHDs of MYST3. This virtual screen gave a highest scoring pose of -7.5, which would be considered a low score, and suggest that the target was difficult.

Although these results do not confirm the hypothesis that tandem PHDs are ligandable, they do not refute it either. The total number of compounds available for screening was 11,324, made up of a 10,000 member compound library and 1,324 fragments, with only 2,542 submitted to the AlphaScreen assay. Although requiring significant time and resource to screen, this number of compounds still only covers a small fraction of possible chemical space and is too small a screening library to definitely state that a target is unligandable. Use of a larger screening library may lead to the identification of primary hits suitable for optimisation.

It is also possible that the use of alternative primary screening methods may lead to the identification of primary hits. Techniques such as surface plasmon resonance (SPR), crystal soaking, or ligand observed NMR are common methods used for primary fragment screening.[179] The use of crystal soaking as a primary screen is particularly appealing due to the low false positive rate and the immediate availability of binding data for identified hits. This work discussed preliminary attempts to develop a crystal-soaking assay using DPF3b. Although this

attempt was not successful, optimisation of the DPF3b construct or the introduction of surface mutations could be used to develop a crystal form suitable for crystal soaking.

Another strategy that could be used for the development of a DPF2 inhibitor is the use of peptidomimetics. The recent development of a peptidomimetic inhibitor of the chromodomain of chromobox homolog 7 (CBX7) demonstrates that it is possible to design inhibitors that mimic histone peptides.[140] A peptide trimer matching the three *N*-terminal amino acids of H3 was found to have an IC_{50} of 120 μM in the DPF2 AlphaScreen assay. This could be used as a starting point for the development of a peptidomimetic DPF2 inhibitor.

Chapter 5 - Investigating the Ligandability of the PHD-JmjC of PHF8

As discussed in Chapter 3, PHDs that exist as part of multi-domain complexes are hypothesised to be more likely to be amenable to inhibition by small molecules. The specific example of the PHD-JmjC multi-domain complex of the histone demethylase PHD finger protein 8 (PHF8), which contains a groove between the PHD and JmjC domains where the histone 3 peptide binds, has previously been suggested as an example of a ligandable PHD.[78] This chapter will describe work to design suitable assays for the PHD-JmjC construct of PHF8, the results of the screening experiments to find inhibitors, and the development of an inhibitor series identified during the screening process.

Biological Function of PHF8

PHF8 (KDM7B) is a JmjC containing demethylase. PHF8 is a member of the KDM7 sub-family of JmjC demethylases which also includes the PHD finger protein 2 (PHF2) and lysine (K)-specific demethylase 7A (KDM7A, KIAA1718).[180] The JmjC domain is responsible for the 2-oxoglutarate and Fe(II) dependent enzymatic activity. JmjC domains catalyse hydroxylation of the methyl group of a methyl lysine and the resulting hemi-aminal collapses to give formaldehyde and the lysine residue with one less methyl group than before the enzymatic reaction. This mechanism differs from lysine specific demethylases (LSDs) which are flavin adenine dinucleotide (FAD) dependent (Chapter 1). Examples of biological functions of PHF8's enzymatic activity are summarised below.

PHF8 has been shown to demethylate the repressive H4K20me1 and H3K9me2 marks. There is also evidence that PHF8 is involved in regulating the transition from G_1 to S within the cell cycle by demethylation of H4K20me1.[181] The PHD of PHF8 was also shown to play an important role in this function by recruiting PHF8 to promoters by binding to the activating H3K4me2 and H3K4me3 marks. This binding specificity for the PHD of PHF8 towards histone 3 methylated at lysine 4 is in keeping with the predicted binding properties of PHDs from sub-family 3

(Chapter 2). Investigation of a PHF8 construct with the PHD deleted showed no activity towards H4K20me1 on mononucleosomes.[181]

Another example of PHF8 controlled gene expression was demonstrated by Feng et al.[182] PHF8 was shown to co-immunoprecipitate with RNA polymerase I (Pol I) and methylation sensitive polymerase chain reaction (PCR) suggests that PHF8 associates with hypomethylated ribosomal RNA (rRNA) genes. It was therefore suggested that PHF8 is involved in controlling transcription of ribosomal DNA (rDNA). PHF8 was also shown to colocalise with H3K4me3 marks during chromatin immunoprecipitation (ChIP) and a Y7A mutation in the aromatic cage of the PHD (Chapter 1) resulted in a loss of PHF8-dependent transcriptional activation. This suggests that a functional PHD is required to interact with H3K4me3 and induce PHF8 dependent transcription, in this case by demethylation of H3K9me2.

Both of these examples suggest that the function of the PHD and JmjC domain of PHF8 are interlinked. This is in agreement with the second part of the histone code hypothesis, which states that "modifications on the same or different histone tails may be interdependent and generate various combinations on any one nucleosome."[4] In both the examples described above the PHD binds to the activating mark H3K4me3 leading to demethylation of a nearby repressive mark, either H4K20me1 or H3K9me2. In both examples the demethylation of repressive marks leads to PHF8 dependent gene transcription.

Disease Associations of PHF8 mutations

Mutations in the *PHF8* gene are linked to the condition Siderius X-linked mental retardation (XLMR) whose symptoms also include cleft lip/palate.[183] Three of these mutations are nonsense mutations within the JmjC domain that would almost certainly lead to a loss of function. A fourth clinically observed mutation in the JmjC domain, F279S, has also shown to destroy the enzymatic activity of PHF8.[184] This suggests that the phenotypes associated with *PHF8* mutations are directly linked to the loss of enzymatic function.

Role of PHD in Enzymatic Activity

JmjC domains are capable of accepting mono-, di-, and trimethylated lysine as substrates. However a mass spectrometry based investigation of substrate selectivity using a construct containing only the JmjC of PHF8 using peptide fragments from H3 and H4 showed that PHF8 is only active for dimethylated lysine.[184] No activity was shown against H3K9me3 or H3K36me1, but the double demethylation of H3K36me2 was detected, which may be due to the incomplete release of H3K36me1 following the first demethylation event.

	JmjC Only	PHD-JmjC
H3K9me2	Activity[184]	Activity
H3K4me3K9me2	Not tested	Greater activity than H3K4me0K9me2[111]
H3K9me3	No activity[184]	No activity
H3K27me2	Activity[184]	Not tested
H3K4me3K27me2	Not tested	Not tested
H3K36me2	Activity[184]	Not tested
H4K20me1	No activity[184]	Not tested
H4K20me1 with H3K4me3 on same nucleosome	No activity[181]	Activity[181]

Table 17. Comparison of the substrate specificity of PHF8 with and without the PHD. The presence of the PHD is required for H4K20me1 demethylation, and the PHD can enhance activity for H3K9me2 by binding to a H3K4me3 on the same peptide.

The activity and specificity of PHF8 is modulated by its PHD in the presence of a H3K4me3 mark (Table 17). The K_M of the PHD-JmjC for a H3K4me3K9me2 peptide is sixteen-fold lower than for a similar peptide without the H3K4me3 mark.[111] This difference in activity is supported by isothermal calorimetry measurements which show that although the PHD-JmjC binds a H3K4me3K9me2 peptide with a K_D of approximately 1 μM; no binding was detected for a H3K4me0K9me2 peptide under the same conditions. This suggests that the PHD is required to recruit PHF8 to H3K4me3 marks, allowing the demethylation of a H3K9me2 mark on the same peptide.

Comparison to related PHD-JmjC complexes

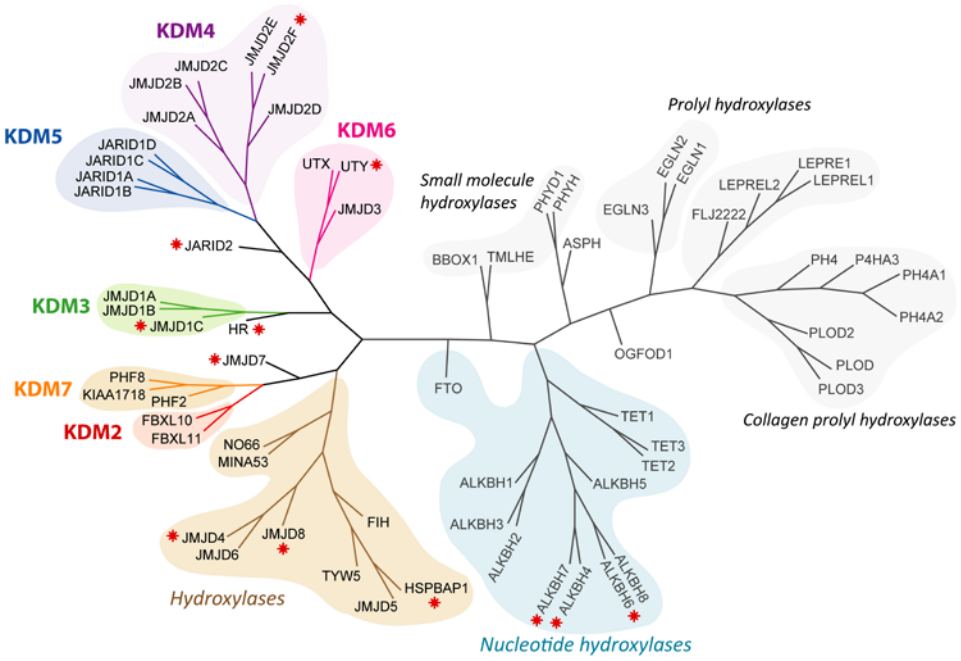


Figure 85. The family of human 2-oxoglutarate dependent oxygenases which includes JmjC demethylases and hydroxylases for a range of biological substrates. The JmjC family is split into six sub-families. Figure taken from *The roles of Jumonji-type oxygenases in human disease*.[180]

The JmjCs of the KDM7 sub-family are closely related to the KDM2 sub-family (Figure 85), and previous studies have shown that the plant growth regulator daminozide is a broad spectrum inhibitor of the KDM2/7 sub-families of demethylases with 60-fold selectivity over other human KDM sub-families.[185] However, the PHDs found in the KDM2 sub-family are not closely related to the KDM7 PHDs. The KDM7 PHDs are found in sub-family 3 of the PHD tree with other H3K4me3 binders and the KDM2 PHDs found in sub-family 8, a sub-family not predicted to be H3K4me3 binders (Chapter 2). Therefore it may be possible to use the PHDs to design inhibitors that are selective for the KDM7 sub-family over the KDM2 sub-family.

Examining the structures of the PHD-JmjCs of the KDM7 subfamily shows that the PHD-JmjC of PHF8 is found in a different domain orientation to that of KDM7A (Figure 86). As discussed above, binding of the PHD of PHF8 to H3k4me3 enhances demethylation of H3K9me2. Contrastingly, in the case of KDM7A the presence of H3K4me3 reduces demethylation activity at

H3K9me2, but increases demethylation activity at H3K27me2 highlighting combinatorial nature of epigenetic domains in determining substrate specificity.

Inspection of the structures reveals that the histone binding face of the PHD of PHF8 makes direct contact with the JmjC domain, forming an enclosed peptide binding groove.[186] Within this groove the aromatic cage of the PHD and the catalytic site of the JmjC are positioned such that H3K4me3 and H3K9me2 can engage these two sites simultaneously. In the case of KDM7A, the PHD's histone binding surface faces away from the JmjC (Figure 86). This structure suggests the aromatic cage of the PHD and catalytic site of the JmjC are too far apart for H3K4me3 and H3K9me2 to be engaged simultaneously. However, the greater distance between K4 and K27 allows H3K4me3 to bind the PHD and H3K27me2 to reach the catalytic site at the same time. This variation in domain orientation offers an explanation for the differing enzymatic substrate specificity of PHF8 and KDM7A and offers the potential to design selective inhibitors.

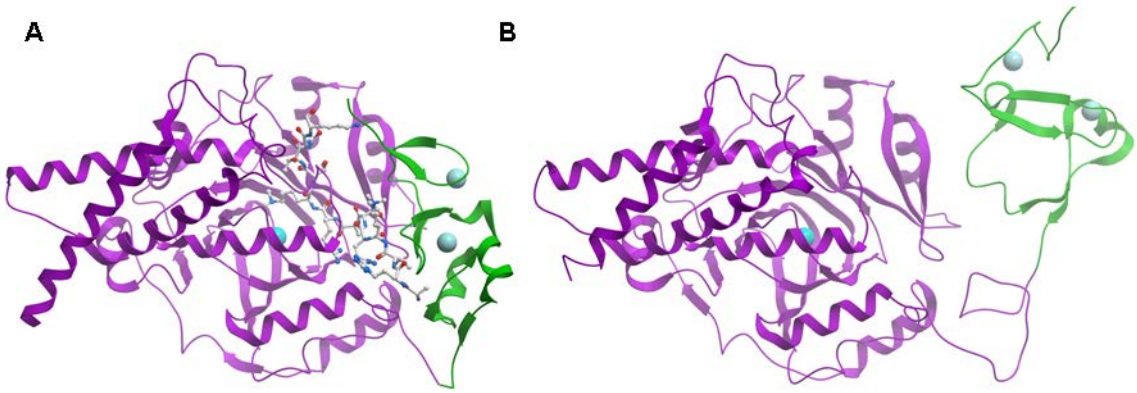


Figure 86. The domain orientation differs between PHF8 and KDM7A. **A.** The PHD-JmjC of PHF8. The histone binding surface of the PHD (green) makes contact with the surface of the JmjC (magenta) creating a binding groove for H3 (white sticks). PDB ID: 3KV4. **B.** The PHD-JmjC of KDM7A. The histone binding surface of the PHD (green) is directed away from the surface of the JmjC (magenta). PDB ID: 3KV5.

Inhibition of JmjC Domains

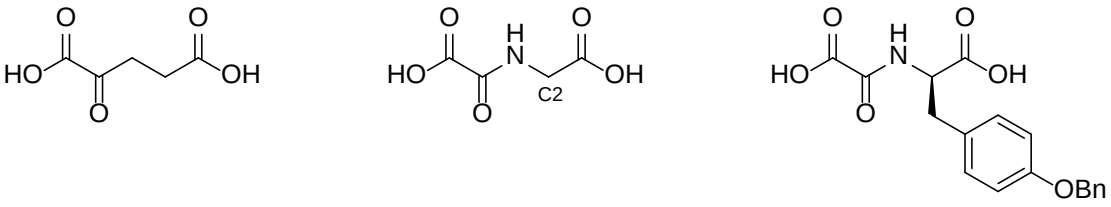
Histone lysine demethylases are promising targets for treating diseases such as cancer.[187] This has led to a large to the development of range of inhibitors[188] and chemical probes.[189–191] This

section will briefly review JmjC inhibitors and discuss the contribution that a PHF8 inhibitor could potentially make to the field.

Co-factor Competitive Inhibitors

The most well established area of JmjC inhibitor research lies in the field of 2-oxoglutarate (2OG) competitive, metal binding inhibitors.

One of the first inhibitors reported for the JmjC domain family is the broad spectrum inhibitor *N*-oxalylglycine (NOG).[192] NOG is a close structural analogue of the essential co-factor 2OG and therefore inhibits many of the non-JmjC enzymes that depend on this co-factor, including many hydroxylases (Figure 85, nucleotide- and prolyl-hydroxylases).[193] Selective inhibitors have been developed based on NOG by substitution at C-2 to give compounds such as compound **66** which exploit hydrophobic pockets near the co-factor binding site that are found in only in the KDM4 sub-family (Figure 87).[194]



2-oxoglutarate *N*-oxalylglycine **65**

Figure 87. The essential JmjC co-factor 2-oxoglutarate (2OG); 2OG's close structural analogue *N*-oxalylglycine (NOG) which is a broad spectrum inhibitor of 2OG dependent oxygenases, including all tested JmjCs; and compound **65**, a derivative of NOG which is a selective inhibitor of members of the KDM4 sub-family.

Pyridine-2,4-dicarboxylic acid (2,4-PDCA) is another broad spectrum JmjC inhibitor.[192] Although the highly polar nature of 2,4-PDCA makes in unsuitable for cellular studies, the dimethyl or diethyl ester pro-drug have been shown to inhibit the demethylation of H3K9me3 in human HEK293T cells.[195]

The JmjC inhibitors described above are highly polar as they bind in the same site as the polar co-factor 2OG. Despite the polarity of the 2OG binding site it is possible to develop cell penetrable JmjC chemical probes. The KDM5 inhibitor KDOAM-25[189] and the pan-inhibitor IOX1[190] are cell active JmjC inhibitors (Figure 87). The JMJD3 inhibitor GSK-J1 is not cell-permeable; however, its ethyl ester is cell-permeable and acts as a pro-drug as it is hydrolysed to give GSK-J1 in cells.[191]

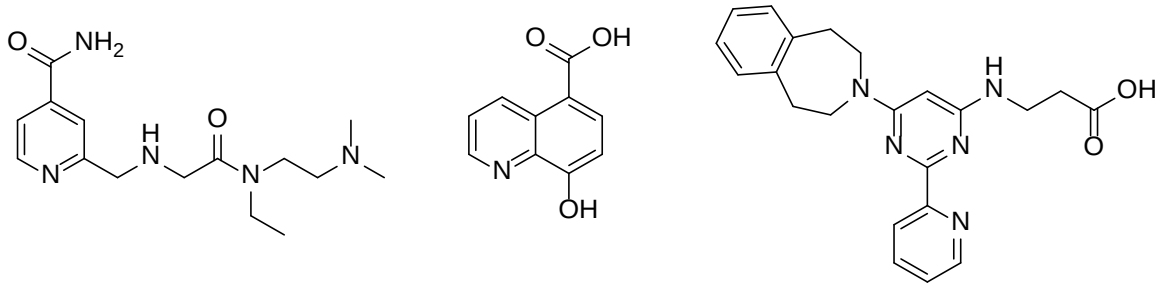


Figure 88. Chemical probes of JmjC domains that bind to the Fe(II) found in the catalytic site of JmjC domains. KDOAM-25 is selective for members of the KDM5 sub-family. IOX1 is a pan-inhibitor which inhibits members of the KDM3, KDM4, KDM5, and KDM6 sub-families. GSK-J1 is selective for JMJD3, a member of the KDM6 sub-family.

Substrate Competitive Inhibitors

Although the majority of work on JmjC inhibitors has focused on co-factor competitive, metal binding ligands, there are some published examples of histone substrate competitive JmjC inhibitors. Initial work in this area focused on modifying the natural peptidic substrate of JmjC domains. For example, Lohse et al. developed inhibitors for members of the KDM4 sub-family by replacing the trimethylated lysine in the substrate with a lysine linked to a uracil, a known metal chelator (Figure 89).[196] By using this strategy to direct a metal chelator to the active site they were able to able to develop inhibitors with micromolar potency.

A similar approach of tethering a metal chelator to a substrate peptide was employed by Woon et al. *N*-oxalyl-D-cysteine (DNOC), a thiol containing derivative of NOG, was screened alongside a series of H3 peptide fragments where a single residue had been changed to cysteine (Figure 89).[197] They were able to identify that a peptide containing a T11C mutation formed a

disulphide bond with DNOC in the presence of KDM4E. Replacing the disulphide with a thioether linkage gave an inhibitor with an IC_{50} of 70 nM.

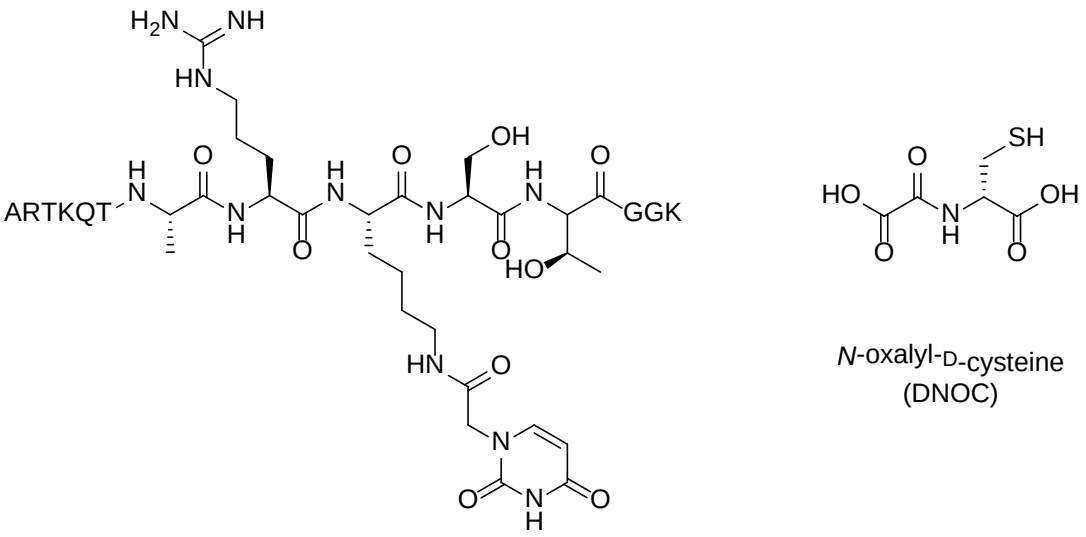


Figure 89. Lohse et al designed a substrate competitive inhibitor by conjugating K9 of a H3 peptide to the known metal chelator uracil.[196] Woon et al used a similar approach of linking the metal chelating DNOC to histone 3 peptides with cysteine mutations via disulphide bond formation.

The only reported small molecule substrate only competitive inhibitor of a JmjC domain is E67-2, a reported inhibitor of KDM7A (Figure 90).[198] Although there is no kinetic evidence to show that this compound is not also co-factor competitive, a crystal structure shows both E67-2 and 2OG bound to KDM7A at separate sites. The inhibitor E67-2 is derived from BIX-01294, an inhibitor of the H3K9 methyltransferase euchromatic histone-lysine *N*-methyltransferase 2 (EHMT2, also known as G9a). The authors rationalised that the substrate binding pocket for a H3K9 methyltransferase must be similar to that of a demethylase capable of demethylating H3K9. It is important to note that that the authors used a mass-spectrometry based assay to measure demethylation of H3K9me2 peptide, with the peptide unmodified at H3K4. As discussed above, the presence of H3K4me3 reduces the activity of KDM7A to H3K9me2, but increases demethylation of H3K27me2. Therefore although they note that the inhibitory effect of E67-2 is similar for a KDM7A PHD-JmjC construct and a JmjC only construct, they fail to consider the full

effects of the PHD and H3K4 methylation state. A more appropriate substrate for use in the mass spectrometry demethylation assay would be one containing H3K4me3 and H3K27me2 marks.

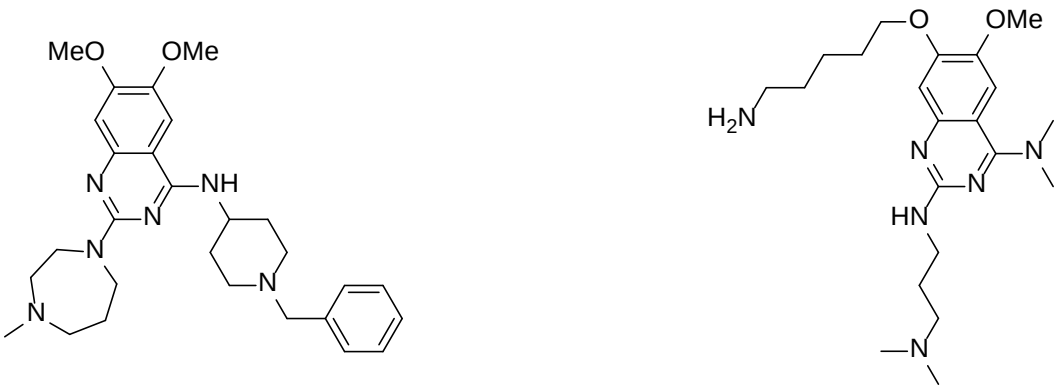


Figure 90. The EHMT2 methyltransferase inhibitor BIX-01294 and the KDM7A inhibitor E67-2. Both molecules are substrate competitive mimetics of H3K9me3.

Inhibitors of PHF8

The KDM7A inhibitor E67-2 was shown to inhibit a JmjC only construct of PHF8, but showed greatly reduced inhibition of a PHD-JmjC construct. This shows the importance of considering both domains when designing PHF8 inhibitors. It also highlights the exciting potential for inhibitor selectivity offered by pockets formed at domain-domain interfaces.

Aside from co-factor competitive pan-inhibitors discussed above, the only reported inhibitors of PHF8 are the plant growth regulator daminozide and hydroxamate **66** (Figure 91).[185,199] Daminozide is a closely related in structure to both 2OG and succinate, a by-product of JmjC catalysed demethylation. Daminozide shows good selectivity for members of the KDM2 and KDM7 sub-families of JmjCs over other tested sub-families.[185]

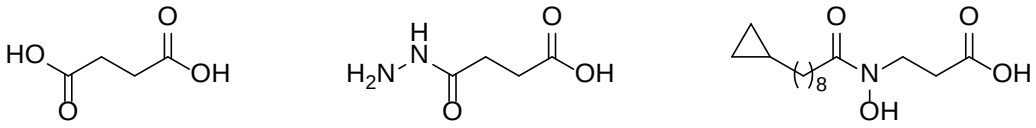


Figure 91. A comparison of the structures of succinate, the by-product of JmjC catalysed demethylation; the KDM2/7 inhibitor daminozide; and hydroxamate inhibitor **66**.

Hydroxamate **66** was designed as a KDM7 sub-family inhibitor based on previous screening results and docking studies. It is predicted to be a metal-binding inhibitor, with the hydroxyl and carbonyl oxygens of the hydroxamate group binding Fe(II) in a bidentate manner. It is most potent against KDM7A (IC_{50} = 0.20 µM), though does also show activity against PHF8 (IC_{50} = 1.2 µM). Hydroxamate **66** also shows activity against KDM2A (IC_{50} = 6.8 µM).

The lack of selectivity between the KDM2 and KDM7 sub-families shown by daminozide and hydroxamate **6** shows the limitations of trying to design selective, 2OG competitive PHF8 inhibitors. This demonstrates the need for the exploration of new strategies such as targeting the enclosed peptide binding groove formed at the PHD-JmjC interface which is unique to PHF8.

Designing Assays for PHF8

There are well established screening assays for JmjC domains that take advantage of the enzymatic activity of the domain. Examples of enzymatic assays of JmjCs include AlphaScreen and mass spectrometry based assays.[185,200]

AlphaScreen assays for JmjC domains differ from the protein-protein interaction AlphaScreen assay described in Chapter 4. As before, a biotinylated H3 peptide containing a demethylase substrate (e.g. H3K9me3) is bound to a streptavidin coated donor bead, which is incubated with the JmjC containing enzyme of interest, the necessary co-factors, a product specific antibody, and acceptor beads coated with the immunoglobulin binder protein A. The JmjC domain demethylates the H3 peptide which is then bound by the product specific antibody. This antibody is bound by protein A, bringing the beads into close contact. Upon stimulation at

680 nm, $^1O_2$ is formed and moves by diffusion between the beads leading to an emission in the 520-620 nm range (Figure 92). If an inhibitor is present during the incubation period, the degree of demethylation will be reduced; the antibody will not recognize the product and the donor and acceptor beads will be physically separated. The unstable $^1O_2$ that is formed will react with solvent before it can reach the acceptor bead and emission will be diminished leading to a loss of signal.[200]

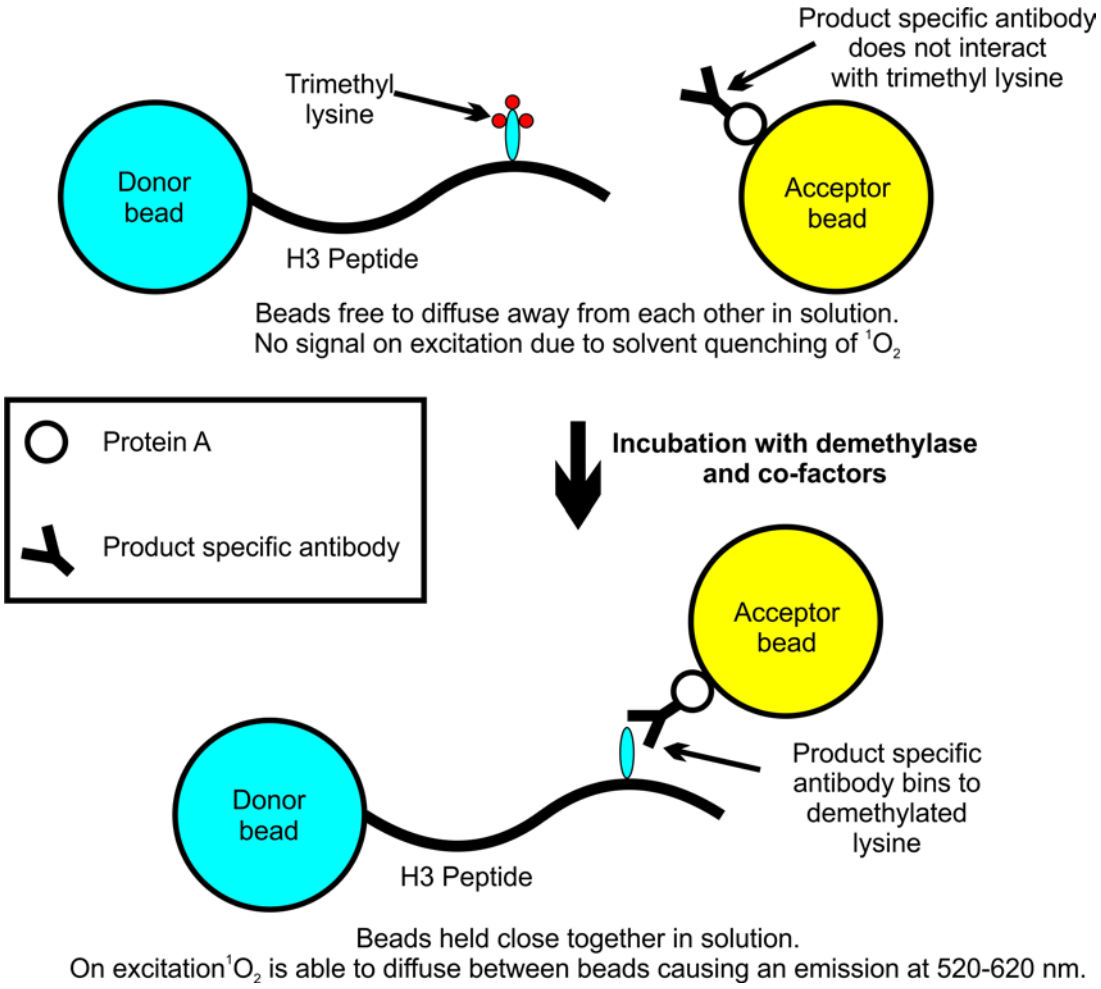


Figure 92. Schematic representation of enzymatic AlphaScreen assay. Initially the beads are not in close contact, but as the enzymatic demethylation proceeds, the product specific antibody is able to bind to H3 bring the beads close together. This allows $^1O_2$ to diffuse between beads causing an emission at 520-620 nm. If an inhibitor is present during the incubation period the degree of demethylation is reduced, leading ultimately to a loss of signal.

Mass spectrometry assays for JmjC domains take advantage of the change of mass of the substrate peptide on demethylation. The enzyme and necessary co-factors are incubated with peptide for a time period determined by the activity of the enzyme. The mixture is then analysed by mass spectrometry to calculate the ratio of methylated to demethylated peptide. If an inhibitor is present during the incubation period, there will be less demethylated peptide compared to the negative control.

Alongside the two assay types described above, it is also possible to study JmjC activity and inhibition using ^{14}C-labelled 2OG turnover assays;[184] by monitoring formaldehyde production using a formaldehyde dehydrogenase (FDH) and nicotinamide adenine dinucleotide (NAD^+) coupled assay;[192] and by cell based immunostaining assays.[191]

The two types of assay described in detail above are not suitable for discriminating between a co-factor competitive inhibitor and a substrate competitive inhibitor in a single concentration point screening assay. As the stated aim of this work was to discover inhibitors that bind at the peptide binding groove at the PHD-JmjC interface, an alternative assay method was sought. A peptide displacement AlphaScreen assay was chosen, similar to that described for DPF2 in Chapter 4. Compounds that are 2OG co-factor competitive should have no effect on the binding of a H3 peptide to the PHD-JmjC of PHF8. Therefore the only compounds that should appear as hits in the assay are those that inhibit substrate binding to the PHD-JmjC of PHF8. Work described above by Horton et al. showed that the K_D for a H3K4me3K9me2 peptide with the PHD-JmjC is approximately 1 μM, whereas no binding was detected for a H3K9me2 peptide under similar consitions. This suggests that H3K4me3 recognition by the PHD of PHF8 is a key compotent of peptide binding, and therefore any substrate competitive inhibitor is likely to inhibit this key interaction.

The PHF8 assay used in this work was developed by Octovia Monteiro (Structural Genomics Consortium, University of Oxford). The assay uses a H3 peptide consisting of the first twenty-one

residues of H3 with a H3K4me3 mark. As in the case of DPF2 described in Chapter 4, the concentrations of both protein and peptide were optimised to maximise signal (Figure 93).

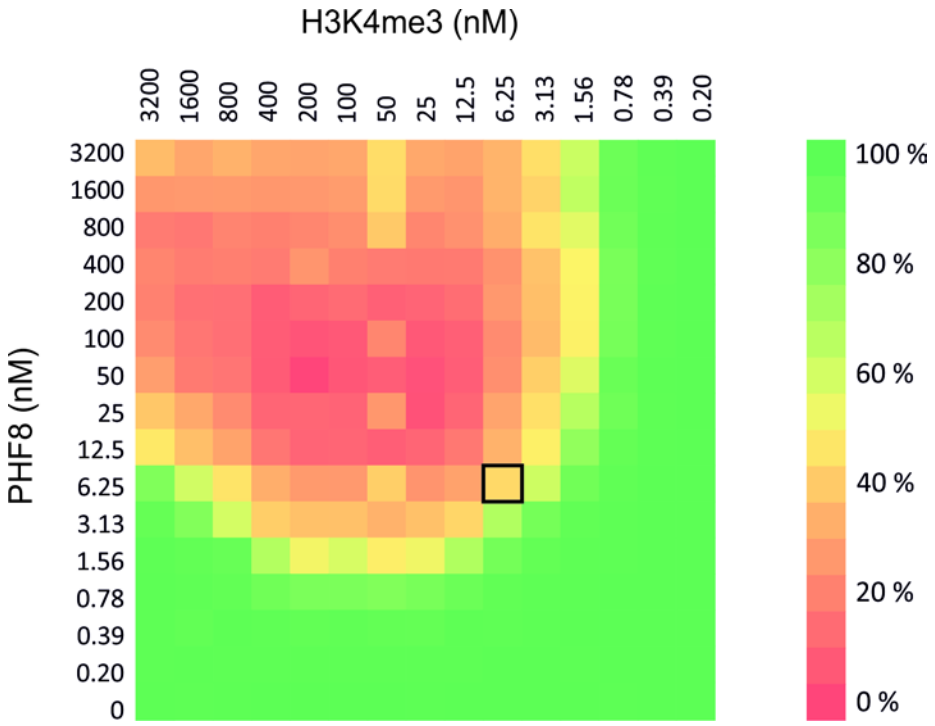


Figure 93. Heat maps showing AlphaScreen signal for varying concentrations of PHF8 and partner peptide. The concentrations of PHF8 and H3K4me3 used in the assay are indicated with a black square. This work was carried out by Octovia Monteiro (Structural Genomics Consortium, University of Oxford).

Library Design

A 10,000 member library was described in Chapter 4 for use in screening against DPF2. This library was used for screening against PHF8 with prior virtual screening to prioritise compounds, as in the case of DPF2. The fragment library described in Chapter 4 was also used for PHF8 screening.

Design of a Library Focused by Virtual Screening

The 10,000 member compound library was made available for use by Daniel Ebner (Target Discovery Institute, University of Oxford) and used in a virtual screen to prioritise compounds for experimental screening with the AlphaScreen assay described above. As was the case for the

DPF2 AlphaScreen assay described in Chapter 4, the PHF8 AlphaScreen assay would have proved suitable for a high throughput screen (HTS) of all 10,000 compounds, but the costs of such a screen would be prohibitive. Therefore it was necessary to use virtual screening to prioritise the library.

Virtual Screen of PHF8

There is only one available structure of PHF8 that contains both the PHD and JmjC domains,[111] although there are other structures of the isolated JmjC domain only available and a solution phase structure of the isolated PHD.[201] For this virtual screen the structure containing both PHD and JmjC was chosen as the aim of this work is to identify compounds that bind at the interface of the PHD and JmjC.

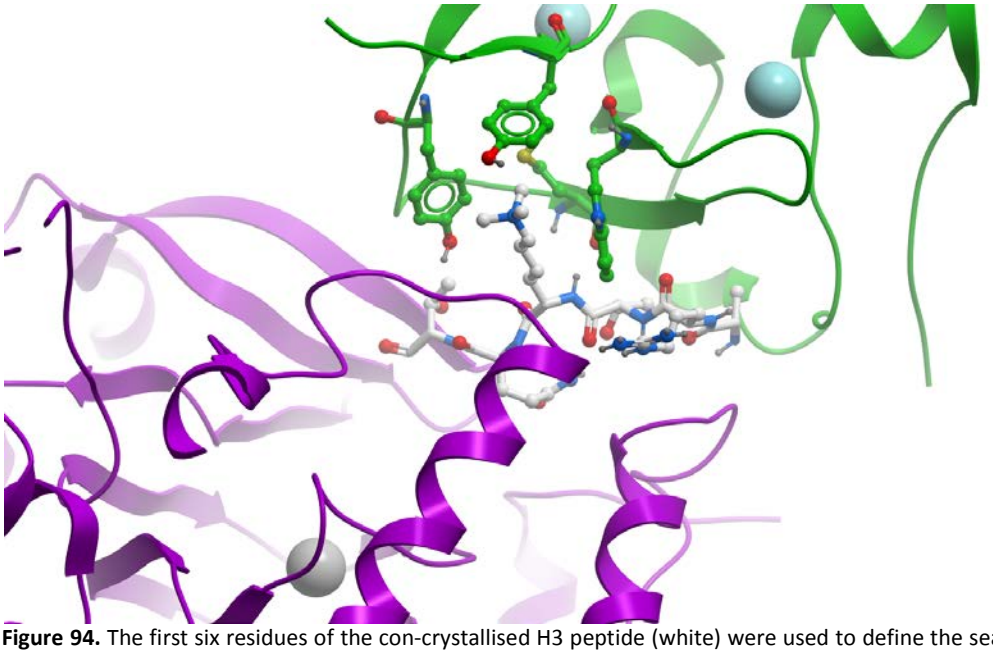


Figure 94. The first six residues of the con-crystallised H3 peptide (white) were used to define the search area for the virtual screen of PHF8. This section of the peptide sits in the groove formed by the PHD (green) and JmjC (magenta). The aromatic cage is shown as green sticks. PDB ID: 3KV4.

The protein was pre-processed using the Schrödinger Protein Preparation Wizard, with water and ethylene glycol molecules removed. The structure contains a H3 peptide consisting of residues 1-14 with H3M4me3 and H3K9me2 marks. The first six of these residues sit in the site

between the PHD and the JmjC and were therefore used to define the search area for the virtual screen (Figure 94).

Results of Virtual Screening

As discussed in Chapter 4, a GlideScore more negative than -10 is considered a good score. In the case of the tandem PHDs of MYST3 studied in Chapter 4, the highest scoring pose was -7.49, suggesting a difficult target. This is in line with the results seen in the rest of Chapter 4, where no primary hits could be identified. In the case of PHF8, nine compounds had a GlideScore of -10 of better, suggesting that PHF8 is a more ligandable target than the tandem PHDs examined in Chapter 4. It is interesting to note that of these nine compounds with GlideScores better than -10, seven of them contain a tertiary amine which is predicted to bind the aromatic cage where H3K4me3 would usually bind (Figure 95).

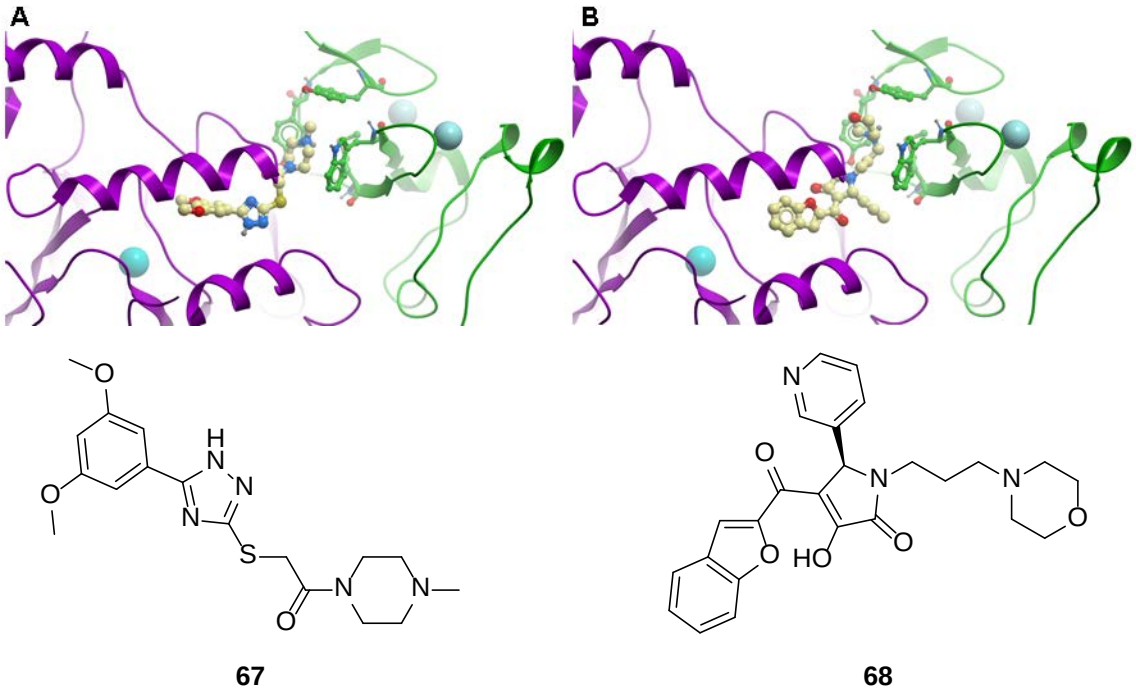


Figure 95. An example of two high scoring compounds in the PHF8 virtual screen. Top hits from the virtual screen which shows a molecule with a tertiary amine bound in the aromatic cage of the PHD of PHF8 (green). **A.** Compound **67** shown with the *N*-methyl piperazine bound in the aromatic cage of PHF8. **B.** Compound **68** shown with the morpholine group bound in the aromatic cage of PHF8.

The top 500 ligand poses were selected for the PHF8 virtual screening library. This set contained only 233 unique compounds, as many of these compounds had multiple high scoring poses. These compounds were inspected for the presence of protein reactive or unstable functional groups; none were identified, and all compounds were progressed to experimental screening.

Summary of Libraries to be Screened

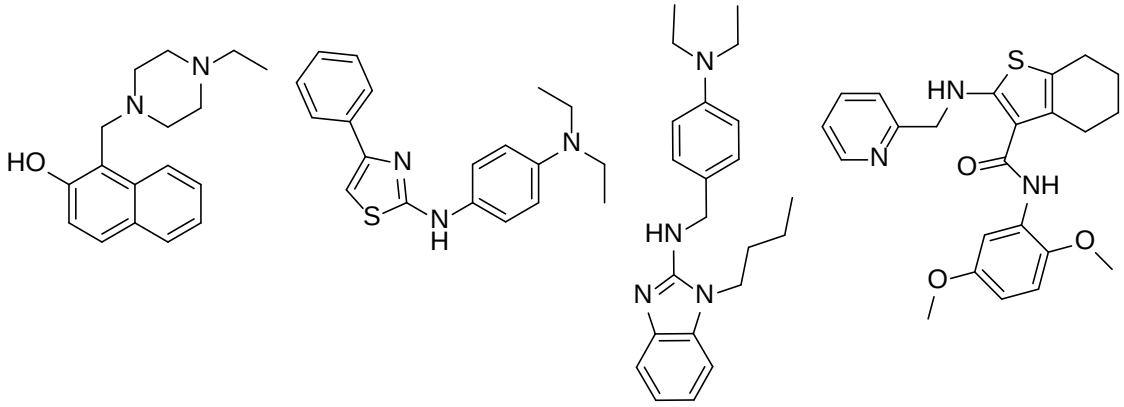
Two distinct libraries were selected for experimental screening using the AlphaScreen assay described above. A 1324 member fragment library also used for DPF2 screening in Chapter 4, and a 233 member library derived from virtual screening experiment described above.

Results of Primary Screen

All compounds were screened in duplicate at a single concentration and the results normalised against DMSO and water controls. Any compound which showed activity in this initial assay was tested using the dually His6 tagged and biotinylated peptide counter screen assay as described in Chapter 4. Compounds which showed a meaningful difference in inhibition between the PHF8 assay and the counter screen were progressed to IC_{50} measurement.

Results from Virtual Screening Library

The 233 compounds from the virtual screening library were screened at 100 µM. Four compounds showed > 80% inhibition at 100 µM (Figure 96). These compounds were tested using the dually His6 tagged and biotinylated peptide counter screen assay. Compound **69** showed the greatest difference between PHF8 inhibition and counter screen inhibition, but this was discounted from further investigation as it is a likely to form an *ortho* quinone methide species on hydrolysis. These species are highly reactive and likely to cause assay interference by protein reactivity. The remaining three compounds showed undesirable activity in the counter screen assay and were therefore not investigated further.

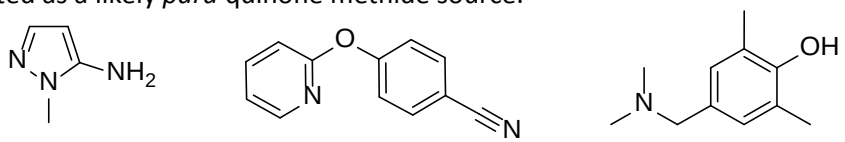


69	**70**	**71**	**72**
PHF8 inhibition at 100 µM = 95% Counter screen inhibition at 100 µM =33%	PHF8 inhibition at 100 µM = 91% Counter screen inhibition at 100 µM = 89%	PHF8 inhibition at 100 µM = 88% Counter screen inhibition at 100 µM = 69%	PHF8 inhibition at 100 µM = 85% Counter screen inhibition at 100 µM = 90%

Figure 96. Most active compounds from virtual screening library. For each compound percentage inhibition at 100 µM is shown. All four compounds were rejected as primary hits due to activity in the counter screen assay, suggesting that they are assay interference compounds.

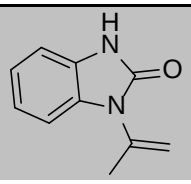
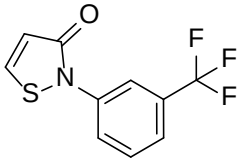
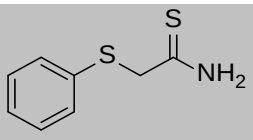
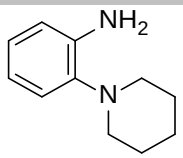
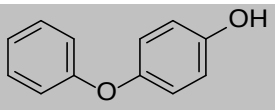
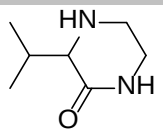
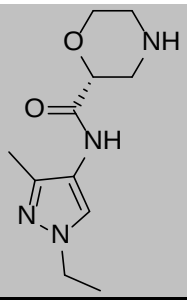
Results from Fragment Library

Fragments were screened in duplicate at 2 mM and compounds that showed greater than 70% inhibition in the DPF2 assay and less than 20% inhibition in the counter screen were progressed to IC_{50} measurement. Sixteen compounds were initially selected for IC_{50} measurement, but on further inspection three were rejected. (Figure 97) Compounds **11** and **14** were rejected on the basis that they also showed activity in the DPF2 AlphaScreen assay, and were therefore likely to be assay interference compounds or non-selective inhibitors. Compound **73** was rejected as a likely *para* quinone methide source.[202]



11 **14** **73**

Figure 97. Compounds that showed a meaningful difference between assay and counter screen inhibition, but were rejected prior to further investigation. Compounds **11** and **14** appeared as hits in the DPF2 assay and are therefore likely assay interference compounds or promiscuous inhibitors, and compound **73** is a potential *para* quinone methide source.

Compound Number	Structure	IC_{50} (μM)	IC_{50} Upper 95% Confidence Limit	IC_{50} Lower 95% Confidence Limit
74		540 μM	860 μM	340 μM
75		510 μM	630 μM	410 μM
76		790 μM	1100 μM	590 μM
77		270 μM	310 μM	220 μM
78		290 μM	440 μM	190 μM
79		540 μM	860 μM	340 μM
80		270 μM	340 μM	200 μM

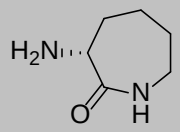
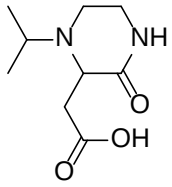
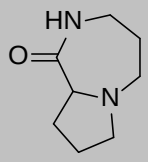
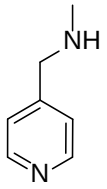
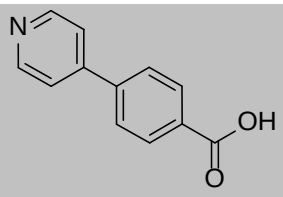
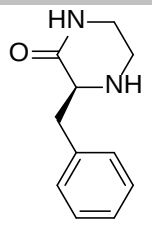
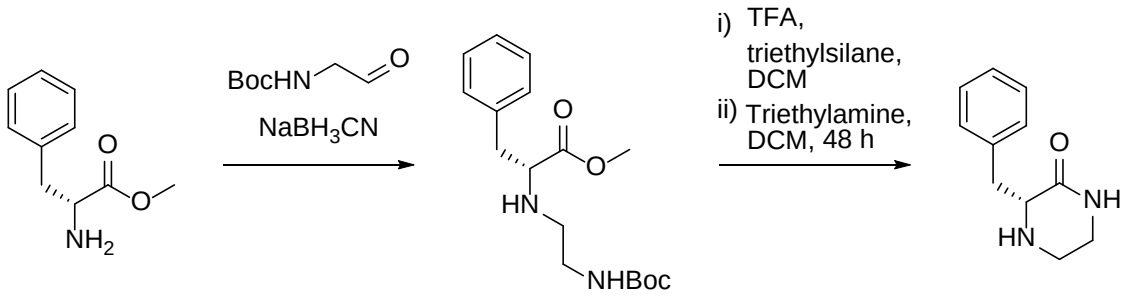
Compound Number	Structure	IC_{50} (μM)	IC_{50} Upper 95% Confidence Limit	IC_{50} Lower 95% Confidence Limit
81		140 μM	170 μM	110 μM
82		510 μM	630 μM	410 μM
83		790 μM	1100 μM	590 μM
84		130 μM	150 μM	110 μM
85		110 μM	130 μM	90 μM
86		260 μM	310 μM	220 μM

Table 18. Fragment hits from PHF8 AlphaScreen assay. IC_{50} measurements were taken for thirteen compounds and ranged from 110 to 790 μM. Compounds where stereochemistry is shown were tested as single enantiomers; compounds where no stereochemistry is indicated were tested as racemates.

The remaining thirteen compounds were progressed for IC_{50} measurement. Of these compounds, ketopiperazine **86** was the most attractive for further investigation. The ketopiperazine core is amenable to derivatisation at either nitrogen, and it is also possible add substituents to the phenyl ring. Compound **86** is also predicted to be uncharged at neutral pH.

Ketopiperazine Series



87 **88** **89**

Scheme 3. Synthesis of the enantiomer of compound **86**. D-Phenylalanine methyl ester **87** was used as the amine partner in a reductive alkylation with *N*-Boc-2-aminoacetaldehyde. The *tert*-butyloxycarbonyl protecting group of the resultant product was removed by hydrolysis with trifluoroacetic acid and the resulting primary amine cyclised in the presence of triethylamine.

The enantiomer of ketopiperazine **86** was synthesised to test if the observed activity of this compound resided in a single enantiomer. If compound **86** was inhibiting the assay by an interference mechanism such as aggregation, protein reactivity, or an AlphaScreen specific mechanism such as singlet oxygen quenching, the enantiomer would be expected to show the same potency. However, if compound **86** showed inhibition by binding to the protein and preventing peptide binding, the enantiomer would be expected to show a different level of activity due to the chirality of the compound binding site.

Compound **89** was synthesised using a method described in a patent granted to Novo Nordisk in 1999 (Scheme 3).[203] D-Phenylalanine methyl ester was coupled in a reductive alkylation with *N*-Boc-2-aminoacetaldehyde. This reaction gave yields in the range of 30-40% after column chromatography. However, these low yields proved satisfactory for the purpose of the synthesis of compound **89**. The isolated product was deprotected using trifluoroacetic acid. It was found

that altering the reported procedure by adding triethylsilane as a *tert*-butyl cation scavenging agent reduced the formation of a brightly coloured by-product which had previously proved difficult to separate. The by-products of this reaction could be removed by trituration in line with a procedure described by Mehta et al.[204] The addition of triethylamine induced the cyclisation of the resulting primary amine to give ketopiperazine **89**.

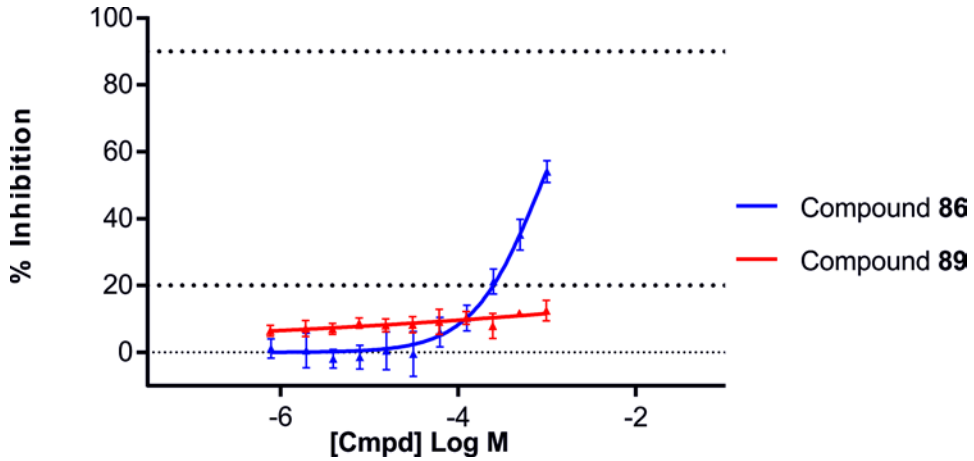
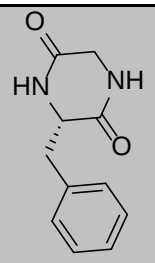
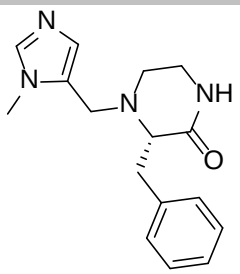
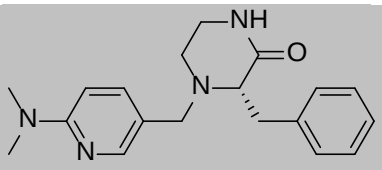
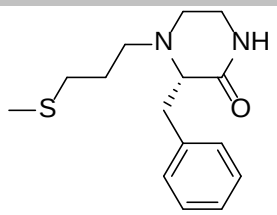
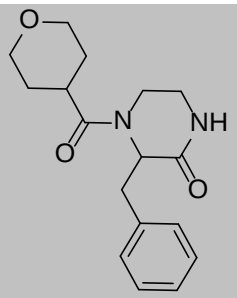
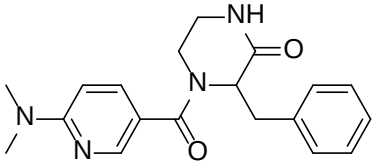


Figure 98. Comparison of the activities of compound **86** and its enantiomer compound **89** as measured by the PHF8 AlphaScreen assay. Although a full binding curve could not be obtained for compound **86**, it clearly shows greater activity that its enantiomer.

A resynthesised batch of ketopiperazine **86** was tested alongside its enantiomer ketopiperazine **89** (Figure 98). Although the resynthesised compound **86** did not show the same activity as the original screening batch, it clearly shows greater activity than its enantiomer **89**. This evidence was deemed sufficient to provoke further investigation into compound **86** and its analogues.

SAR by Catalogue

A series of analogues retaining the 3-benzyl ketopiperazine core of compound **86** with alkyl or acyl substituents at position 4 were purchased and their IC_{50} values measured. The IC_{50} of the counter screen assay was also measured to check for assay interference compounds (Table 19).

Compound Number	Structure	IC_{50} (μM)	IC_{50} Upper 95% Confidence Limit	IC_{50} Lower 95% Confidence Limit	Counter Screen Inhibition (IC_{50})
90		> 1 mM			No inhibition
91		190 μM	450 μM	77 μM	120 μM
92		310 μM	360 μM	270 μM	400 μM
93		77 μM	100 μM	58 μM	290 μM
94		150 μM	88 μM	250 μM	No inhibition
95		900 μM	1400 μM	700 μM	No inhibition

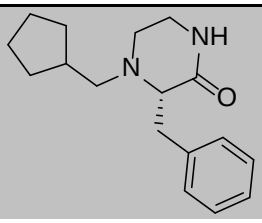
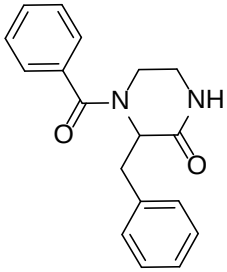
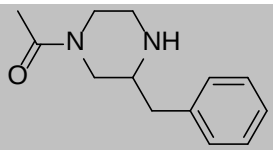
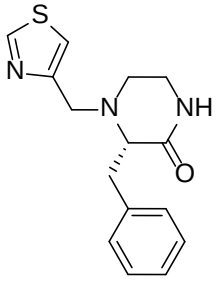
Compound Number	Structure	IC_{50} (μM)	IC_{50} Upper 95% Confidence Limit	IC_{50} Lower 95% Confidence Limit	Counter Screen Inhibition (IC_{50})
96		110 μM	180 μM	65 μM	280 μM
97		750 μM			750 μM
98		29 μM	33 μM	25 μM	80 μM
99		150 μM	190 μM	110 μM	800 μM

Table 19. Analogues of compound **86** were purchased and screened using the AlphaScreen assay. The compounds were also tested using the dually His6 tagged and biotinylated peptide counter screen assay. Compounds where stereochemistry is shown were tested as single enantiomers; compounds where no stereochemistry is indicated were tested as racemates.

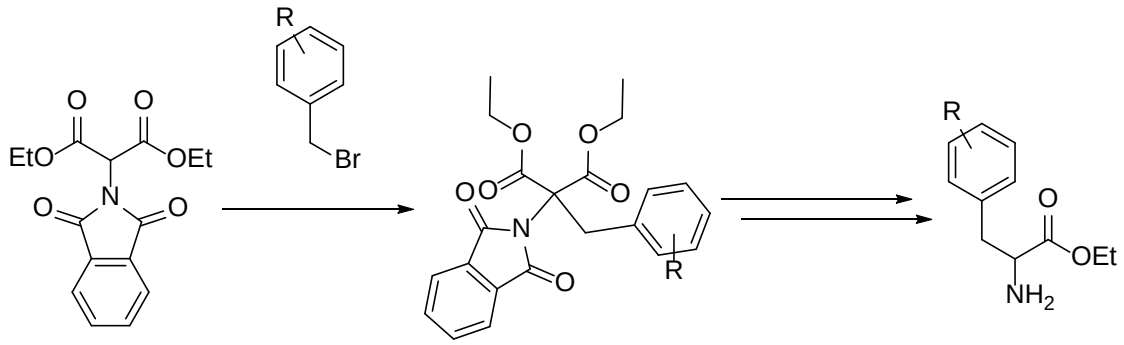
Of the analogues screened the most promising was compound **94**. Compound **94** had a measured IC_{50} of 150 μM with no inhibition observed in the counter screen. This was more potent than compound **86** (IC_{50} = 260 μM) and suggests that *N*-acylation of compound **86** leads to an increase in potency. However, comparison with *N*-acyl compound **97** (IC_{50} = 750 μM) which contains a phenyl ring in place of a tetrahydropyran suggests that any increase in potency is dependent on the nature of the acyl group.

N-alkyl compounds such as compounds **93** (IC_{50} = 77 µM) and **96** (IC_{50} = 110 µM) showed an even greater increase in potency than *N*-acyl compound **94**. However, these compounds also showed activity in the counter screen assay making it difficult to determine whether the inhibition seen in the PHF8 assay is genuine.

Synthesis of Analogues of Compound 86

The analogues of compound **86** purchased for testing gave some insight into the effects of modifications around the ketopiperazine core on PHF8 inhibition; however, no compounds with modifications on the phenyl ring of compound **86** were available for purchase. Therefore a synthetic strategy was devised to allow the parallel synthesis of multiple analogues of compound **86** with modifications on the phenyl ring.

Diethyl phthalimidomalonate **100** is a commercially available masked amino acid. This substrate has been previously shown to act as a nucleophile in S_N2 substitution reaction with alkyl iodides and allyl bromides.[205,206] Given this pattern of reactivity it was hypothesised that benzyl bromides would also act as a suitable substrate for nucleophilic substation. Given the wide availability of benzyl bromide electrophiles a wide range of analogues of compound **101** could be prepared. Hydrolysis- decarboxylation of the malonate and removal of the phthalimide protecting group would then give analogues of amino acid ester **102** (Scheme 4). These analogues could then be used as a starting material for ketopiperazine synthesis as outlined in Scheme 3.



100 **101** **102**

Scheme 4. Proposed synthetic route to phenyl substituted analogues of phenylalanine. Such analogues could be used for ketopiperazine synthesis using the route outlined in Scheme 3

A selection of sixteen benzyl bromides were used (Table 20). These contained small, common substituents at *ortho-*, *meta-*, and *para-* positions. These were chosen as they were all commercially available and would provide sufficient diversity in order to investigate SAR in the final compounds.

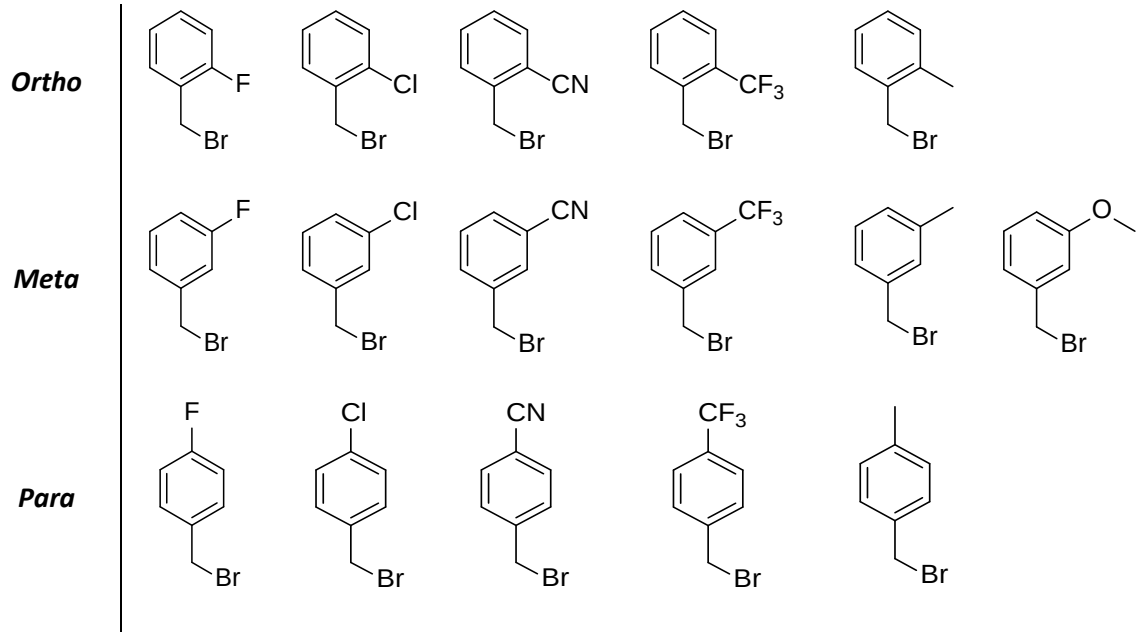
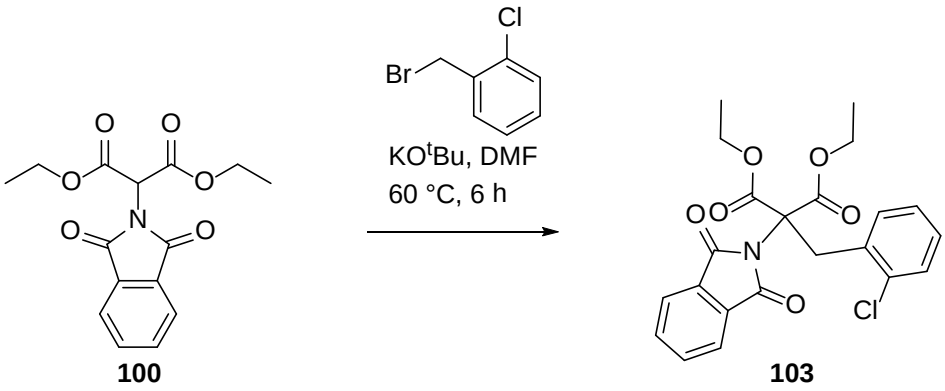


Table 20. Commercially available benzyl bromides chosen for benzylation of diethyl phthalimidomalonate **100**.

The benzylation of diethyl phthalimidomalonate **100** proved suitable for parallel chemistry as diethyl phthalimidomalonate **100** and potassium *tert*-butoxide could be pre-prepared as solutions in DMF. These solutions could be mixed prior to the addition of the appropriate benzyl bromide allowing for multiple reactions to be set up in a fast and simple procedure (Scheme 5).

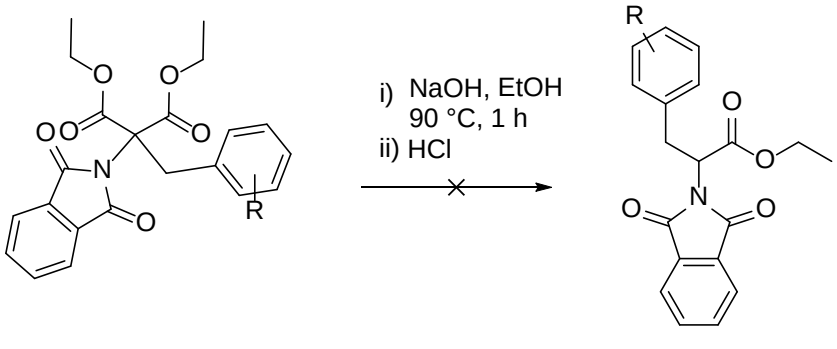
The alkylation reactions were worked up in parallel using phase separators equipped with hydrophobic frits and the resulting crude residues proved suitable for use without further purification. Overall the procedure allowed for sixteen reactions to be completed with only two evaporation steps and without the need for column chromatography in yields of typically 60-80%.



100 **103**

Scheme 5. Example reaction of diethyl phthalimidomalonate **101** with **a** substituted benzyl bromide (85% yield).

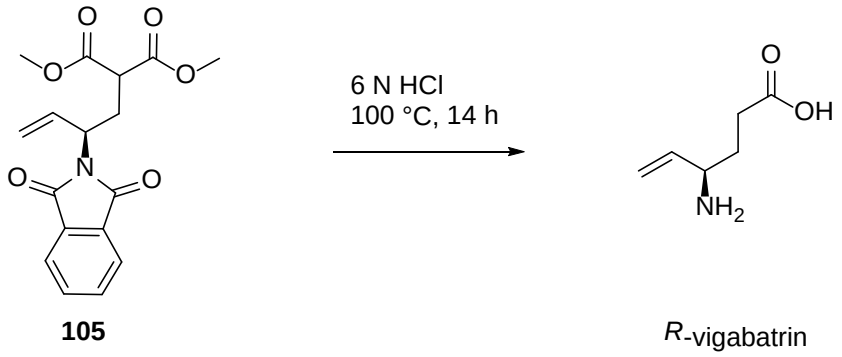
Following the successful synthesis of sixteen analogues of compound **101**, reaction conditions for the hydrolysis-decarboxylation of the malonate and removal of the phthalimide protecting group were investigated.



101 **104**

Scheme 6. Basic hydrolysis followed by acidic work did not deliver analogues of **104** in good yield and purity. Partial hydrolysis of the phthalimide group was observed, and malonate hydrolysis did not go to completion.

Initially, a procedure to hydrolyse-decarboxylate analogues of compound **101** to give analogues of compound **104** was sought. This was on the basis that compound **104** is uncharged and was therefore hypothesised to be amenable to column chromatography. A procedure based on a method described in a patent granted to GlaxoSmithKline in 2007 was used,[207] but LCMS analysis of the reaction mixture indicted that hydrolysis of the malonate was incomplete and the phthalimide was partially hydrolysed (Scheme 6). Therefore this approach was rejected and an alternative method for proceeding from analogues of **101** sought.

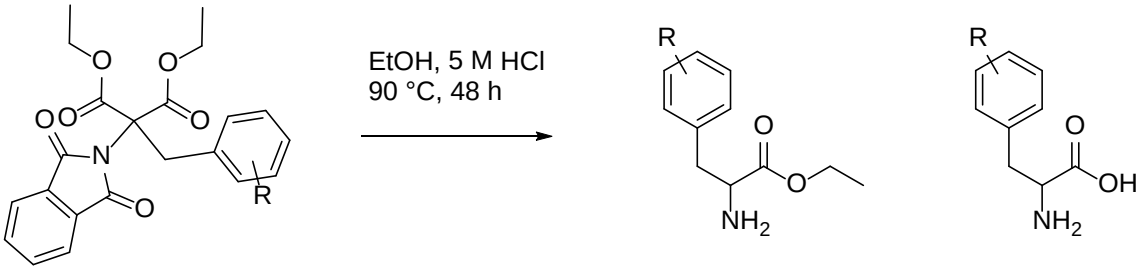


105 *R*-vigabatrin

Scheme 7. Final step in the synthesis of *R*-vigabatrin described by Trost et al. This procedure removes the phthalimide protecting group and performs a hydrolysis-decarboxylation of the malonate in a single step.

Trost et al. have published a synthesis of the anti-epileptic gamma-amino acid vigabatrin.[208] In this synthesis the final step is the removal of a phthalimide protecting group and hydrolysis-decarboxylation of a malonate in a single step (Scheme 7). This procedure was hypothesised to be suitable for the hydrolysis of analogues of compound **101**. However, it does have the disadvantage of hydrolysing both esters of the malonate and therefore requiring an esterification of the product amino acid before use as a starting material in the ketopiperazine synthesis shown in Scheme 3.

The analogues of **101** were found to be sparingly soluble in 5 M HCl, and were therefore dissolved in neat EtOH prior to dilution with acid. Monitoring of the reaction by LCMS suggested that the presence of EtOH did not inhibit the hydrolysis of the phthalimide group, but did slow the hydrolysis of the malonate. It was therefore necessary to allow the EtOH to evaporate in order to drive the reaction to completion (Scheme **8**). The presence of EtOH resulted in mixtures of amino acid ethyl esters **106** and amino acids **107** being produced. Nitrile containing compounds were found to be incompatible with the reaction conditions and were not carried forward to subsequence steps.



101 **106** **107**

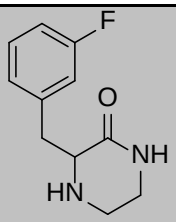
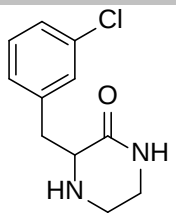
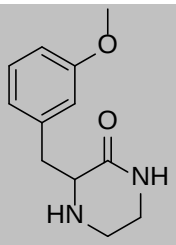
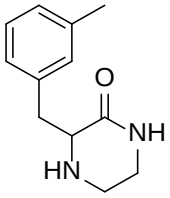
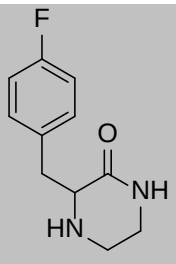
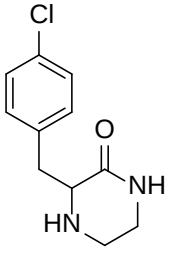
Scheme 8. Acidic hydrolysis of analogues of **101** gave mixtures of amino acid ethyl esters **106** and amino acids **107**.

This mixture of products did not prove to be problematic as it was possible to esterify the acid **108** in the presence of the ethyl ester **107** to give homogenous samples of amino acid ethyl esters for use in ketopiperazine synthesis.

Of the remaining thirteen amino acid ethyl esters, seven were successfully carried through to the corresponding ketopiperazine. Of the six compounds that failed, two were due to mechanical loss, with four others failing at the reductive alkylation step.

Analogue SAR

The seven ketopiperazine compounds were tested in the PHF8 AlphaScreen assay and also in the corresponding counter screen (Table 21). None of the compounds tested showed evidence of being active. Only compound **111** showed activity in the PHF8 assay, but also showed similar activity in the counter-screen.

Compound Number	Structure	IC_{50} (µM)	IC_{50} Upper 95% Confidence Limit	IC_{50} Lower 95% Confidence Limit	Counter Screen Inhibition (IC_{50})
108		No activity			No activity
109		No activity			No activity
110		No activity			No activity
111		360 µM	410 µM	280 µM	350 µM
112		No activity			No activity
113		No activity			No activity

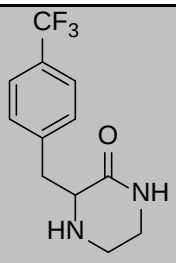
Compound Number	Structure	IC_{50} (μM)	IC_{50} Upper 95% Confidence Limit	IC_{50} Lower 95% Confidence Limit	Counter Screen Inhibition (IC_{50})
114		No activity			No activity

Table 21. Analogues of compound **86** were synthesised and screened using the AlphaScreen assay. The compounds were also tested using the dually His6 tagged and biotinylated peptide counter screen assay. Compounds were synthesised and screened as racemates.

This data suggest that substitution on the phenyl ring of compound **86** leads to a loss of activity. The lack of *ortho* substituted compounds in the library means that the possibility that *ortho* substitution would lead to an increase in activity cannot be ruled out.

Secondary Assays for Ketopiperazine Compounds

A secondary assay was sought to test ketopiperazine **86**, its enantiomer ketopiperazine **89**, and the tetrahydropyran-4-carboxylate ester **94**. A prototype mass spectrometry assay in development by Finn Wolfreys (Structural Genomics Consortium, University of Oxford) was tested. The assay involves incubating an enzymatically active PHD-JmjC construct of PHF8 with a substrate peptide, the necessary co-factors, and varying concentrations of the inhibitor to be tested. After a suitable incubation period the enzymatic reaction is quenched by the addition of formic acid and the ratio of substrate peptide to product peptide measured by mass spectrometry.

The compounds discovered by the PHF8 AlphaScreen assay should prevent peptide binding to the PHD-JmjC of PHF8. These compounds should therefore inhibit the enzymatic conversion of a

substrate peptide containing H3K4me3 and H3K9me2 marks to a product peptide containing H3K4me3 and H3K9me1 marks.

The prototype assay did not prove suitable, probably due to the low enzymatic activity of the recombinant PHF8 enzyme used coupled to the ion suppressing effects of DMSO. These effects combined to result in a low product peptide signal, which did not provide a suitable signal to noise ratio for testing compound inhibition.

Summary and Future Work

The SiteMap analysis of PHDs carried out in Chapter 3 lead to the hypothesis that PHDs are more ligandable when found in tandem with another domain. Chapter 3 also described an analysis of novel potential small molecule binding sites at domain-domain interfaces in epigenetic proteins. The work carried out in this chapter describes screening efforts and subsequence hit optimisation for the PHD-JmjC of PHF8. An AlphaScreen assay developed at the Structural Genomics Consortium was used to screen a fragment library and a library designed by virtual screening against the PHD-JmjC of PHF8.

These screening efforts sought to identify compounds that bound at the PHD-JmjC interface and therefore prevented these tandem domains engaging their histone peptide substrate. The targeting of this peptide binding site which is found only in PHF8 would allow inhibitors to be developed that show a greater degree of selectivity than inhibitors that target the 2OG binding site.

Ketopiperazine **86** was identified from a fragment library as a PHF8 inhibitor. The enantiomer of this compound was synthesised and tested, with its inactivity suggesting that compound **86** was a genuine inhibitor rather than an assay interference compound. A series of analogues exploring SAR around the ketopiperazine ring were selected for purchase and tested. Another set of compounds with substitutions on the phenyl ring were synthesised in parallel using a six step

synthetic route. Substitutions on the phenyl ring resulted in a loss of inhibitory activity against PHF8; however, tetrahydropyran-4-carboxylate ester **94** showed an increase in activity compared to compound **86**.

Further development of the mass spectrometry assay described above would allow these compounds to be confirmed as PHF8 inhibitors. By varying the concentrations of 2OG and substrate peptide a mass spectrometry assay could be used to confirm the hypothesis that these compounds are substrate competitive rather than 2OG competitive. A co-crystal structure would also confirm the binding location of this series of inhibitors and provide structural information to inform compound optimisation.

Overall this chapter describes the discovery of a series of ketopiperazine containing fragments which inhibit PHF8. It is hypothesised that these compounds bind at the domain-domain interface of the PHD and JmjC domain. This is consistent with the hypothesis made in Chapter 3 that sites found at domain-domain interfaces are ligandable.

Chapter 6 – Summary and Conclusions

This thesis set out to provide a family wide analysis of human PHDs and to assess the potential of members of the PHD family to bind small molecule inhibitors. The purpose of such work was to allow prioritisation of which PHDs should be subjected to screening campaigns to help accelerate the development of chemical probes targeting this family of protein domains.

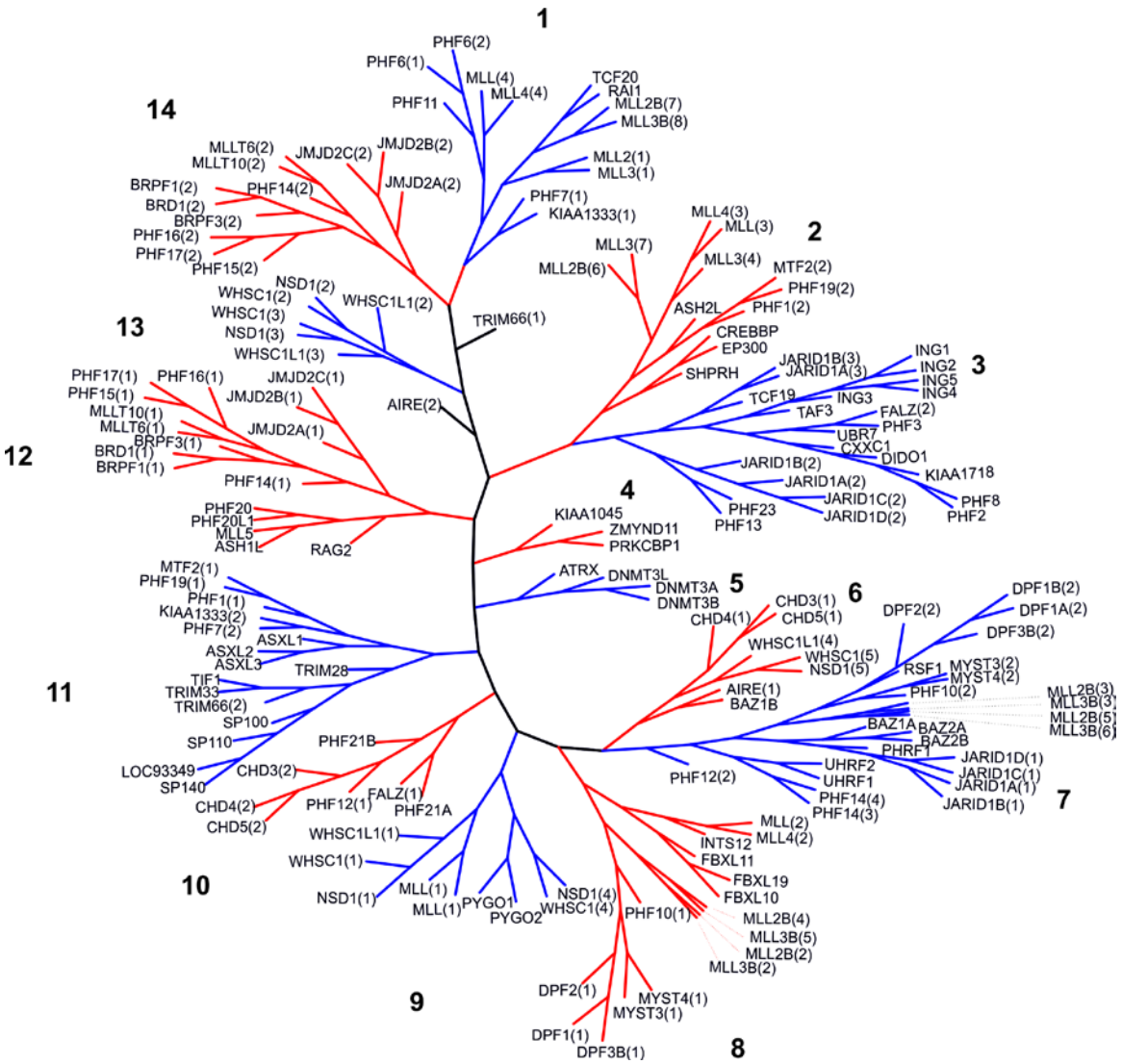


Figure 99. Phylogenetic tree of the human PHD family constructed in Chapter 2. The 14 sub-families of the PHD tree are labelled and alternatively coloured red and blue. TRIM66(1) and AIRE(2) are atypical and not part of any sub-family.

This thesis starts by defining a set of PHDs within the human proteome, this is the first time that such a list of PHDs has been published and includes full details of how it was compiled. 173 unique human PHDs were identified, and an innovative knowledge based alignment method was

used to establish a phylogenetic tree (Figure 99). The phylogenetic tree constructed as a result of this work can be divided into fourteen sub-families within which is it possible to identify conserved residues that can be used to predict the binding specificity of PHDs within that sub-family. For example, sub-family 3 members contain residues capable of forming an aromatic cage for H3K4me3 recognition (Figure 100).

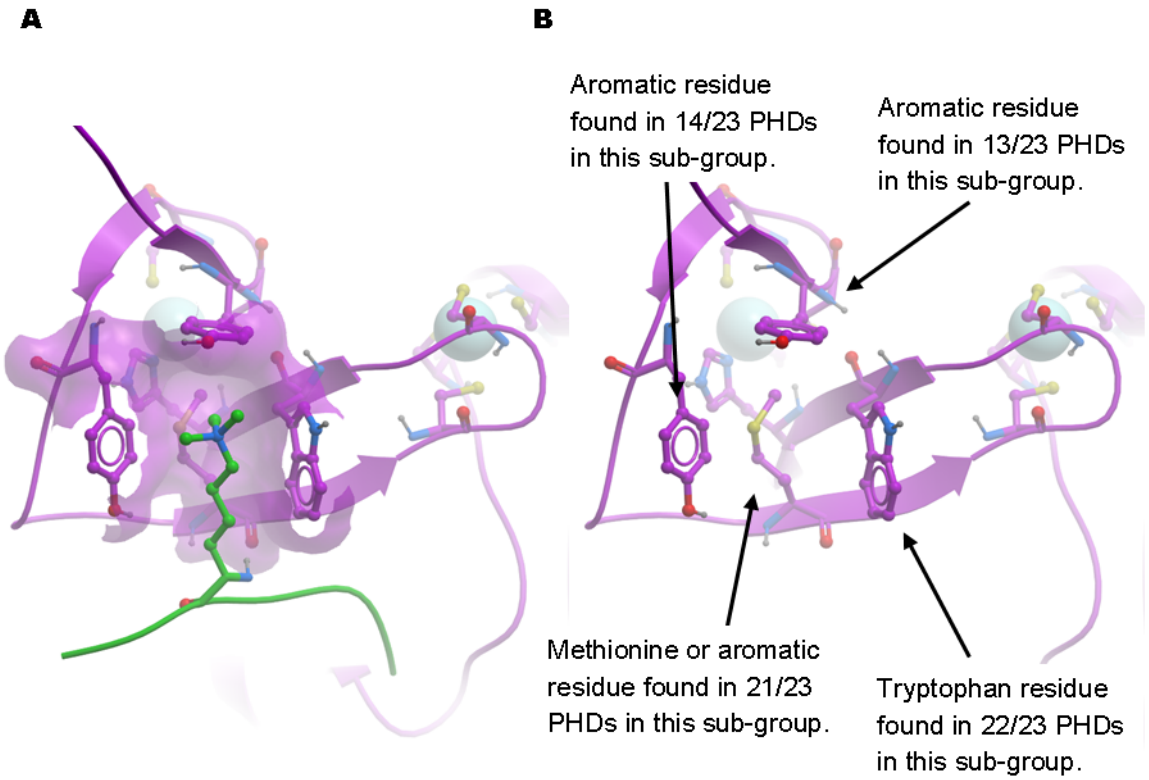


Figure 100. Depiction of aromatic cage residues found in sub-family 3. **A** The trimethylated lysine of histone 3 (green) is shown in a hydrophobic pocket formed by three aromatic residues and a methionine. **B.** Depiction of the same aromatic cage with the histone 3 peptide removed. Each key cage-forming residue is labelled with its prevalence within sub-family 3. The PHD shown is the PHD of PHD Finger Protein 8 (PHF8, PDB ID: 3KV4).

The phylogenetic tree presented in Chapter 2 will aid the development of screening panels for determining specificity for future development of PHD inhibitors and provide a valuable resource for those studying the biochemistry of these domains.

A key conclusion from this family wide analysis is that the majority of PHDs are not predicted to be H3K4me3 binders. This prediction can be made on the basis that PHDs with aromatic cage

residues known to be crucial to H3K4me3 recognition are clustered within sub-family 3. For many of the remaining members of the family, H3 engagement is dependent on the methylation state of H3K4, as tri-methylation at this position prevents binding. These PHDs should therefore be thought of as 'K3K4me0 readers'. This contrasts widely held view that PHDs are 'methyl lysine readers', the results of this family wide analysis suggests that it would be more correct to describe PHDs in general as 'readers of the modification state of residues near the *N*-terminus of histone 3'.

Following on from the identification of 173 human PHDs, an analysis of the family was carried out to determine which PHDs were more amenable to inhibition by small molecule ligands. The large size of the PHD family precludes the use of purely experimental means to survey ligandability; therefore a computational method was sought. The Schrödinger SiteMap program[126,130] was chosen for this purpose as it had previously been used to studying the ligandability of other epigenetic reader families, allowing for direct comparison between PHDs and other epigenetic reader domains.[78,128,129]

The analysis of individual PHD domains suggested that in comparison to other epigenetic reader domains, PHDs are difficult targets for small molecule inhibition. However, it was possible to identify some PHDs that were more ligandable than others. A ranking of PHDs by median DScore[130] shows that the closely related PHDs of PYGO1 and PYGO2 are ranked fourth and third respectively in terms of PHD ligandability (Figure 101). Since the completion of this work these two PHDs were the targets of the first reported small molecule inhibitors with structural evidence of binding.[74] This suggests that it is possible to discover small molecule ligands at least for PHDs towards the top of the ligandability ranking.

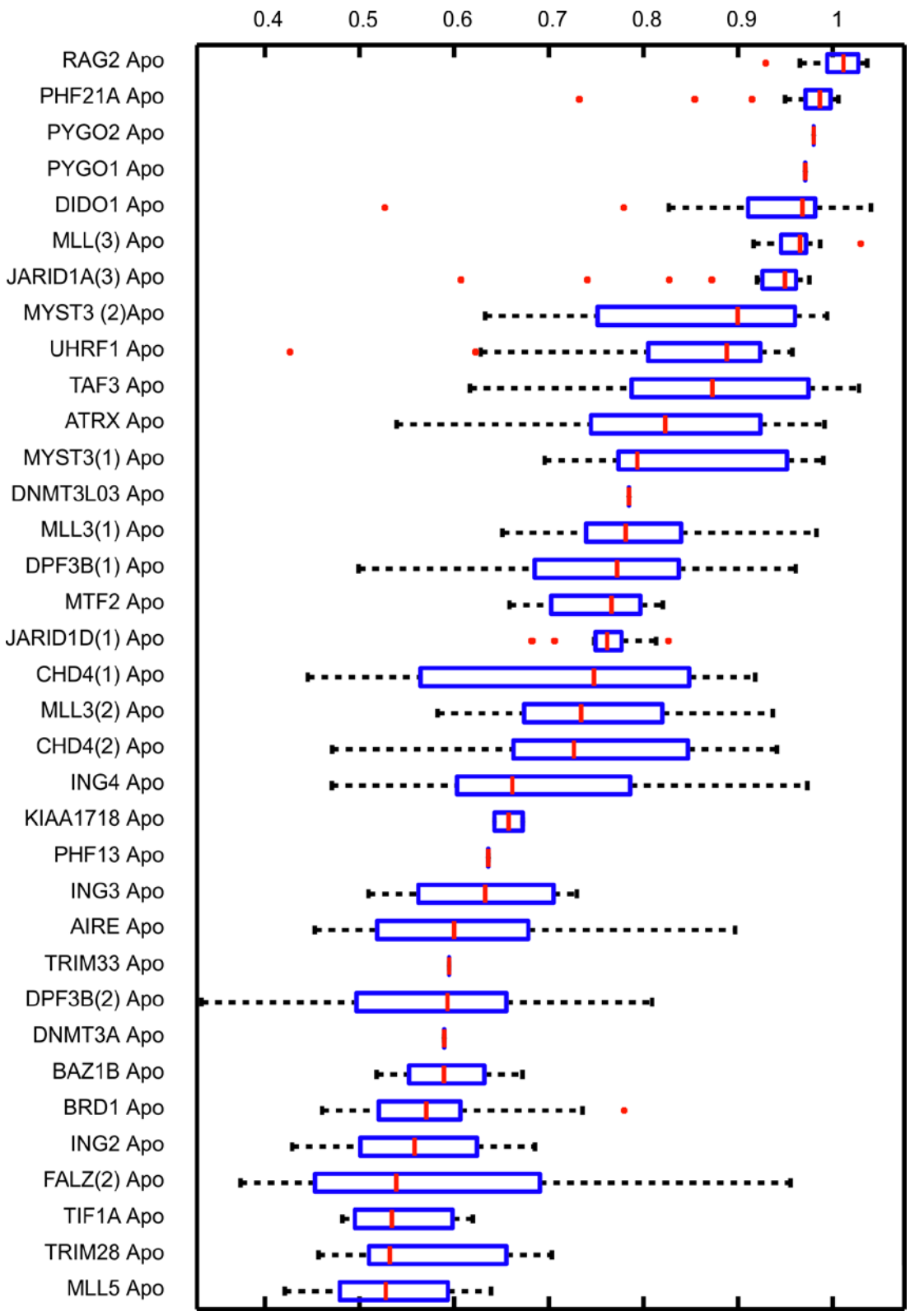


Figure 101. Box plots showing the range of DScores for each *apo* PHD, calculated using Parameter Set 5 (Chapter 3). The plots are sorted by the median value with the most ligandable number at the top. The box plots are marked with the median value and the edges of the box represent the inter-quartile range. The whiskers extend to the most extreme data not considered an outlier, and outliers are plotted individually. An outlier is classed as any data point more than 1.5 inter-quartile ranges below the first quartile or above the third quartile.

As many PHDs are known to occur near other epigenetic domains including another PHD, a SiteMap study was conducted on tandem PHD domains. This suggested that tandem PHDs are more ligandable than single PHDs, with larger, more enclosed binding sites which are less hydrophilic than binding sites found in single PHD domains (Figure 102).

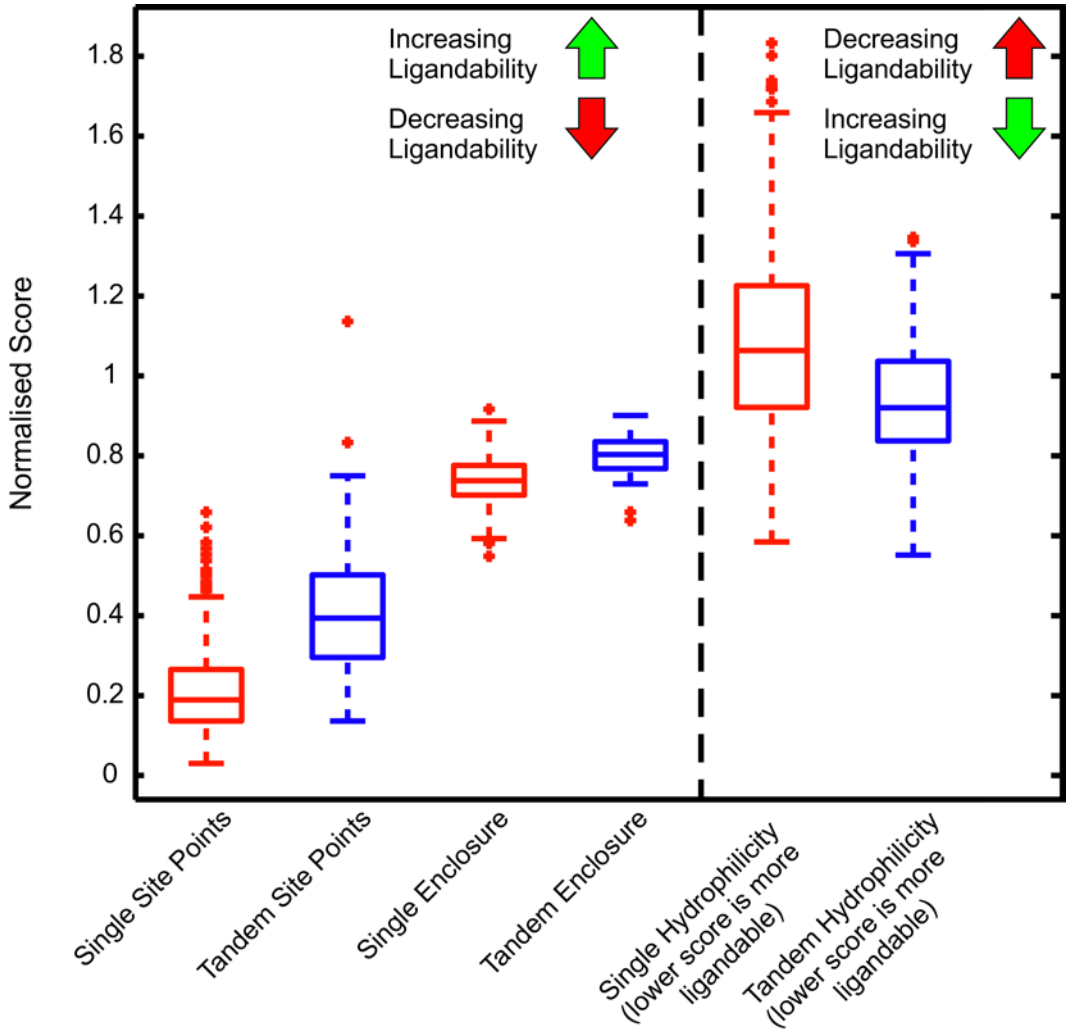


Figure 102. Box plots showing the number of site points, the enclosure, and the hydrophilicity score for tandem and single PHDs. These results predict that tandem PHDs are more ligandable than single PHDs. All values have been normalised with respect to the mean values for these parameters derived from 342 sites with submicromolar inhibitors used to train DScore and SiteScore. The box plots are marked with the median value and the edges of the box represent the inter-quartile range. The whiskers extend to the most extreme data not considered an outlier, and outliers are plotted individually. An outlier is classed as any data point more than 1.5 inter quartile ranges below the first quartile or above the third quartile.

The results suggesting that tandem PHDs are more ligandable than single PHDs led to two further computational investigations. The first sought to investigate whether the pattern of tandem domains being more ligandable than single domains was also observed for Tudor domains. The second sought to discover novel potential small molecule binding sites at domain-domain interfaces in other epigenetic proteins. The study of Tudor domains suggested that the methyl lysine binding aromatic cage found in most Tudor domains was too small to act as a ligandable site alone. In contrast, potentially ligandable sites were identified in the multiple Tudor domains of SETDB1, TP53BP1, and SPIN1 spanning both methyl lysine binding sites.

The investigation of domain-domain interfaces revealed a potential small molecule binding site at the interface of the bromodomain and PWWP domain of ZMYND11. This site specifically binds residues 29-31 of the H3.3 isoform allowing for differentiation over the more common H3.1 isoform. This study also identified an inter-domain binding site in the PHD-bromodomain of SP100 (Figure 103). A fragment soaking assay performed by colleagues at the Structural Genomics Consortium identified three ligands which bound at this site. Although these ligands would not be expected to orthosterically compete with the natural histone peptide ligand, they may be allosteric inhibitors. Although inhibitory activity has not been shown for these ligands, the crystallographic evidence of their binding at the site identified by SiteMap proves some validation of this computational method. This crystallographic evidence also offers a proof of principle that sites at domain-domain interfaces are ligandable.

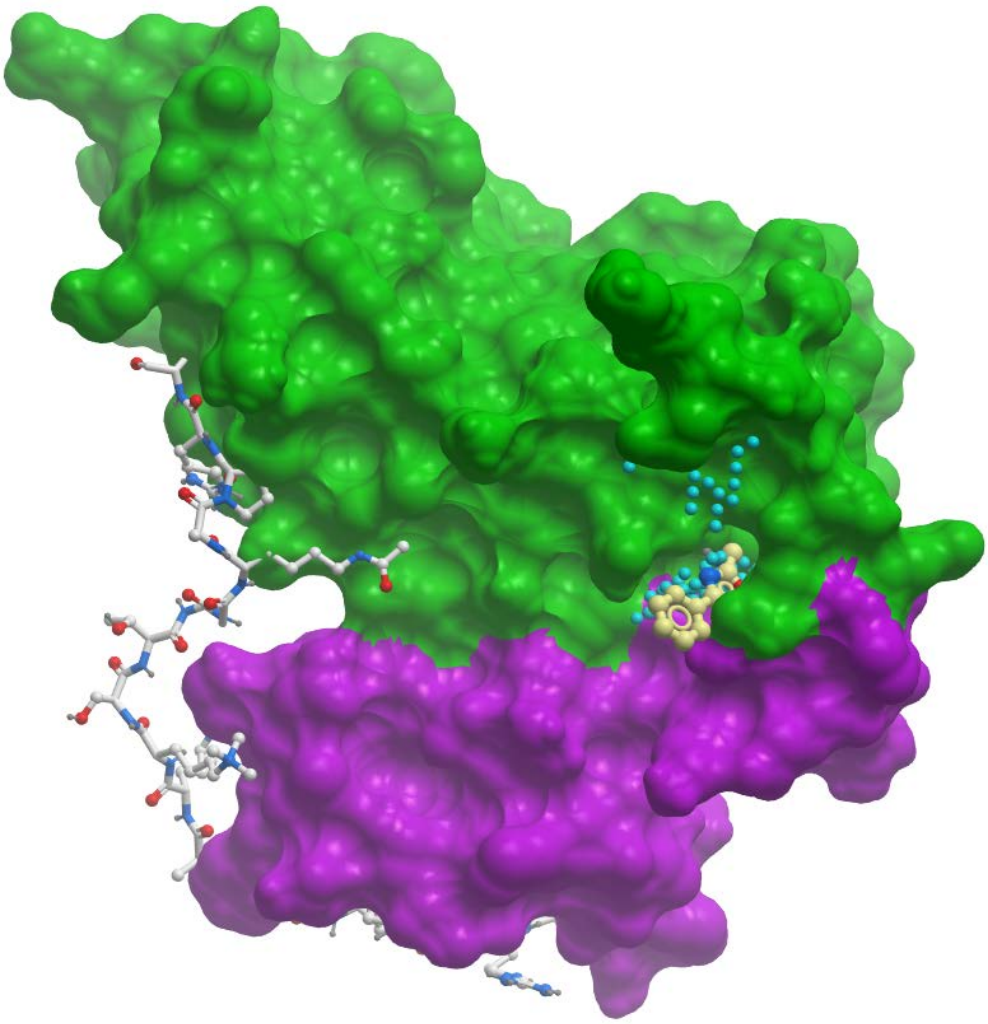


Figure 103. The PHD-bromodomain of SP100 is shown with the site points as identified by SiteMap (cyan). Ligand **5** was identified by a crystal soaking experiment, and binds in this predicted site (yellow). The PHD is shown in magenta, and the bromodomain is shown in green. The peptide ligand shown is taken from the structure of the related PHD-Brd of TRIM33. PDB ID: 3U5M.

As well as the computational investigations summarised above, the result that tandem PHDs are more ligandable than single PHDs lead to experimental efforts to discover small molecule inhibitors of tandem PHDs. Chapter 4 described the design and optimisation of an AlphaScreen assay suitable for identifying molecules which inhibit the binding of the tandem PHD of DPF2 to histone 3 peptides (Figure 104).

Owing to the high false positive rates of AlphaScreen assays[150] a suitable secondary assay was sought in order to confirm any primary hits identified in the AlphaScreen assay. Nuclear Magnetic Resonance (NMR) and crystal soaking assay were investigated as these assays provide direct evidence of ligand binding and can give structural information which may be used for ligand optimisation.

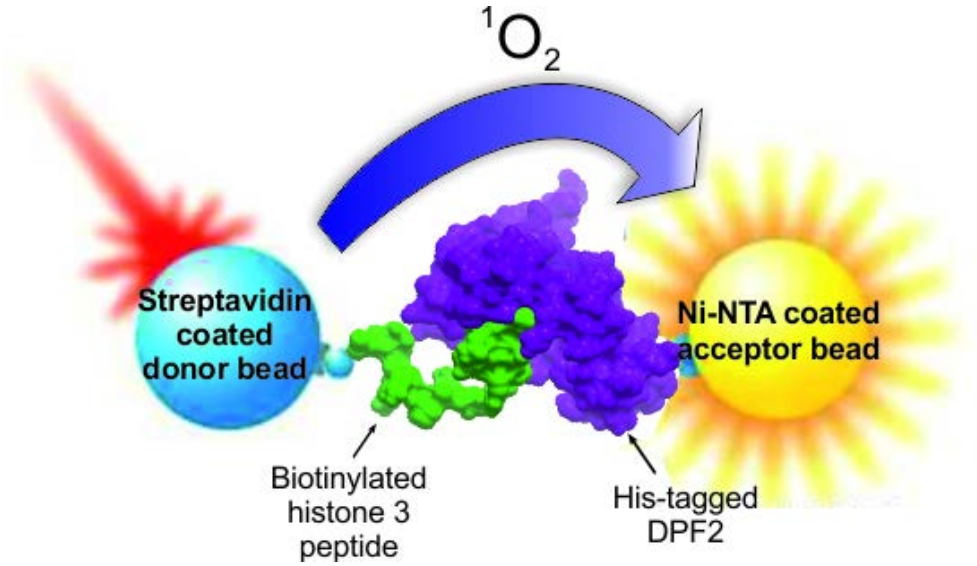


Figure 104. A schematic representation of an AlphaScreen assay consisting of a biotinylated peptide and His6-tagged protein. When the protein and peptide are in contact they bring the donor and acceptor beads close together. When the donor bead is excited with red light, single oxygen is generated, which diffuses to the acceptor bead causing an emission between 520 nm and 620 nm. This assay was used as a primary assay in screening efforts against DPF2.

Development of an NMR assay was hindered by the failure to express DPF2 in minimal media therefore efforts were focused on the development of a crystal soaking assay. Initial attempts to crystallise DPF2 were unsuccessful, but a crystallisable construct of the closely related tandem PHDs of DPF3b was provided by Jinrong Min (Structural Genomics Consortium, University of Toronto). Using this construct crystals of DPF3b were produced which allowed the structure to be solved to a resolution of 1.75 Å (Figure 105). However, these crystals could not be produced in a reproducible, solvent tolerant manner. Therefore a crystal soaking assay for DPF3b could not be developed.

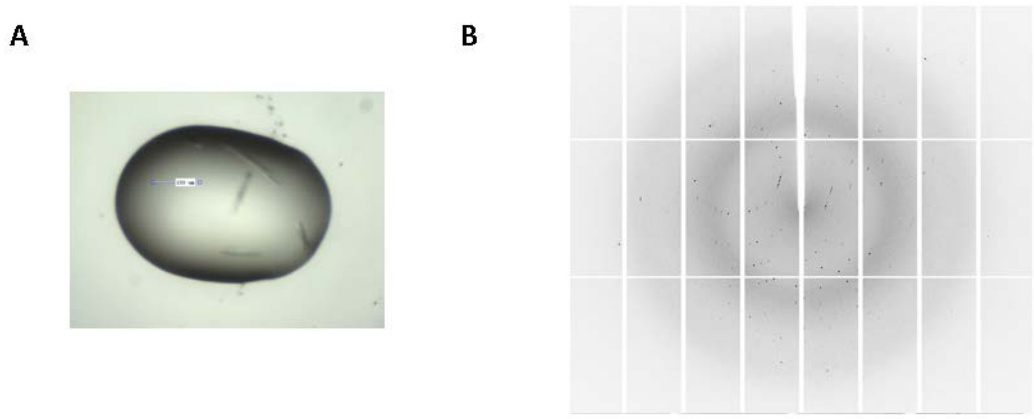


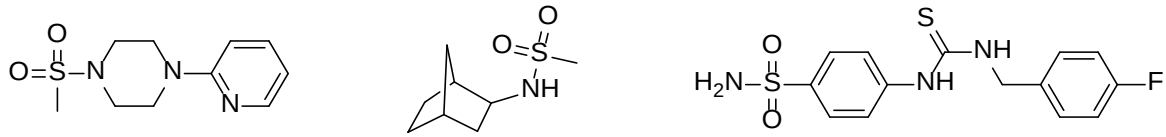
Figure 105. A. Crystals of DPF3b produced in crystallisation conditions based on those developed by SGC Toronto. **B.** Diffraction pattern of DPF3b crystals showing diffraction to 1.75 Å.

Several small molecule libraries were collated for screening against DPF2 using the AlphaScreen assay. A collection of 1324 fragments were screened at 2 mM which revealed a series of *N*-mesyl compounds as promising hits. However, resynthesised batches of the two most potent *N*-mesyl compounds **11** and **14** showed no activity (Figure 106).

A 10,000 member library was screened using several virtual screening techniques to prioritise the library for experimental screening using the AlphaScreen assay. The virtual screening method used was validated on the triple Tudor domains of SPIN1, using a ligand set with known active compounds. This virtual screen ranked fourteen of the twenty-four confirmed hits within the top 5% of ranked compounds.

The tandem PHDs of MYST3 were used as a surrogate for DPF2 as no structure for these tandem PHDs were available. The compounds were screened using Glide[164] and 437 compounds chosen for experimental screening. A further two libraries were designed by a collaborator (Jan Domanski; Department of Biochemistry, University of Oxford). This work used the same MYST3 crystal structure but subjected it to molecular dynamics (MD) simulation. Transient pockets were identified in the MD trajectory that were not seen in the static crystal structure which were used for virtual screening. This work provided two further libraries consisting of 398 and 383 compounds respectively. Compound **36** was discovered as a result of this screen. After the

screening of three purchased analogues, twenty-two analogues were synthesised in parallel with the resynthesis of compound **36** and two of the purchased compounds. Unfortunately none of these compounds showed activity.



9 **12** **36**

Figure 106. Compounds identified by the DPF2 AlphaScreen assay that showed no activity when resynthesised.

Although the no inhibitor of DPF2 was discovered during the work described in Chapter 4, these negative results are not sufficient to suggest that DPF2 is unligandable. It is possible that an alternative primary assay may be more suitable for identifying weak fragment hits which could be optimised to give potent inhibitors. The use of alternative constructs or surface mutations may facilitate more consistent crystallisation of DPF3b for use in crystal soaking assays, and use of alternative plasmid vectors or expression organism may allow for ^{15}N-labelled DPF2 expression for NMR assays. These may offer more suitable primary assays that the AlphaScreen assay as they offer direct evidence of binding. Another possibility for DPF2 inhibitor design would be to use a peptide mimetic approached similar to the one used for the production of inhibitors of the chromodomain of CBX7.

A previously identified example of a ligandable PHD found at a domain-domain interface is that of PHD-JmjC of PHF8.[78] This site is involved in the recognition of H3K4me3, and binding of H3K4me3 at this site directs the nearby H3K9 towards the catalytic site of the JmjC domain, which is capable of demethylating H3K9me2 (Figure 107). The domain orientation of the PHD and JmjC of PHF8 is different from that seen in closely related PHD-JmjCs.[111] The PHDs of the other members of the KDM7 demethylase family - PHF2 and KDM7A (KIAA1718) - are very similar to the PHD of PHF8. Similarly the JmjC of PHF8 is similar to the JmjCs found in PHF2 and KDM7A, as well as JmjCs found in the KDM2 sub-family.[180] Previously published inhibitors

targeting the JmjC only of these demethylases show poor selectivity between members of the KDM2 and KDM7 families.[185,199] Therefore targeting the unique PHD-JmjC interface of PHF8 is a potential solution for the development of selective PHF8 inhibitors.

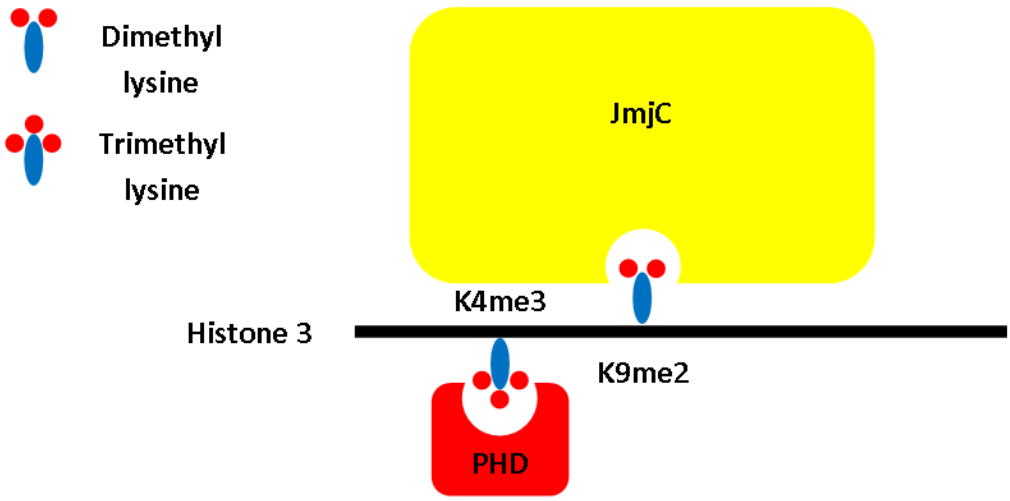
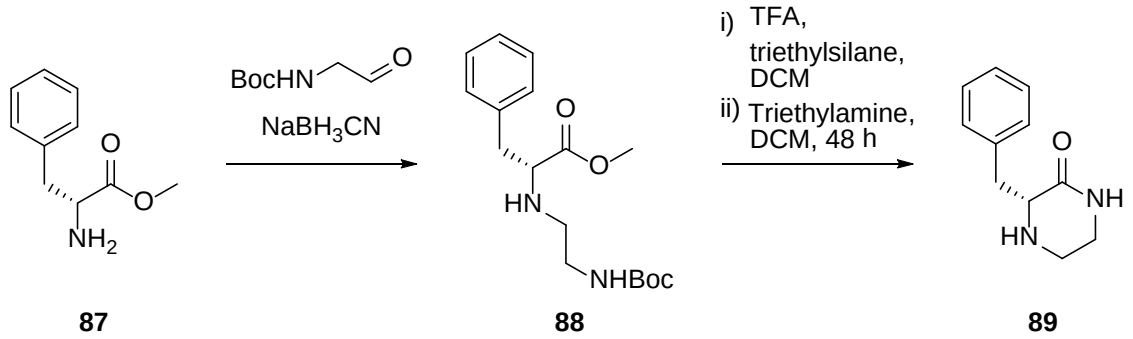


Figure 107. Schematic representation of the PHD and JmjC of PHF8. H3K4me3 binds at a site at the interface of the PHD and JmjC. This directs H3K9me2 towards the catalytic site of the JmjC domain.

An AlphaScreen assay for PHF8 had previously been developed within the department. This uses a H3K4me3K9me0 peptide, as the binding to PHF8 is expected to be driven almost entirely by the interaction of H3K4me3 at the PHD-JmjC interface. Therefore any compound that displaces this peptide is likely to be binding at the PHD-JmjC interface, rather than in the JmjC's active site. This assay was used to screen the same fragment library used for DPF2, as well as a 233-member library designed for PHF8 by virtual screening using Glide.

Once again, no hits were identified from the virtual screening library, despite the method being validated using SPIN1. It is likely that the failure to identify active compounds in the DPF2 and PHF8 virtual screening libraries may be down to the size and identity of the library used in the virtual screen.

Despite the failure of the virtual screening library to produce hits, the fragment screen performed on PHF8 identified ketopiperazine **86** as a potential fragment inhibitor (IC_{50} = 260 µM). In order to confirm this as a genuine hit rather than an assay interference compound its enantiomer, compound **89**, was synthesised (Scheme 9) and tested (Figure 108).



Scheme 9. Synthesis of the enantiomer of compound **86**. D-Phenylalanine methyl ester **87** was used as the amine partner in a reductive alkylation with *N*-Boc-2-aminoacetaldehyde. The Boc protecting group of the resultant product was removed by hydrolysis with trifluoroacetic acid and the resulting primary amine cyclised in the presence of triethylamine.

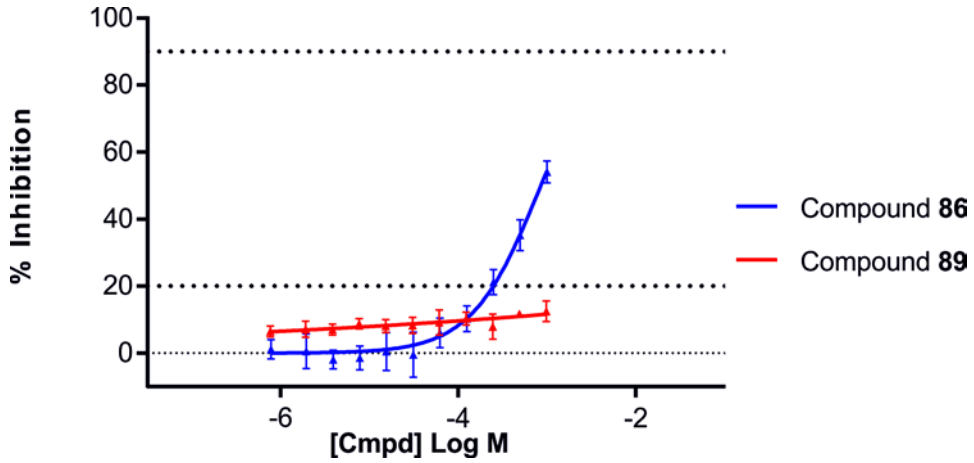
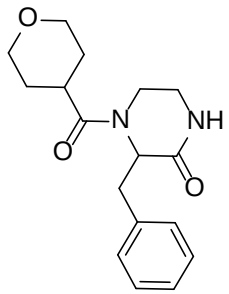


Figure 108. Comparison of the activities of compound **86** and its enantiomer compound **89** as measured by the PHF8 AlphaScreen assay. Although a full binding curve could not be obtained for compound **86**, it clearly shows greater activity that its enantiomer.

The inactivity of the enantiomer suggests that compound **86** was inhibiting the PHF8 AlphaScreen assay by binding to PHF8, rather than by an assay interference method which would be independent of the chirality of the molecule.

On this evidence, a series of analogues were purchased to explore SAR around the ketopiperazine ring, and a series of analogues synthesised to establish SAR around the phenyl ring. Substitution at the *meta*- and *para*-positions lead to a loss of activity, and of the purchased analogues only compound **94** showed increased activity of the original hit (Figure 109).



94
IC_{50} = 150 µM

Figure 109. Compound **94** was the most potent compound in the PHF8 AlphaScreen assay.

The lack of secondary assay meant that these compounds could not be investigated further. However, it is hoped that the continued development of a PHF8 mass spectrometry assay within the group may allow these compounds to be studied further. Of particular interest will be the use of such an assay to determine if these compounds are peptide competitive.

This thesis set out to investigate the ligandability of human PHDs. Initially a definitive list of human PHDs was compiled, which was aligned and used to estimate a phylogenetic tree of human PHDs. Following the definition of a definitive PHD family tree, a ligandability analysis was performed using SiteMap which predicted that PHDs that form part of multi-domain complexes are more ligandable that single PHDs. A collaboration with colleagues at the Structural Genomics Consortium identified four small molecule ligands that bind at the PHD-bromodomain interface of SP100 at a binding site predicted by SiteMap. The tandem PHDs of DPF2 were investigated,

and although efforts to discover small molecule inhibitors of the tandem PHDs of DPF2 proved unsuccessful, the scale of the screening performed is not sufficient to suggest that these tandem PHDs are unligandable. A similar screening campaign was carried out against the PHD-JmjC of PHF8. This led to the discovery of ketopiperazine **86** as a putative inhibitor. It is hoped that this may lead to the development of a peptide competitive inhibitor of PHF8 that would be highly selective over other JmjC containing enzymes.

General Experimental Details

Virtual Screening

Library Filtering

All ligand sets to be used in virtual ligand screening were filtered to remove compounds with reactive groups using the Schrödinger REOS (Rapid Elimination of Swill) node in Knime version 2.8.2.[209]

Ligand Preparation

3-dimensional ligand structures were prepared from a 2-dimensional SDF file using Schrödinger LigPrep version 2.6 (Schrödinger, LLC, New York, NY, 2013). Structures were generated in a desalted form with the ionisation states set at pH 7. Structures were minimised using the OPLS_2005 force field.

Virtual Ligand Screening

All virtual ligand screening was carried out on an Intel® Core™i7-3770 CPU with 8 GB of RAM using Schrödinger Glide[163,164] version 5.9 (Schrödinger, LLC, New York, NY, 2013). The receptor site was defined around the co-crystallised peptide. Glide SP (Standard precision) was used. Ligands were allowed to be flexible, sampling nitrogen inversions and ring conformations. Epik state penalties were included in the docking score.[210]

AlphaScreen

To each well of a 384 assay plate was added the relevant volume of compound solution in DMSO. Protein and peptide were diluted in assay buffer (100 mM NaCl, 25 mM HEPES, 0.1% BSA, and 0.05% CHAPS at pH 7.5) to 1.66 times desired final assay concentration and 12 µL of this solution added to each well and the plates allowed to incubate at room temperature for 30 mins. 8 µL of a solution containing 12.5 µg/mL of donor and acceptor beads were added to each well and the plates incubated for 1 h at room temperature in the dark. Plates were read using a

Pherastar plate reader. IC_{50} measurements were recorded over eleven concentration points, each measured in duplicate.

Synthetic Organic Chemistry

Reactions involving moisture sensitive reagents were carried out under a nitrogen atmosphere using standard vacuum line techniques. Water was deionized by an Elga DV 25 system. All other solvents and reagents were used as supplied (analytical or HPLC grade) without prior purification. Organic layers were dried over Na_2SO_4. Thin layer chromatography was performed on aluminium plates coated with 60 F_{254} silica gel. Plates were visualised using UV light (254 nm) or 1% aq. $KMnO_4$. Flash column chromatography was performed on a Biotage Isolera One flash column chromatography platform. NMR spectra were recorded on a Bruker Avance spectrometer in the deuterated solvent stated. The field was locked by external referencing to the relevant deuteron resonance. Chemical shifts (δ) are reported in ppm and coupling constants (*J*) in Hz. Low-resolution mass spectra were recorded on a Waters SQ Detector 2.

LC/MS analysis was carried out on a Waters system equipped with a Waters 2545 Binary Gradient Module, a SecurityGuard™ ULTRA cartridges for EV0-C18 UHPLC, column guard a Kinetex 5 µm EVO C18 100 Å 100 x 3.0 mm column, and a Waters SQ Detector 2. The standard solvent gradient is described below. Solvent A consists of 93 % H_2O, 5 % acetonitrile, and 2 % 0.5 M ammonium acetate adjusted to pH 6 with glacial acetic acid. Solvent B consists of 18 % H_2O, 80 % acetonitrile, and 2 % 0.5 M ammonium acetate adjusted to pH 6 with glacial acetic acid.

Time (min)	Flow rate (mL/min)	Solvent A percentage	Solvent B percentage
Initial	2.00	95.0	5.0
0.35	2.00	95.0	5.0
1.35	2.00	5.0	95.0
2.10	2.00	5.0	95.0
2.20	2.00	95.0	5.0
3.00	2.00	95.0	5.0

The system also included a Waters 2489 UV/Visible Detector and a Waters 2424 ELS Detector. Retention times are reported based on the relevant peak in the 254 nm spectrum. All synthesised compounds submitted for screening showed no impurities in the total ion count chromatogram.

Preparatory scale LC/MS was performed on the same system equipped with a SecurityGuard™ PREP Cartridge column guard, a Kinetex 5 µm EVO C18 100 Å 150 x 21.2 mm column and a Waters 2767 Sample Manager. The standard solvent gradient is described below. Solvent A and Solvent B are as used for analytical LC/MS described above.

Time (min)	**Flow rate (mL/min)**	**Solvent A percentage**	**Solvent B percentage**
Initial	20.00	90.0	10.0
2.50	20.00	90.0	10.0
5.00	20.00	10.0	90.0
12.00	20.00	10.0	900
12.50	20.00	90.0	10.0
15.00	20.00	90.0	10.0

Melting point analysis was carried out on a Stuart SMP40 apparatus, with the reported melting point corresponding to the clear point as identified by the Stuart SMP40 system.

High resolution mass spectrometry data was collected on an Agilent 6530 QTOF. Predicted masses were calculated using http://www.sisweb.com/referenc/tools/exactmass.htm.

Infrared spectroscopy was carried out on a Thermo Scientific Nicolet iS5 FT-IR spectrometer fitted with a iD7-ATR accessory.

Experimental Details for Chapter 2 - PHD Family Analysis

HMMER and PSI-BLAST

HMMER version 3.0 (March 2010) was run on a local Linux server. The initial hidden Markov model was constructed using a multiple sequence alignment of forty-one PHDs whose structures had been deposited in the Protein Data Bank. The RefSeq database used was updated on 22/10/2012. Output alignments from HMMER were converted from Stockholm to FASTA format using the web service found at http://sequenceconversion.bugaco.com/converter/biology/sequences, and manually adjusted using ICM (Molsoft) and used for subsequent searches using HMMER. Subsequent protein BLAST searches were ran using Protein-Protein BLAST 2.2.28+.

Sequence Alignment

PHD sequences were aligned in ICM utilising ICM's structural superimposition function and secondary structure prediction from http://www.compbio.dundee.ac.uk/www-jpred. Sequence logos were generated using http://weblogo.berkeley.edu/logo.cgi.

Phylogenetic Tree Construction

Phylogenetic tree construction was performed using MEGA6 (Build#:6140226) on a Dell OptiPlex 9010 running Windows 7. The 'Find Best DNA/Protein Models (ML)' function was used to select a suitable substitution model. The phylogenetic tree was constructed with the WAG + G + I model with 500 bootstrap replications.

Experimental Details for Chapter 3 - Computational Assessment of the Ligandability of PHDs and Other Epigenetic Reader Domains

General Site Map Procedure

The general procedure for the use of SiteMap on an individual structure is described below. Structures were imported into the Schrödinger Maestro environment in pdb file format. Additional chains, domains, and water molecules were removed. In the case of the first bromodomain of Bromodomain Containing 1 (BRD1) five water molecules were left in place as these are known to be conserved across the bromodomain family. The structures were process using the Schrödinger Protein Preparation Wizard, and bound ligands removed. SiteMap was run using default parameters unless otherwise specified.

In the case of single PHD structures, the structures were superimposed using the ICM *superimpose* command with the alignment described in Chapter 2 used to guide the superimposition. The structures were then re-downloaded using PyMol; extra chains, extra domains, and waters removed; and superimposed with the identical structure created using ICM. This allowed ICM's superimposition tool to be used (which was found to be superior for superimposing similar sequences to PyMol's and Maestro's superimposition tools) while negating an intrinsic problem with .pdb files created by ICM (these files do not contain CONECT data which causes errors when the file is opened in Maestro). This facilitated separation of potential ligand binding sites at the protein-peptide interface from sites identified at other positions.

Schrödinger SiteMap (version 2.8, Schrödinger, LLC, New York, NY, 2013) was run from the command line using a Linux server, and data extracted from the output files using KNIME 2.8.2. Sites were clustered using their centroids in Matlab (©1984-2012 MathWorks, version R2012a) using the kmeans functions. This facilitated selection of potential ligand binding sites at the protein-peptide interface.

Experimental Details for Chapter 4 - Assessing the Ligandability of Tandem PHDs

Virtual Screening

Protein Structure Preparation

The structure of homologous tandem PHDs of KAT6A (MYST3, MOZ) was used as a surrogate for the tandem PHDs of DPF2, for which no structures were available. The structure (PDB ID 3V43) was prepared using Schrödinger Protein Preparation Wizard (Epik version 2.4; Impact version 5.9; Prime version 3.2, Schrödinger, LLC, New York, NY, 2013). Missing side chains were not built as they were not near the peptide binding site. All water molecules were deleted. H-bonds were optimised for pH 7 ionisation states. The structure was refined using the OPLS_2005 force field restraining the final structure to a RMSD of 0.3 Å compared to the initial structure.

Protein Production

E. coli Rosetta strain cells transfected with a plasmid containing a construct shown to have good expression were provided as glycerol stocks for each of DPF2 and PHF10 by Pavel Savitsky (SGC). The constructs used encoded for a His-GST-TEV tagged protein. Overnight cultures were prepared from glycerol stocks in 100 mL growth media (12 g peptone, 24 g yeast extract, 4 mL glycerol, 100 mL of buffer containing 0.17 M KH_2PO_4 and 0.72 M K_2HPO_4 in a total volume of 1 L ddH_2O) containing 50 µg/mL kanamycin and 34 µg/mL chloramphenicol. 10 mL of overnight culture were added to each of 6 x 1 L of growth media. The cultures were grown at 37 °C until a mean OD_{600} of 2 was observed. The cultures were cooled for an hour at 18 °C and induced with IPTG (to a final concentration of 0.2 mM). After 16 h the media was pelleted, and the pellet resuspended in 75 mL lysis buffer (50 mM HEPES, 500 mM NaCl, 5% glycerol, 0.5 mM TCEP) per 1 L of culture. The resuspended cells were mechanically homogenised, and the homogenate pelleted. The resulting supernatant was purified by affinity chromatography at 4 °C using nickel-NTA resin, and the fractions analysed by SDS-PAGE and a NanoDrop spectrophotometer.

Combined fractions of suitable purity were incubated overnight at 4 °C with TEV protease at a molar ratio of 1:20. This step was omitted in the preparation of tagged protein for AlphaScreen. Size exclusion chromatography was used to further purify the protein; a GE Superdex 200 column was used for the tagged protein and a GE Superdex 75 column used for the untagged protein.

BioLayer Interferometry

BioLayer Interferometry experiments were performed using a ForteBio OctetRed384 system equipped with Streptavidin Dip and Read™ Biosensors. Sensors were loaded with the relevant biotinylated peptide and allowed to equilibrate in buffer (25 mM HEPES pH 7.5, 100 mM NaCl) for 60 s prior to use. Tips were submerged in a solution containing DPF2 (20 µM) for 240 s, then submerged in buffer for 300 s to allow for the dissociation of bound protein.

Chemical Biotinylation

Recombinant protein was chemically biotinylated using an ImmunoProbe™ Biotinylation kit (Sigma Aldrich) following the supplied procedure. Untagged DPF2 was diluted to 56 µM in the supplied 0.1 M sodium phosphate buffer (pH 7.2) to a total volume of 1 mL. The supplied BAC-SulfoNHS reagent (5 mg) was dissolved in 30 µL DMSO, 0.1 M sodium phosphate buffer (pH 7.2) was added to give a total volume of 0.5 mL (10 mg/mL, 18 mM). 19 µL of BAC-SulfoNHS solution was added to the DPF2 solution (6:1 molar ratio) and the solution incubated at RT for 30 mins with gentle shaking.

The supplied 3 mL Sephadex G-25M column was equilibrated with 30 mL of 0.01 M PBS and the reaction mixture loaded onto the column. The column was eluted with 9 x 1 mL 0.01 M PBS and the fractions analysed for the presence of protein by measuring absorbance at 280 nm. Fractions containing protein were pooled and concentrated to 4.99 mg/mL. Analysis of the concentrated protein by mass spectrometry showed the singly biotinylated protein to be the

dominant species, with contaminating unmodified protein, as well as doubly and triply biotinylated protein.

Native Mass Spectrometry for Zinc Ejection Assay

Native mass spectrometry was performed on an Agilent 6530 QTOF. The mass spectrometer was set to 1 GHz extended mass range mode which is suitable for detecting ions with an m/z ratio up to 20,000. Once set to this mode the spectrometer was allowed to stabilise for 20 mins before calibration with the supplied calibration mixture.

Protein samples were desalted by passing through three Micro Bio-Spin columns (Bio-Rad) pre-equilibrated with 50 mM NH_4OAc. Protein samples were directly infused into the mass spectrometer, with no chromatography or solid-phase extraction, using a syringe pump set at 1000 µL/h.

Protein Crystallisation

Aliquots of the purified proteins were set up for crystallisation using a mosquito crystallisation robot (TTP Labtech, Royston UK). Coarse screens were typically setup onto Greiner 3-well plates using three different drop ratios of precipitant solution to protein solution per condition (100+50 nL, 75+75 nL and 50+100 nL). The production of multiple DPF3b crystals for solvent screening was performed using larger drops consisting of 150 nL of protein solution and 150 nL of precipitant (0.07 M HEPES pH 7.2; 1.4 M sodium citrate tribasic; 3% glycerol). Crystallisation was carried out using the sitting drop vapour diffusion method at 4 °C. Crystals were harvested at 4 °C and immediately flash frozen with liquid nitrogen.

Collection of X-ray Diffraction Data and Data Processing

Crystals were tested for diffraction using a Bruker Microstar fitted with an Apex II detector.

X-ray diffraction data was collected at the Diamond Light Synchrotron beamline I04-1. Diffraction images were indexed using iMosflm, with further data processing using the CCP4

package. The crystal structure of DPF3b was solved using Phaser[211] by molecular replacement using the crystal structure of DPF3b provided by Jinrong Min (Structural Genomics Consortium, University of Toronto).

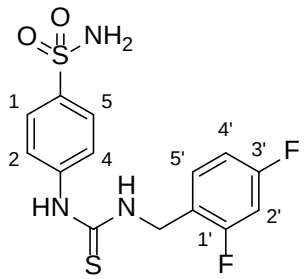
Synthesis and Characterisation

General Procedure for the Synthesis of Thioureas

4-isothiocyanatobenzenesulfonamide (25 mg, 0.13 mmol) was suspended in 2.5 mL of dry acetonitrile. An equimolar amount of the appropriate benzylamine was added and the reaction mixture allowed to stir at room temperature for 2.5 h. The reaction mixture was analysed by LC/MS to confirm that it had gone to completion, and the solvent removed *in vacuo*. The resultant residue was either recrystallised from hot ethanol, purified using cation exchange chromatography, or used without further purification, as stated. In the case of compounds that were recrystallised, the crystals were dissolved in an appropriate solvent in order to transfer them to a storage container. The solvent was then removed *in vacuo* prior to the melting point measurement.

Carbon spectra where the number of peaks observed is less than the number of peaks expected due to overlap are labelled with an asterisk.

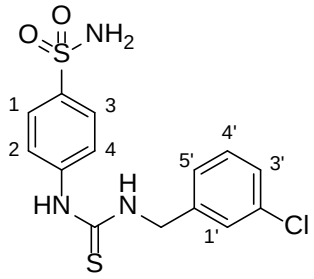
4-(3-(2,4-Difluorobenzyl)thioureido)benzenesulfonamide



Used without further purification. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.75 (d, *J* = 5.4 Hz, 2 H, C**H**$_2$), 7.10 (td, *J* = 8.4, 2.3 Hz, 1 H), 7.21 - 7.26 (m, 1 H), 7.28 (s., 2 H, N**H**$_2$), 7.47 (app. q, *J* = 8.0 Hz, 1 H), 7.68 (app. d, *J* = 8.9 Hz, 2 H), 7.75 (app. d, *J* = 8.7 Hz, 2 H), 8.43 (t, *J* = 5.3 Hz, 1 H, CH$_2$N**H**), 9.96 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d6) δ ppm 41.09 (**C**H$_2$), 104.17 (t, *J* = 26.4 Hz), 111.74 (dd, *J* = 21.3, 3.7 Hz), 121.87 - 122.75 (m), 126.73, 131.33, 139.27, 142.97, 159.25 (d, *J* = 12.5 Hz), 160.78 (d, *J* = 11.7 Hz), 161.76 (d, *J* = 13.2 Hz), 163.15 (d, *J* = 11.7

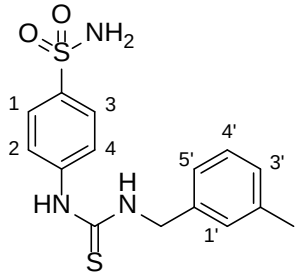
Hz), 181.34 (**C**=S); mp 189 °C; LC/MS *m/z* (ESI$^+$) 358.21 [M + H]$^+$, t_R = 1.42 min; HRMS (ESI$^+$) observed 358.0494, calculated for $C_{14}H_{14}F_2N_3O_2S_2^+$ [M + H]$^+$ 358.0496; ν_{max} cm^{-1} (neat) 3273 (N-H), 1545 (SO_2NH_2), 1323 (C=S), 1155 (SO_2).

4-(3-(4-Chlorobenzyl)thioureido)benzenesulfonamide



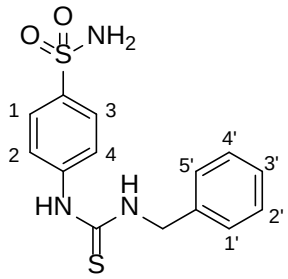
Used without further purification. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.77 (d, *J* = 5.6 Hz, 2 H, C**H**$_2$), 7.28 (s., 2 H, N**H**$_2$), 7.30 - 7.35 (m, 2 H), 7.38 (d, *J* = 7.5 Hz, 1 H), 7.41 (s., 1 H), 7.67 (app. d, *J* = 8.8 Hz, 2 H), 7.72 - 7.78 (m, 2 H), 8.50 (br. s., 1 H, CH_2N**H**), 9.97 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d6) δ ppm 46.95 (**C**H$_2$), 122.47, 126.62, 126.73, 127.32, 127.67, 130.65, 133.39, 139.28, 141.87, 142.98, 181.36 (**C**=S); mp 160 °C; LC/MS *m/z* (ESI$^+$) 356.23 [M + H]$^+$, t_R = 1.43 min; HRMS (ESI$^+$) observed 356.0291, calculated for $C_{14}H_{15}ClN_3O_2S_2^+$ [M + H]$^+$ 356.0294; ν_{max} cm^{-1} (neat) 3351 (N-H), 1542 (SO_2NH_2), 1305 (C=S), 1148 (SO_2).

4-(3-(3-Methylbenzyl)thioureido)benzenesulfonamide



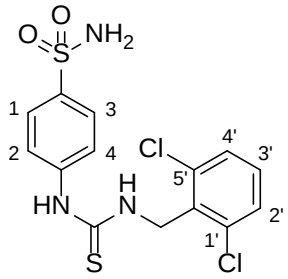
Recrystallised from EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.31 (s., 3 H, C**H**$_3$), 4.71 (d, *J* = 5.4 Hz, 2 H, C**H**$_2$), 7.09 (d, *J* = 7.5 Hz, 1 H), 7.12 - 7.18 (m, 2 H), 7.24 (d, *J* = 7.5 Hz, 1 H), 7.27 (s., 2 H, N**H**$_2$), 7.69 (app. d, *J* = 8.8 Hz, 2 H), 7.74 (m, *J* = 8.9, 2.0 Hz, 2 H), 8.40 (br. s., 1 H, CH_2N**H**), 9.88 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 21.51 (**C**H$_3$), 47.65 (**C**H$_2$), 122.18, 125.15, 126.69, 128.13, 128.62, 128.75, 137.90, 138.89, 143.14, 181.06 (**C**=S)*; mp 191 °C; LC/MS *m/z* (ESI$^+$) 336.29 [M + H]$^+$, t_R = 1.42 min; HRMS (ESI$^+$) observed 336.0842, calculated for $C_{15}H_{18}N_3O_2S_2^+$ [M + H]$^+$ 336.084; ν_{max} cm^{-1} (neat) 3267 (N-H), 1535 (SO_2NH_2), 1315 (C=S), 1157 (SO_2).

4-(3-Benzylthioureido)benzenesulfonamide



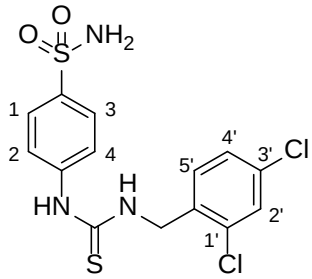
Recrystallised from EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.75 (d, J = 5.4 Hz, 2 H, C**H**$_2$), 7.24 - 7.31 (m, 2 H), 7.33 - 7.38 (m, 4 H), 7.68 (app. d, J = 9.0 Hz, 2 H), 7.74 (app. d, J = 8.8 Hz, 2 H), 8.44 (br. s., 1 H, CH$_2$N**H**), 9.90 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 47.64 (**C**H$_2$), 122.29, 126.70, 127.48, 128.01, 128.80, 139.03, 143.10, 181.16 (**C**=S)*; mp 197 °C; LC/MS *m/z* (ESI$^+$) 322.25 [M + H]$^+$, t_R = 1.37 min; HRMS (ESI$^+$) observed 344.0508, calculated for $C_{14}H_{15}N_3O_2S_2Na^+$ [M + Na]$^+$ 344.0503; ν_{max} cm^{-1} (neat) 3274 (N-H), 1538 (SO_2NH_2), 1316 (C=S), 1157 (SO_2).

4-(3-(2,6-Dichlorobenzyl)thioureido)benzenesulfonamide



Recrystallised from EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.87 (d, J = 3.9 Hz, 2 H, C**H**$_2$), 7.26 (s., 2 H, N**H**$_2$), 7.43 (dd, J = 8.7, 7.5 Hz, 1 H), 7.56 (d, J = 7.8 Hz, 2 H), 7.73 (s., 4 H), 8.21 (br. s., 1 H, CH$_2$N**H**), 9.73 (s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 44.40 (**C**H$_2$), 121.71, 126.66, 129.18, 131.25, 133.06, 136.15, 138.97, 143.21, 180.84 (**C**=S); mp 205 °C; LC/MS *m/z* (ESI$^+$) 390.17 [M + H]$^+$,t_R = 1.46 s; HRMS (ESI$^+$) observed 389.9915, calculated for $C_{14}H_{14}Cl_2N_3O_2S_2^+$ [M + H]$^+$ 389.9905; ν_{max} cm^{-1} (neat) 3307 (N-H), 1534 (SO_2NH_2), 1321 (C=S), 1150 (SO_2).

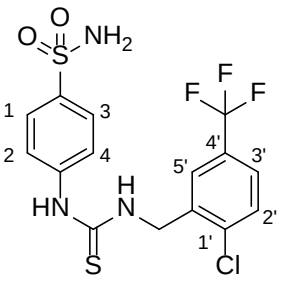
4-(3-(2,4-Dichlorobenzyl)thioureido)benzenesulfonamide



Recrystallised from EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.78 (br. s., 2 H, C**H**$_2$), 7.28 (s., 2 H, N**H**$_2$), 7.40 (d, J = 8.3 Hz, 1 H, C(5')**H**), 7.46 (dd, J = 8.4, 2.0 Hz, 1 H C(4')**H**), 7.64 (d, J = 2.1 Hz, 1 H, C(2')**H**), 7.69 (app. d, J = 8.8 Hz, 2 H), 7.76 (app. d, J = 8.8 Hz, 2 H), 8.46 (br. s., 1 H, CH$_2$N**H**), 10.10 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 45.07 (**C**H$_2$), 122.47, 126.76, 127.73, 129.05, 132.70, 133.33, 135.56, 139.38, 142.91,

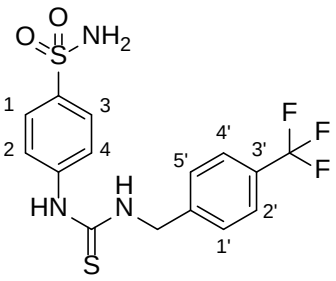
181.56 (**C**=S)*; mp 203 °C; LC/MS *m/z* (ESI^+) 390.00 $[M + H]^+$, t_R = 1.52 min; HRMS (ESI^+) observed 389.9905, calculated for $C_{14}H_{14}Cl_2N_3O_2S_2^+$ $[M + H]^+$ 389.9905; ν_{max} cm^{-1} (neat) 3328 (N-H), 3234 (N-H), 1534 (SO_2NH_2), 1308 (C=S), 1149 (SO_2).

4-(3-(2-Chloro-5-(trifluoromethyl)benzyl)thioureido)benzenesulfonamide



Purified using HPLC/MS. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 4.87 (s., 2 H, C**H**$_2$), 7.28 (br. s., 2 H, N**H**$_2$), 7.56 - 7.84 (m, 7 H), 8.62 (br. s., 1 H, CH_2N**H**), 10.24 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 45.37 (**C**H_2), 122.43, 122.99, 125.70, 125.90, 126.76, 127.98, 128.30, 130.80, 136.83, 138.09, 139.36, 143.00, 181.72 (**C**=S); mp 203 °C; LC/MS *m/z* (ESI^+) 424.08 $[M + H]^+$, t_R = 1.52 min; HRMS (ESI^+) observed 424.0173, calculated for $C_{15}H_{14}ClF_3N_3O_2S_2^+$ $[M + H]^+$ 424.0168; ν_{max} cm^{-1} (neat) 3262 (N-H), 1535 (SO_2NH_2), 1327 (C=S), 1151 (SO_2).

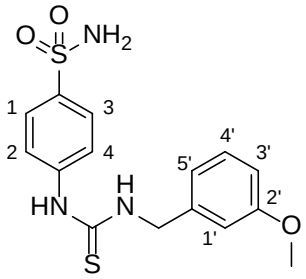
4-(3-(4-(Trifluoromethyl)benzyl)thioureido)benzenesulfonamide



Recrystallised from EtOH. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 4.85 (s., 2 H, C**H**$_2$), 7.29 (s., 2 H, N**H**$_2$), 7.56 (d, *J* = 8.1 Hz, 2 H), 7.67 (d, *J* = 8.8 Hz, 2 H), 7.70 - 7.78 (m, 4 H), 8.56 (br. s., 1 H, CH_2N**H**), 10.02 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 47.12 (**C**H_2), 122.56, 123.48, 125.59, 125.63, 126.76, 127.81, 128.12, 128.46, 139.35, 142.90, 144.27, 181.51 (**C**=S); mp 194 °C; LC/MS *m/z* (ESI^+) 390.19 $[M + H]^+$, t_R = 1.48 min; HRMS (ESI^+) observed 390.0563, calculated for $C_{15}H_{15}F_3N_3O_2S_2^+$ $[M + H]^+$ 390.0558; ν_{max} cm^{-1} (neat) 3299 (N-H), 3161 (N-H), 1526 (SO_2NH_2), 1320 (C=S), 1159 (SO_2).

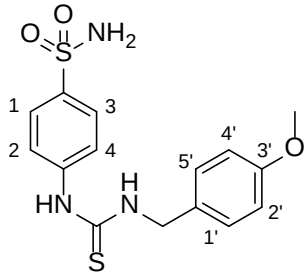
4-(3-(3-Methoxybenzyl)thioureido)benzenesulfonamide

Recrystallised from EtOH. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 3.76 (s., 3 H, C**H**$_3$), 4.72 (d, *J* = 5.4 Hz, 2 H, C**H**$_2$), 6.85 (dd, *J* = 7.8, 2.2 Hz, 1 H), 6.88 - 6.96 (m, 2 H), 7.22 - 7.31 (m, 3 H), 7.68 (app. d, *J* = 9.0 Hz, 2 H), 7.74 (app. d, *J* = 8.8 Hz, 2 H), 8.42 (br. s., 1 H, CH_2N**H**), 9.90 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 47.62 (**C**H_2), 55.49 (O**C**H_3), 112.81, 113.73, 120.16,



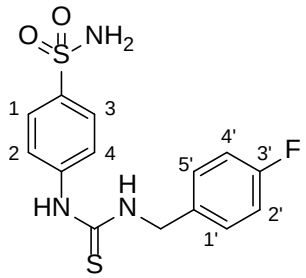
122.26, 126.70, 129.90, 140.59, 143.10, 159.78, 181.16 (**C**=S)*; mp 191 °C; LC/MS *m/z* (ESI^+) 352.26 $[M + H]^+$, t_R = 1.38 min; HRMS (ESI^+) observed 352.0781, calculated for $C_{15}H_{18}N_3O_3S_2^+$ $[M + H]^+$ 352.079; ν_{max} cm^{-1} (neat) 3260 (N-H), 1540 (SO_2NH_2), 1316 (C=S), 1158 (SO_2).

4-(3-(4-Methoxybenzyl)thioureido)benzenesulfonamide



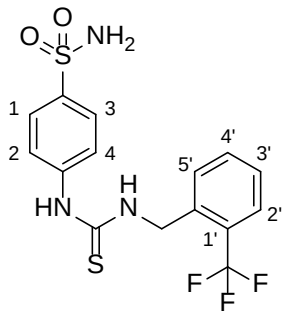
Purified using cation exchange chromatography. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 3.74 (s., 3 H), 4.66 (d, *J* = 5.3 Hz, 2 H, C**H**$_2$), 6.91 (app. d, *J* = 8.8 Hz, 2 H), 7.26 (s., 2 H, N**H**$_2$), 7.29 (app. d, *J* = 8.7 Hz, 2 H), 7.67 (app. d, *J* = 8.8 Hz, 2 H), 7.73 (app. d, *J* = 8.8 Hz, 2 H), 8.35 (br. s., 1 H, CH_2N**H**), 9.84 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 47.17 (**C**H_2), 55.56 (O**C**H_3), 114.22, 122.18, 126.70, 129.50, 130.83, 139.04, 143.13, 158.90, 180.85 (**C**=S)*; mp 201 °C; LC/MS *m/z* (ESI^+) 352.28 $[M + H]^+$, t_R = 1.37 min; HRMS (ESI^+) observed 352.0786, calculated for $C_{15}H_{18}N_3O_3S_2^+$ $[M + H]^+$ 352.079; ν_{max} cm^{-1} (neat) 3243 (N-H), 1512 (SO_2NH_2), 1318 (C=S), 1151 (SO_2).

4-(3-(4-Fluorobenzyl)thioureido)benzenesulfonamide



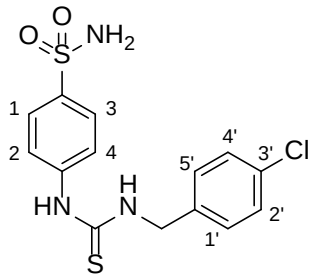
Recrystallised from EtOH. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 4.73 (d, *J* = 4.6 Hz, 2 H, C**H**$_2$), 7.18 (app. tt, *J* = 8.9, 2.2 Hz, 2 H), 7.27 (s., 2 H, N**H**$_2$), 7.37 - 7.44 (m, 2 H), 7.67 (app. d, *J* = 8.8 Hz, 2 H), 7.74 (app. d, *J* = 8.8 Hz, 2 H), 8.45 (br. s., 1 H, CH_2N**H**), 9.91 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 46.84 (**C**H_2), 115.50 (d, **C**F), 122.35, 126.72, 129.99, 130.07, 135.32, 139.19, 143.02, 160.54, 162.95, 181.15 (**C**=S); mp 198 °C; LC/MS *m/z* (ESI^+) 340.15 $[M + H]^+$, t_R = 1.38 min; HRMS (ESI^+) observed 340.0584, calculated for $C_{14}H_{15}FN_3O_2S_2^+$ $[M + H]^+$ 340.059; ν_{max} cm^{-1} (neat) 3272 (N-H), 1511 (SO_2NH_2), 1316 (C=S), 1152 (SO_2).

4-(3-(2-(Trifluoromethyl)benzyl)thioureido)benzenesulfonamide



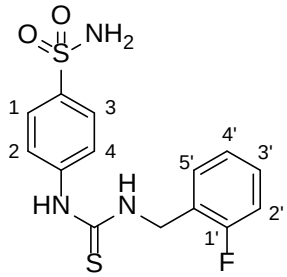
Purified using HPLC/MS. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.96 (s., 2 H, C**H**$_2$), 7.28 (br. s., 2 H, N**H**$_2$), 7.46 - 7.52 (m, 1 H), 7.55 (d, *J* = 7.8 Hz, 1 H), 7.65 - 7.78 (m, 6 H), 8.54 (br. s., 1 H, CH$_2$N**H**), 10.19 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 44.25 (**C**H$_2$), 122.54, 126.29 (m), 126.74, 127.86, 128.98, 133.10, 137.55, 139.34, 142.99, 181.72 (**C**=S)*; mp 197 °C; LC/MS *m/z* (ESI$^+$) 390.16 [M + H] $^+$, t_R = 1.47 min; HRMS (ESI$^+$) observed 390.056, calculated for $C_{15}H_{15}F_3N_3O_2S_2^+$ [M + H]$^+$ 390.0558; ν_{max} cm^{-1} (neat) 3361 (N-H), 1531 (SO_2NH_2), 1307 (C=S), 1146 (SO_2).

4-(3-(4-Chlorobenzyl)thioureido)benzenesulfonamide



Recrystallised from EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.74 (d, *J* = 5.5 Hz, 2 H, C**H**$_2$), 7.28 (s., 2 H, N**H**$_2$), 7.37 (app. d, *J* = 8.6 Hz, 2 H), 7.42 (app. d, *J* = 8.8 Hz, 2 H), 7.66 (app. d, *J* = 9.0 Hz, 2 H), 7.74 (app. d, *J* = 8.7 Hz, 2 H), 8.47 (br. s., 1 H, CH$_2$N**H**), 9.95 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm46.85 (**C**H$_2$), 122.44, 126.73, 128.70, 129.81, 131.92, 138.27, 139.25, 142.97, 181.28 (**C**=S)*; mp 204 °C; LC/MS *m/z* (ESI$^+$) 356.19 [M + H] $^+$, t_R = 1.45 min; HRMS (ESI$^+$) observed 356.0289, calculated for $C_{14}H_{15}ClN_3O_2S_2^+$ [M + H]$^+$ 356.0294; ν_{max} cm^{-1} (neat) 3258 (N-H), 1534 (SO_2NH_2), 1320 (C=S), 1159 (SO_2).

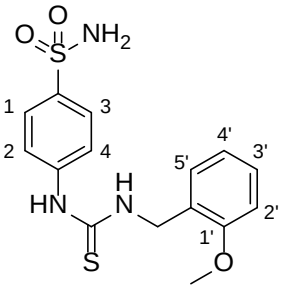
4-(3-(2-Fluorobenzyl)thioureido)benzenesulfonamide



Purified by HPLC/MS. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 4.79 (s., 2 H, C**H**$_2$), 7.17 - 7.24 (m, 2 H), 7.27 (s., 2 H, N**H**$_2$), 7.33 (m, 2 H), 7.42 (td, *J* = 7.7, 1.7 Hz, 1 H), 7.70 (app. d, *J* = 8.8 Hz, 2 H), 7.75 (app. d, *J* = 8.8 Hz, 1 H), 8.46 (br. s., 1 H, CH$_2$N**H**), 9.99 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 41.53 (**C**H$_2$), 115.59 (d,

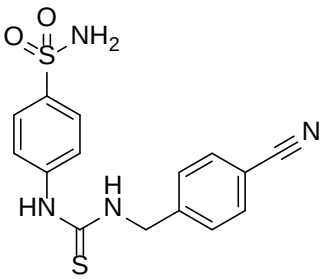
J = 22.0 Hz), 122.29, 124.78 (d, *J* = 3.7 Hz), 125.76 (d, *J* = 14.7 Hz), 126.70, 129.54 (d, *J* = 8.1 Hz), 130.16, 139.20, 143.06, 159.36, 161.79, 181.35 (**C**=S); mp 194 °C; LC/MS *m/z* (ESI^+) 340.23 $[M + H]^+$, t_R = 1.38 min; HRMS (ESI^+) observed 340.0595, calculated for $C_{14}H_{15}ClN_3O_2S_2^+$ $[M + H]^+$ 340.0590; ν_{max} cm^{-1} (neat) 3266 (N-H), 1492 (SO_2NH_2), 1317 (C=S), 1148 (SO_2).

4-(3-(2-Methoxybenzyl)thioureido)benzenesulfonamide



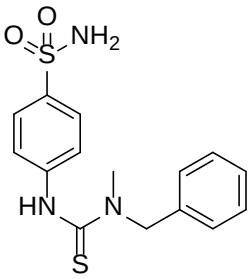
Recrystallised from EtOH. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 3.85 (s., 3 H), 4.69 (d, *J* = 5.0 Hz, 2 H, C**H**$_2$), 6.94 (td, *J* = 7.4, 0.8 Hz, 1 H), 7.03 (d, *J* = 7.7 Hz, 1 H), 7.22 - 7.32 (m, 4 H), 7.73 (m, 4 H), 8.21 (br. s., 1 H, CH_2N**H**), 9.92 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 43.25 (**C**H_2), 55.86 (O**C**H_3), 111.07, 120.62, 121.97, 126.21, 126.68, 128.98, 143.22, 157.35, 181.00 (**C**=S)*; mp 194 °C; LC/MS *m/z* (ESI^+) 352.26 $[M + H]^+$, t_R = 1.39 min; HRMS (ESI^+) observed 352.078, calculated for $C_{15}H_{18}N_3O_3S_2^+$ $[M + H]^+$ 352.0790; ν_{max} cm^{-1} (neat) 3345 (N-H), 1522 ($SO2NH_{2)}$, 1311 (C=S), 1149 (SO_2).

4-(3-(4-Cyanobenzyl)thioureido)benzenesulfonamide



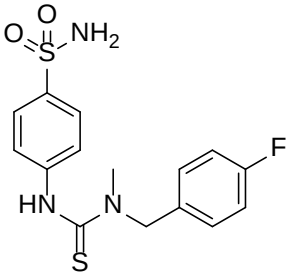
Triturated with EtOH. ^1H NMR (400 MHz, DMSO-d_6) δ ppm 4.78 (s., 2 H, C**H**$_2$), 7.21 (s., 2 H, N**H**$_2$), 7.45 (app. d, *J* = 8.3 Hz, 2 H), 7.58 (app. d, *J* = 8.7 Hz, 2 H), 7.68 (app. d, *J* = 8.7 Hz, 2 H), 7.76 app. (d, *J* = 8.4 Hz, 2 H), 8.48 (br. s., 1 H, CH_2N**H**), 9.98 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 47.22 (**C**H_2), 109.99, 119.39, 122.65, 126.77, 128.59, 132.68, 139.41, 142.85, 145.36, 181.58 (**C**=S); mp 211 °C; LC/MS *m/z* (ESI^+) 347.34 $[M + H]^+$, t_R = 1.34 min; HRMS (ESI^+) observed 347.0643, calculated for $C_{15}H_{15}N_4O_2S_2^+$ $[M + H]^+$ 347.0636; ν_{max} cm^{-1} (neat) 3309 (N-H), 1541 ($SO2NH_2$), 1302 (C=S), 1157 (SO_2).

4-(3-Benzyl-3-methylthioureido)benzenesulfonamide



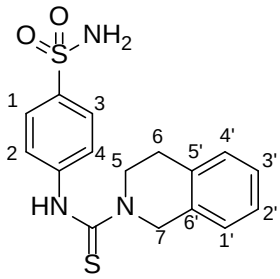
Triturated with EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 3.21 (s., 3 H), 5.18 (s., 2 H, C**H**$_2$), 7.23 - 7.43 (m, 7 H), 7.55 (app. d, *J* = 9.0 Hz, 2 H), 7.75 (app. d, *J* = 8.9 Hz, 2 H), 9.42 (s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 56.46, 125.41, 126.06, 127.72, 127.74, 129.01, 137.55, 139.76, 144.67, 182.23 (**C**=S)*; mp 219 °C; LC/MS *m/z* (ESI$^+$) 336.30 [M + H]$^+$, t_R = 1.39 min; HRMS (ESI$^+$) observed 336.0837, calculated for $C_{15}H_{18}N_3O_2S_2^+$ [M + H]$^+$ 336.0840; ν_{max} cm^{-1} (neat) 3245 (N-H), 1527 (SO_2NH_2), 1298 (C=S), 1145 (SO_2).

4-(3-(4-Fluorobenzyl)-3-methylthioureido)benzenesulfonamide



Triturated with EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 3.18 (s., 3 H, C**H**$_3$), 5.15 (s., 2 H, C**H**$_2$), 7.21 (t, *J* = 8.9 Hz, 2 H), 7.29 (s., 2 H, N**H**$_2$), 7.39 (dd, *J* = 8.7, 5.6 Hz, 2 H), 7.54 (d, *J* = 8.7 Hz, 2 H), 7.74 (d, *J* = 8.8 Hz, 2 H), 9.41 (s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d6) δ ppm 26.81 (N**C**H$_3$), 55.74 (**C**H$_2$), 115.67, 115.88, 125.49, 126.06, 129.83, 129.91, 133.74, 139.82, 144.62, 160.71, 182.18 (**C**=S)*; mp 200 °C; LC/MS *m/z* (ESI$^+$) 354.24 [M + H]$^+$, t_R = 1.42 min; HRMS (ESI$^+$) observed 354.0741, calculated for $C_{15}H_{17}FN_3O_2S_2^+$ [M + H]$^+$ 354.0746; ν_{max} cm^{-1} (neat) 3268 (N-H), 1528 (SO_2NH_2), 1297 (C=S), 1146 (SO_2).

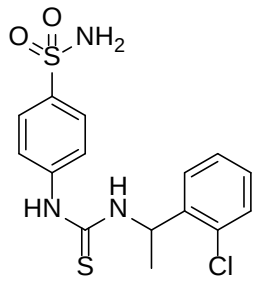
N-(4-Sulfamoylphenyl)-3,4-dihydroisoquinoline-2(1H)-carbothioamide



Triturated with EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.89 (t, *J* = 5.9 Hz, 2 H, C(6)**H**$_2$), 4.01 (t, *J* = 5.9 Hz, 2 H, C(5)**H**$_2$), 4.98 (s., 2 H, C(7)**H**$_2$), 7.11 - 7.19 (m, 4 H), 7.20 (s., 2 H, N**H**$_2$), 7.45 (app. d, *J* = 8.8 Hz, 2 H), 7.66 (app. d, *J* = 8.9 Hz, 2 H), 9.48 (s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 28.62 (**C**(6)H$_2$), 46.73 (**C**(5)H$_2$), 50.70 (**C**(7)H$_2$), 124.79, 126.15, 126.74, 126.79, 127.26, 128.59, 133.77, 135.47, 139.46, 144.63, 181.22 (**C**=S); mp 233 °C; LC/MS *m/z* (ESI$^+$) 384.29 [M + H]$^+$, t_R = 1.41 min; HRMS (ESI$^+$) observed 348.0842, calculated for

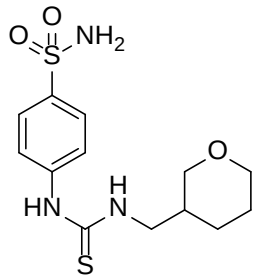
$C_{16}H_{18}N_3O_2S_2^+$ [M + H]$^+$ 348.084; ν_{max} cm^{-1} (neat) 3334 (N-H), 3254 (N-H), 1528 (SO_2NH_2), 1298 (C=S), 1159 (SO_2).

4-(3-(1-(2-Chlorophenyl)ethyl)thioureido)benzenesulfonamide



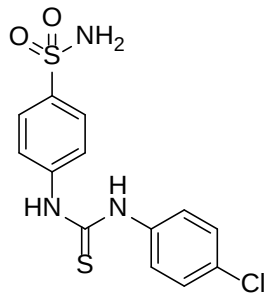
Triturated with EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 1.38 (d, *J* = 7.0 Hz, 3 H), 5.66 (quin, *J* = 7.1 Hz, 1 H, C**H**CH$_3$), 7.19 (s., 2 H, N**H**$_2$), 7.22 (dd, *J* = 7.6, 1.7 Hz, 1 H), 7.29 (td, *J* = 7.5, 1.3 Hz, 1 H), 7.37 (td, *J* = 8.2, 1.6 Hz, 2 H), 7.60 - 7.69 (m, 4 H), 8.48 (d, *J* = 7.5 Hz, 1 H, CHCH$_3$N**H**), 9.75 (s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 21.25 (**C**H$_3$), 50.96 (**C**H$_2$), 121.81, 126.68, 127.46, 127.94, 128.91, 129.83, 132.09, 138.92, 141.97, 143.23, 180.30 (**C**=S); mp 191 °C; LC/MS *m/z* (ESI$^+$) 370.18 [M + H]$^+$, t_R = 1.45 min; HRMS (ESI$^+$) observed 370.9455, calculated for $C_{15}H_{17}ClN_3O_2S_2^+$ [M + H]$^+$ 370.0451; ν_{max} cm^{-1} (neat) 3311 (N-H), 1541 (SO_2NH_2), 1322 (C=S), 1144 (SO_2).

4-(3-((Tetrahydro-2H-pyran-3-yl)methyl)thioureido)benzenesulfonamide



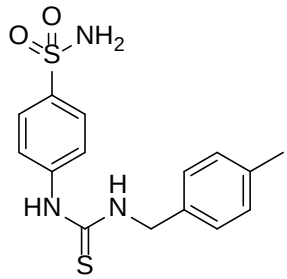
Triturated with EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 1.19 - 1.33 (m, 1 H), 1.40 - 1.54 (m, 1 H), 1.55 - 1.66 (m, 1 H), 1.73 - 1.83 (m, 1 H), 1.83 - 1.96 (m, 1 H), 3.08 - 3.20 (m, 1 H), 3.28 - 3.47 (m, 3 H), 3.65 - 3.83 (m, 2 H, C**H**$_2$), 7.26 (s., 2 H, N**H**$_2$), 7.65 (app. d, *J* = 8.8 Hz, 2 H), 7.73 (app. d, *J* = 8.8 Hz, 2 H), 8.07 (br. s., 1 H, CH$_2$N**H**), 9.74 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 25.24, 27.43, 35.80, 46.38, 67.95, 70.95, 122.00, 126.71, 138.92, 143.17, 181.06 (**C**=S); mp 192 °C; LC/MS *m/z* (ESI$^+$) 330.26 [M + H]$^+$, t_R = 1.18 min; HRMS (ESI$^+$) observed 330.0937, calculated for $C_{13}H_{20}N_3O_3S_2^+$ [M + H]$^+$ 330.0946; ν_{max} cm^{-1} (neat) 3240 (N-H), 1533 (SO_2NH_2), 1321 (C=S), 1153 (SO_2).

4-(3-(4-Chlorophenyl)thioureido)benzenesulfonamide



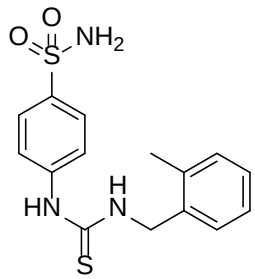
Triturated with EtOH. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 7.30 (s., 2 H, N**H**$_2$), 7.41 (app. d, *J* = 8.7 Hz, 2 H), 7.53 (app. d, *J* = 8.8 Hz, 2 H), 7.68 (app. d, *J* = 8.7 Hz, 2 H), 7.77 (app. d, *J* = 8.9 Hz, 2 H), 10.10 (s., 1 H), 10.14 (s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 123.19, 125.82, 126.66, 128.88, 129.05, 138.68, 139.68, 142.96, 180.14 (**C=S**); mp 191 °C; LC/MS *m/z* (ESI$^+$) 342.17 [M + H]$^+$, t_R = 1.40 s; HRMS (ESI$^+$) observed 342.0133, calculated for $C_{13}H_{13}ClN_3O_2S_2^+$ [M + H]$^+$ 342.0138; ν_{max} cm^{-1} (neat) 3339 (N-H), 3198 (N-H), 1530 (SO_2NH_2), 1338 (C=S), 1157 (SO_2). Data consistent with literature values.[212]

4-(3-(4-Methylbenzyl)thioureido)benzenesulfonamide



Purified using HPLC/MS. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.22 (s., 3 H), 4.62 (d, *J* = 4.4 Hz, 2 H, C**H**$_2$), 7.09 (d, *J* = 7.8 Hz, 2 H), 7.16 (s., 2 H, N**H**$_2$), 7.19 (br. s., 2 H), 7.60 (app. d, *J* = 8.8 Hz, 2 H), 7.66 (app. d, *J* = 8.7 Hz, 2 H), 8.31 (br. s., 1 H, CH_2N**H**), 9.79 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 21.17 (**C**H$_3$), 47.43 (**C**H$_2$), 122.18, 126.70, 128.04, 129.34, 135.91, 136.59, 139.07, 143.12, 181.02 (**C**=S); mp 216 °C; LC/MS *m/z* (ESI$^+$) 336.21 [M + H]$^+$, t_R = 1.42 min; HRMS (ESI$^+$) observed 336.0831, calculated for $C_{15}H_{18}N_3O_2S_2^+$ [M + H]$^+$ 336.084; ν_{max} cm^{-1} (neat) 3333 (N-H), 3255 (N-H), 1494 (SO_2NH_2), 1317 (C=S), 1152 (SO_2).

4-(3-(2-Methylbenzyl)thioureido)benzenesulfonamide



Purified using HPLC/MS. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.32 (s., 3 H), 4.70 (d, *J* = 5.0 Hz, 2 H, C**H**$_2$), 7.20 (m, 2 H), 7.27 (m, 2 H), 7.73 (s., 4 H), 8.29 (br. s., 1 H, CH_2N**H**), 9.85 (br. s., 1 H, ArN**H**); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 19.18 (**C**H$_3$), 45.88 (**C**H$_2$), 126.30, 126.68, 127.64, 130.50, 181.00 (**C**=S)*; mp 176 °C; LC/MS *m/z* (ESI$^+$) 336.23 [M + H]$^+$,

t_R = 1.40 min; HRMS (ESI^+) observed 336.0830, calculated for $C_{15}H_{18}N_3O_2S_2^+$ $[M + H]^+$ 336.084; ν_{max} cm^{-1} (neat) 3333 (N-H), 3254 (N-H), 1493 (SO_2NH_2), 1329 (C=S), 1159 (SO_2).

Experimental Details for Chapter 5 - Investigating the Ligandability of the PHD-JmjC of PHF8

Virtual Screening

Protein Structure Preparation

The structure of PHF8 containing the PHD and JmjC domains (PDB ID 3KV4) was prepared using Schrödinger Protein Preparation Wizard (Epik version 2.4; Impact version 5.9; Prime version 3.2, Schrödinger, LLC, New York, NY, 2013). Missing side chains were not built as they were not near the peptide binding site. All ethylene glycol molecules were deleted. Four water molecules were retained, as they appeared to form a well-defined network in the peptide binding cavity. The retained waters are residues 487, 491, 659, and 660 in the original PDB file. H-bonds were optimised for pH 7 ionisation states. The structure was refined using the OPLS_2005 force field restraining the final structure to a RMSD of 0.3 Å compared to the initial structure. The *N*-oxalylglycine and residues 7-14 of the bound peptide were deleted after structure minimisation.

Synthesis and Characterisation

Synthesis of Diethyl 2-Benzyl-2-(1,3-Dioxoisoindolin-2-yl) Malonates

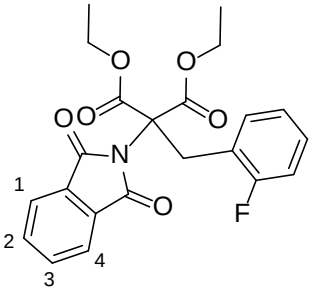
General Procedure

Diethyl 2-(1,3-dioxoisoindolin-2-yl)malonate (14 g, 46 mmol) was dissolved in anhydrous DMF (20 mL) to give a 2.3 M solution. Potassium *tert*-butoxide (5.12 g, 46 mmol) was dissolved in anhydrous DMF (20 mL) to give a 2.3 M solution. 1 mL of each of 2.3 M diethyl 2-(1,3-dioxoisoindolin-2-yl)malonate solution and 2.3 M potassium *tert*-butoxide were added to a sealed vial and the mixture allowed to stir at room temperature for 1 h. The appropriate benzyl-bromide (1 equiv.) was added either as a solution in 0.5 mL of DMF for solid benzyl bromides or as a neat liquid for liquid benzyl bromides. The mixture was allowed to stir at 60 °C for 7 h. The solvent was removed *in vacuo*. The crude material was dissolved in 2 mL of DCM and

2 mL of H_2O. The biphasic mixture was extracted with 3 x 2 mL DCM and the combined organic layers washed with 3 x 5 mL of H_2O, 5 mL of brine, and dried *in vacuo.*

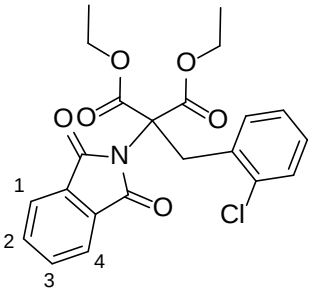
Carbon spectra where the number of peaks observed is less than the number of peaks expected due to overlap are labelled with an asterisk.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(2-fluorobenzyl)malonate



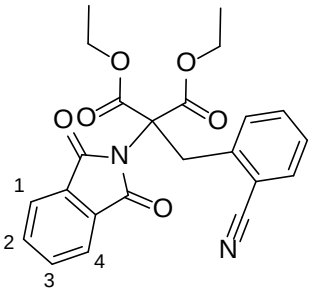
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.29 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 3.87 (s., 2 H, C**H**$_2$Ar), 4.20 - 4.44 (m, 4 H, C**H**$_2$CH$_3$), 6.70 (t, *J* = 9.3 Hz, 1 H), 7.03 (td, *J* = 7.5, 1.1 Hz, 1 H), 7.09 - 7.17 (m, 1 H), 7.61 (td, *J* = 7.7, 1.7 Hz, 1 H), 7.69 - 7.82 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.84, 31.03 (d, *J* = 1.5 Hz), 61.96 - 64.11 (m), 67.86, 114.58 (d, *J* = 22.7 Hz), 122.06 (d, *J* = 14.7 Hz), 123.31, 123.84 (d, *J* = 2.9 Hz), 129.01 (d, *J* = 8.1 Hz), 131.58, 133.98 (d, *J* = 4.4 Hz), 134.11, 161.77 (d, *J* = 248.0 Hz), 165.60, 166.69; LC/MS *m/z* (ESI$^+$) 414.32 $[M + H]^+$, t_R = 1.79 min; HRMS (ESI$^+$) observed 436.1191, calculated for $C_{22}H_{20}FNO_6Na^+$ $[M + Na]^+$ 436.1172.

Diethyl 2-(2-chlorobenzyl)-2-(1,3-dioxoisoindolin-2-yl)malonate



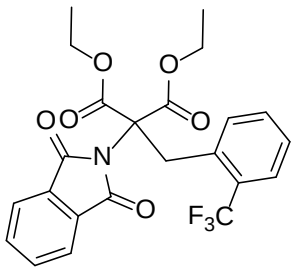
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 1.15 (t, *J* = 7.1 Hz, 6 H, C**H**$_3$), 3.87 (s., 2 H, C**H**$_2$Ar), 4.12 - 4.32 (m, 4 H, C**H**$_2$CH$_3$), 7.17 - 7.25 (m, 3 H), 7.55 (d, *J* = 7.3 Hz, 1 H), 7.84 - 7.91 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 14.07, 34.83, 63.21, 67.85, 124.02, 127.45, 129.40, 129.51, 130.93, 133.00, 133.56, 134.62, 135.78, 165.54, 166.77; LC/MS *m/z* (ESI$^+$) 430.29 $[M + H]^+$, t_R = 1.93 min; HRMS (ESI$^+$) observed 452.0899, calculated for $C_{22}H_{20}ClNO_6Na^+$ $[M + Na]^+$ 452.0876.

Diethyl 2-(2-cyanobenzyl)-2-(1,3-dioxoisoindolin-2-yl)malonate



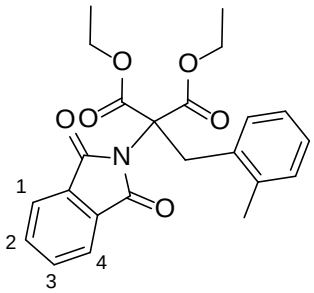
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.29 (t, *J* = 7.1 Hz, 6 H, C**H**$_3$), 4.09 (s., 2 H, C**H**$_2$Ar), 4.26 - 4.42 (m, 4 H, C**H**$_2$CH$_3$), 7.25 - 7.31 (m, 1 H), 7.42 - 7.48 (m, 2 H), 7.68 (d, *J* = 7.8 Hz, 1 H), 7.73 - 7.83 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.81, 36.68, 63.21, 67.92, 114.53, 117.89, 123.58, 127.61, 131.44, 132.15, 132.30, 132.70, 134.39, 139.07, 165.40, 166.81; LC/MS *m/z* (ESI$^+$) 421.35 $[M + H]^+$, t_R = 1.68 min; HRMS (ESI$^+$) observed 443.124, calculated for $C_{23}H_{20}N_2O_6Na^+$ $[M + Na]^+$ 443.1219.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(2-(trifluoromethyl)benzyl)malonate



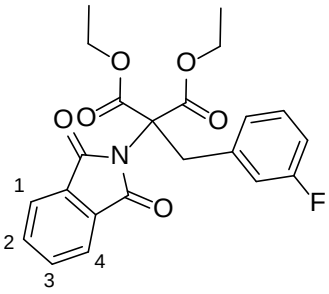
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.23 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 4.21 (s., 2 H, C**H**$_2$Ar), 4.24 - 4.36 (m, 4 H, C**H**$_2$CH$_3$), 7.30 (t, *J* = 7.8 Hz, 1 H), 7.41 (t, *J* = 7.8 Hz, 1 H), 7.56 (d, *J* = 7.8 Hz, 1 H), 7.64 (d, *J* = 7.9 Hz, 1 H), 7.72 - 7.75 (m, 2 H), 7.78 - 7.83 (m, 2 H); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.69, 33.88, 62.94, 68.89, 123.48, 125.53 - 126.18 (m), 126.89, 131.29, 131.46, 132.06, 134.30, 134.47 (d, *J* = 5.1 Hz), 166.05, 167.24; LC/MS *m/z* (ESI$^+$) 464.34 $[M + H]^+$, t_R = 1.91 min; HRMS (ESI$^+$) observed 486.1161, calculated for $C_{23}H_{20}F_3NO_6Na^+$ $[M + Na]^+$ 486.1140.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(2-methylbenzyl)malonate



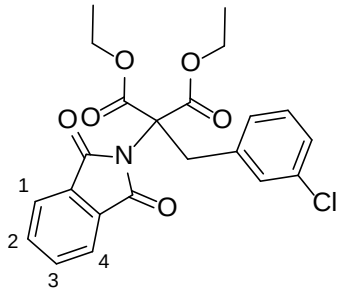
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.27 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 2.26 (s., 3 H, ArC**H**$_3$), 3.97 (s., 2 H, C**H**$_2$Ar), 4.17 - 4.39 (m, 4 H, C**H**$_2$CH$_3$), 6.91 - 6.96 (m, 1 H), 6.96 - 7.03 (m, 2 H), 7.30 (d, *J* = 7.2 Hz, 1 H), 7.68 - 7.79 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.81, 19.86, 34.29, 62.90, 68.62, 123.33, 125.42, 126.88, 130.17, 130.91, 131.41, 133.65, 134.18, 137.78, 166.25, 166.92; LC/MS *m/z* (ESI$^+$) 410.37 $[M + H]^+$, t_R = 1.87 min; HRMS (ESI$^+$) observed 432.1438, calculated for $C_{23}H_{23}NO_6Na^+$ $[M + Na]^+$ 432.1423.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(3-fluorobenzyl)malonate



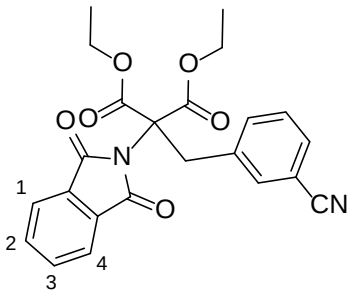
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.30 (t, *J* = 7.3 Hz, 6 H, C**H**$_3$), 3.79 (s., 2 H, C**H**$_2$Ar), 4.26 - 4.47 (m, 4 H, C**H**$_2$CH$_3$), 6.74 - 6.83 (m, 1 H), 6.96 - 7.06 (m, 3 H), 7.67 - 7.81 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.86, 37.49 (d, *J* = 1.5 Hz), 62.99, 68.24, 113.99 (d, *J* = 21.3 Hz), 117.97 (d, *J* = 21.3 Hz), 123.39, 126.55 (d, *J* = 2.9 Hz), 129.19 (d, *J* = 8.1 Hz), 131.27, 134.27, 137.41 (d, *J* = 7.3 Hz), 162.33 (d, *J* = 245.8 Hz) 165.72, 166.69; LC/MS *m/z* (ESI$^+$) 414.33 $[M + H]^+$, t_R = 1.83 min; HRMS (ESI$^+$) observed 436.1189, calculated for $C_{22}H_{20}FNO_6Na^+$ $[M + Na]^+$ 436.1172.

Diethyl 2-(3-chlorobenzyl)-2-(1,3-dioxoisoindolin-2-yl)malonate



^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.31 (t, *J* = 7.3 Hz, 6 H, C**H**$_3$), 3.77 (s., 2 H, C**H**$_2$Ar), 4.23 - 4.45 (m, 4 H, C**H**$_2$CH$_3$), 7.02 (d, *J* = 7.6 Hz, 1 H), 7.07 (dt, *J* = 8.1, 1.3 Hz, 1 H), 7.17 (dt, *J* = 7.6, 1.6 Hz, 1 H), 7.20 (t, *J* = 1.7 Hz, 1 H), 7.70 - 7.79 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.87, 37.44, 63.02, 68.24, 123.42, 127.20, 129.09, 129.17, 131.03, 131.27, 133.64, 134.30, 136.99, 165.68, 166.73; LC/MS *m/z* (ESI$^+$) 430.31 $[M + H]^+$, t_R = 1.92 min; HRMS (ESI$^+$) observed 452.0896, calculated for $C_{22}H_{20}ClNO_6Na^+$ $[M + Na]^+$ 452.0876.

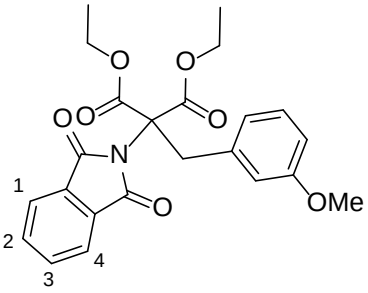
Diethyl 2-(3-cyanobenzyl)-2-(1,3-dioxoisoindolin-2-yl)malonate



^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.21 (t, *J* = 7.1 Hz, 6 H, C**H**$_3$), 3.73 (s., 2 H, C**H**$_2$Ar), 4.17 - 4.34 (m, 4 H, C**H**$_2$CH$_3$), 7.08 - 7.13 (m, 1 H), 7.31 (dt, *J* = 7.8, 1.3 Hz, 1 H), 7.46 (s., 2 H), 7.62 - 7.69 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.86, 37.44, 63.17, 67.96, 111.98, 118.52, 123.54, 128.70, 130.87, 131.06,

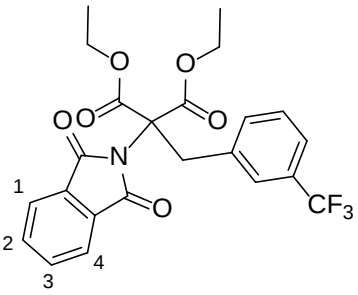
134.54, 135.53, 136.63, 165.53, 166.67; LC/MS *m/z* (ESI^+) 421.33 $[M + H]^+$, t_R = 1.74 min; HRMS (ESI^+) observed 443.1239, calculated for $C_{23}H_{20}N_2O_6Na^+$ $[M + Na]^+$ 443.1219.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(3-methoxybenzyl)malonate



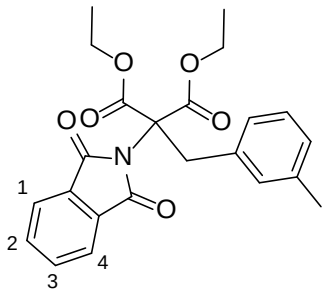
^1H NMR (400 MHz, $CDCl_3$) δ ppm 1.31 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 3.54 (s., 3 H, OC**H**$_3$), 3.77 (s., 2 H, C**H**$_2$Ar), 4.25 - 4.46 (m, 4 H, C**H**$_2$CH$_3$), 6.63 (dd, *J* = 7.8, 2.0 Hz, 1 H), 6.79 - 6.84 (m, 2 H), 6.95 (d, *J* = 7.8 Hz, 1 H), 7.68 - 7.80 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.89, 37.71, 54.87, 62.90, 68.49, 113.47, 115.74, 123.29, 123.39, 128.78, 131.45, 134.13, 136.37, 159.07, 165.87, 166.67; LC/MS *m/z* (ESI^+) 426.34 $[M + H]^+$, t_R = 1.80 min; HRMS (ESI^+) observed 448.1387, calculated for $C_{23}H_{23}NO_7Na^+$ $[M + Na]^+$ 448.1372.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(3-(trifluoromethyl)benzyl)malonate



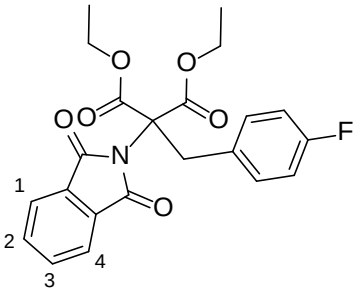
^1H NMR (400 MHz, $CDCl_3$) δ ppm 1.31 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 3.84 (s., 2 H, C**H**$_2$Ar), 4.25 - 4.44 (m, 4 H, C**H**$_2$CH$_3$), 7.26 (t, *J* = 7.8 Hz, 1 H), 7.31 (s., 1 H), 7.35 (d, *J* = 7.7 Hz, 1 H), 7.62 (d, *J* = 7.7 Hz, 1 H), 7.72 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.85, 37.43, 63.07, 68.16, 123.40, 123.84 (q, *J* = 3.7 Hz), 127.39 (q, *J* = 3.7 Hz), 128.40, 131.17, 134.34, 134.73 (app. d, *J* = 1.5 Hz), 136.02, 165.62, 166.71; LC/MS *m/z* (ESI^+) 464.34 $[M + H]^+$, t_R = 1.92 min; HRMS (ESI^+) observed 486.1161, calculated for $C_{23}H_{20}F_3NO_6Na^+$ $[M + Na]^+$ 486.1140.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(3-methylbenzyl)malonate



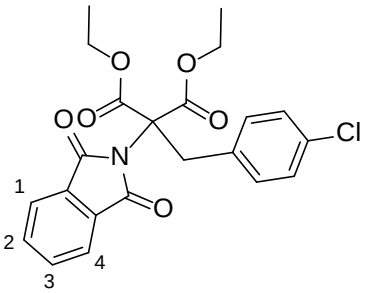
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.31 (t, *J* = 6.6 Hz, 6 H, C**H**$_3$), 2.00 (s., 3 H, ArC**H**$_3$), 3.75 (s., 2 H, C**H**$_2$Ar), 4.21 - 4.43 (m, 4 H, C**H**$_2$CH$_3$), 6.88 (d, *J* = 7.5 Hz, 1 H), 6.93 (s., 1 H), 6.97 (t, *J* = 7.5 Hz, 1 H), 7.09 (d, *J* = 7.6 Hz, 1 H), 7.68 - 7.80 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.89, 20.94, 37.54, 62.84, 68.62, 123.24, 127.64, 127.78, 128.07, 131.49, 131.64, 134.08, 134.81, 137.28, 165.87, 166.71; LC/MS *m/z* (ESI$^+$) 410.37 $[M + H]^+$, t_R = 1.87 min; HRMS (ESI$^+$) observed 432.1438, calculated for $C_{23}H_{23}NO_6Na^+$ $[M + Na]^+$ 432.1423.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(4-fluorobenzyl)malonate



^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.30 (t, *J* = 7.0 Hz, 6 H, C**H**$_3$), 3.76 (s., 2 H, C**H**$_2$Ar), 4.25 - 4.43 (m, 4 H, C**H**$_2$CH$_3$), 6.74 (t, *J* = 8.7 Hz, 2 H), 7.22 (dd, *J* = 8.8, 5.6 Hz, 2 H), 7.68 - 7.78 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm; 13.87, 36.96, 62.94, 68.32, 114.71 (d, *J* = 21.3 Hz), 123.38, 130.65 (d, *J* = 2.9 Hz), 131.25, 132.48 (d, *J* = 8.1 Hz), 134.26, 161.97 (d, *J* = 244.3 Hz), 165.82, 166.70; LC/MS *m/z* (ESI$^+$) 414.34 $[M + H]^+$, t_R = 1.83 min; HRMS (ESI$^+$) observed 436.1186, calculated for $C_{22}H_{20}FNO_6Na^+$ $[M + Na]^+$ 436.1172.

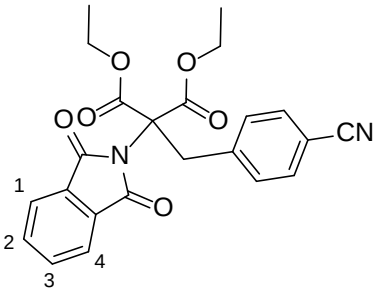
Diethyl 2-(4-chlorobenzyl)-2-(1,3-dioxoisoindolin-2-yl)malonate



^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.30 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 3.76 (s., 2 H, C**H**$_2$Ar), 4.22 - 4.43 (m, 4 H, C**H**$_2$CH$_3$), 7.04 (dt, *J* = 8.6, 2.7 Hz, 2 H), 7.20 (dt, *J* = 8.6, 2.4 Hz, 2 H), 7.70 - 7.77 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.87, 37.15, 62.97, 68.19, 123.44, 128.04, 131.23, 132.27, 133.02, 133.46,

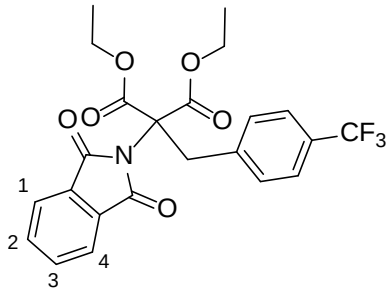
134.31, 165.77, 166.70; LC/MS *m/z* (ESI^+) 430.30 $[M + H]^+$, t_R = 1.92 min; HRMS (ESI^+) observed 452.0896, calculated for $C_{22}H_{20}ClNO_6Na^+$ $[M + Na]^+$ 452.0876.

Diethyl 2-(4-cyanobenzyl)-2-(1,3-dioxoisoindolin-2-yl)malonate



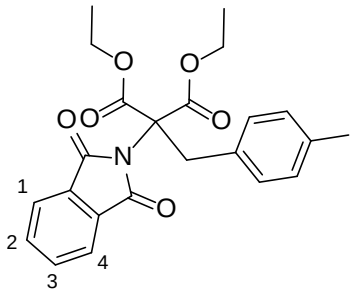
^1H NMR (400 MHz, $CDCl_3$) δ ppm 1.30 (t, *J* = 7.2 Hz, 6 H, C**H**$_3$), 3.84 (s., 2 H, C**H**$_2$Ar), 4.23 - 4.43 (m, 4 H, C**H**$_2$CH$_3$), 7.38 (d, *J* = 4.9 Hz, 4 H), 7.75 (s., 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.86, 37.91, 63.18, 67.89, 111.01, 118.78, 123.53, 127.30, 131.71, 132.64, 134.52, 140.70, 165.54, 166.62; LC/MS *m/z* (ESI^+) 421.38 $[M + H]^+$, t_R = 1.74 min; HRMS (ESI^+) observed 443.1237, calculated for $C_{23}H_{20}N_2O_6Na$ $[M + Na]^+$ 443.1219.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(4-(trifluoromethyl)benzyl)malonate



^1H NMR (400 MHz, $CDCl_3$) δ ppm 1.29 (t, *J* = 7.0 Hz, 6 H, C**H**$_3$), 3.84 (s., 2 H, C**H**$_2$Ar), 4.22 - 4.42 (m, 4 H, C**H**$_2$CH$_3$), 7.34 (dd, *J* = 27.6, 8.4 Hz, 4 H), 7.66 - 7.81 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.85, 37.61, 63.06, 68.09, 123.42, 124.73 (q, *J* = 3.7 Hz), 126.87, 129.27 (q, *J* = 33.0 Hz), 131.17, 131.29, 134.38, 139.15, 165.69, 166.70; LC/MS *m/z* (ESI^+) 464.37 $[M + H]^+$, t_R = 1.94 min; HRMS (ESI^+) observed 486.116, calculated for $C_{23}H_{20}F_3NO_6Na^+$ $[M + Na]^+$ 486.1140.

Diethyl 2-(1,3-dioxoisoindolin-2-yl)-2-(4-methylbenzyl)malonate



^1H NMR (400 MHz, $CDCl_3$) δ ppm 1.29 (t, *J* = 7.3 Hz, 6 H, C**H**$_3$), 2.17 (s., 3 H, ArC**H**$_3$), 3.74 (s., 2 H, C**H**$_2$Ar), 4.23 - 4.43 (m, 4 H, C**H**$_2$CH$_3$), 6.85 (d, *J* = 7.9 Hz, 2 H), 7.10 (d, *J* = 8.1 Hz, 2 H), 7.65 - 7.76 (m, 4 H, C(1-4)**H**); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 13.89,

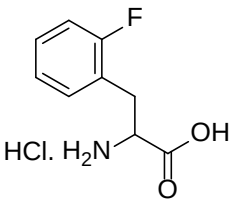
20.97, 37.29, 62.82, 68.56, 123.30, 128.57, 130.77, 131.44, 131.75, 134.08, 136.53, 165.92, 166.75; LC/MS *m/z* (ESI^+) 410.37 $[M + H]^+$, t_R = 1.88 min; HRMS (ESI^+) observed 432.1441$^+$, calculated for $C_{23}H_{23}NO_6Na$ $[M + Na]^+$ 432.1423.

Hydrolysis of Diethyl 2-(1,3-Dioxoisoindolin-2-yl)-2-Benzyl Malonate

The crude residue of the proceeding synthesis of diethyl 2-benzyl-2-(1,3-dioxoisoindolin-2-yl)malonate was dissolved in 7.5 mL of EtOH. The material was transferred to a 20 mL vial equipped with a stirrer bar and 7.5 mL of 5 M HCl was added. The vials were sealed and the mixture stirred at 80 °C for 48 h. At this point the vial caps were loosened and the mixtures heated at 80 °C for a further 24 h. *Ortho*-phthalic acid was precipitated by the addition of cold water, and extracted with 3 x 5 mL of EtOAc. Samples of the aqueous layer were neutralised with 1 M NaOH and submitted for LC/MS and HRMS analysis. The remaining aqueous layer was evaporated *in vacuo* to give the desired product. In some cases the product was isolated as a mixture of free acid and ethyl ester. These cases were carried through to the next step without purification. No NMR data is given for these examples.

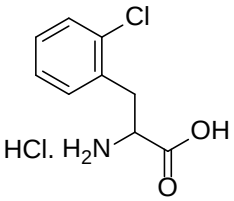
Carbon spectra where the number of peaks observed is less than the number of peaks expected due to overlap are labelled with an asterisk.

2-Amino-3-(2-fluorophenyl)propanoic acid. HCl



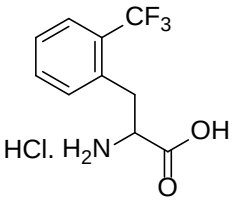
^{1}H NMR (400 MHz, D_2O) δ ppm 3.13 (dd, *J* = 14.8, 7.3 Hz, 1 H), 3.28 (dd, *J* = 14.9, 6.0 Hz, 1 H), 4.23 (dd, *J* = 7.3, 6.0 Hz, 1 H), 6.98 - 7.12 (m, 2 H), 7.14 - 7.33 (m, 2 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 29.40, 53.08, 115.63 (d, *J* = 22.0 Hz), 120.75 (d, *J* = 15.4 Hz), 124.88 (d, *J* = 2.9 Hz), 130.24 (d, *J* = 8.1 Hz), 131.79 (d, *J* = 3.7 Hz), 161.19 (d, *J* = 244.3 Hz), 171.06; LC/MS *m/z* (ESI^+) 184.16 $[M + H]^+$, t_R = 0.28 min; HRMS (ESI^+) observed 228.0418, calculated for $C_9H_9FNO_2Na_2^+$ $[M + 2Na - H]^+$ 228.0408. ^{1}NMR data consistent with literature values.[213]

2-Amino-3-(2-chlorophenyl)propanoic acid. HCl



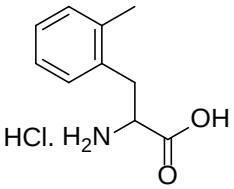
^{1}H NMR (400 MHz, D_2O) δ ppm 3.19 (dd, *J* = 14.5, 8.6 Hz, 1 H), 3.45 (dd, *J* = 14.5, 6.4 Hz, 1 H), 4.21 (dd, *J* = 8.7, 6.5 Hz, 1 H), 7.24 - 7.34 (m, 3 H), 7.41 - 7.48 (m, 1 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 34.00, 53.45, 127.60, 129.61, 129.87, 131.76, 132.24, 133.92, 172.08; LC/MS m/z (ESI^+) 200.16 $[M + H]^+$, t_R = 0.45 min; HRMS (ESI^+) observed 244.0124, calculated for $C_9H_9ClNO_2Na_2^+$ $[M + 2Na - H]^+$ 244.0112. ^{1}NMR data are consistent with literature values.[213]

2-Amino-3-(2-trifluoromethylphenyl)propanoic acid. HCl



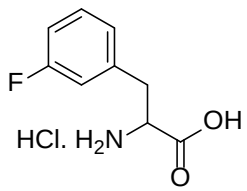
LC/MS *m/z* (ESI^+) 234.19 $[M + H]^+$, t_R = 0.69 min; HRMS (ESI^+) observed 278.0389, calculated for $C_{10}H_9F_3NO_2Na_2^+$ $[M + 2Na - H]^+$ 278.0376.

2-Amino-3-(2-methylphenyl)propanoic acid. HCl



^{1}H NMR (400 MHz, D_2O) δ ppm 2.28 (s., 3 H), 3.07 (dd, *J* = 14.7, 9.0 Hz, 1 H), 3.35 (dd, *J* = 14.5, 6.1 Hz, 1 H), 4.14 (dd, *J* = 8.8, 6.4 Hz, 1 H), 7.16 - 7.19 (m, 2 H), 7.21 - 7.25 (m, 2 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 18.28, 33.59, 53.40, 126.50, 128.11, 130.17, 130.93, 132.71, 137.29, 171.95; LC/MS *m/z* (ESI^+) 180.20 $[M + H]^+$, t_R = 0.46 min; HRMS (ESI^+) observed 224.0668, calculated for $C_{10}H_{12}NO_2Na_2^+$ $[M + 2Na - H]^+$ 224.0659. ^{1}NMR data consistent with literature values.[214]

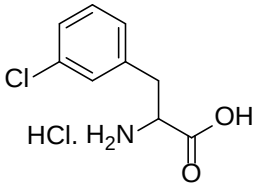
2-Amino-3-(3-fluorophenyl)propanoic acid. HCl



^{1}H NMR (400 MHz, D_2O) δ ppm 3.14 (dd, *J* = 14.5, 7.8 Hz, 1 H), 3.28 (dd, *J* = 14.4, 5.1 Hz, 1 H), 4.20 (dd, *J* = 7.8, 5.5 Hz, 1 H), 6.99 - 7.08 (m, 3 H), 7.34 (td, *J* = 7.9, 6.2 Hz, 1 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 35.42, 54.32, 114.72 (d, *J* = 21.3 Hz), 116.09 (d, *J* = 21.3 Hz), 125.24 (d, *J* = 2.2 Hz), 130.87 (d, *J* = 8.1 Hz), 136.66 (d, *J* = 8.1 Hz), 162.75 (d, *J* = 243.6 Hz), 171.79; LC/MS *m/z* (ESI^+) 184.16 $[M + H]^+$,

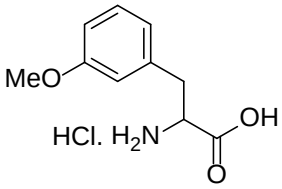
t_R = 0.39 min; HRMS (ESI$^+$) observed 228.0418, calculated for $C_9H_9FNO_2Na_2^+$ [M + 2Na - H]$^+$ 228.0408. ^{1}NMR data are consistent with literature values.[213]

2-Amino-3-(3-chlorophenyl)propanoic acid. HCl



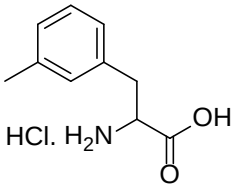
^{1}H NMR (400 MHz, D_2O) δ ppm 3.12 (dd, *J* = 14.7, 7.7 Hz, 1 H), 3.26 (dd, *J* = 14.7, 5.6 Hz, 1 H), 4.17 (dd, *J* = 7.7, 5.6 Hz, 1 H), 7.13 - 7.20 (m, 1 H), 7.26 - 7.34 (m, 3 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 35.40, 54.42, 127.74, 127.91, 129.22, 130.55, 134.11, 136.38, 171.87; LC/MS m/z (ESI$^+$) 200.17 [M + H]$^+$, t_R = 0.57 min; HRMS (ESI$^+$) observed 244.0123, calculated for $C_9H_9ClNO_2Na_2^+$ [M + 2Na - H]$^+$ 244.0112. ^{1}NMR data are consistent with literature values.[213]

2-Amino-3-(3-methoxyphenyl)propanoic acid. HCl



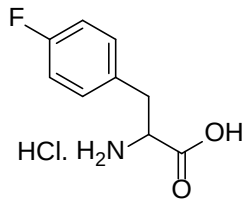
^{1}H NMR (400 MHz, D_2O) δ ppm 3.12 (dd, *J* = 14.5, 7.7 Hz, 1 H), 3.25 (dd, *J* = 14.5, 5.9 Hz, 1 H), 3.76 (s., 3 H), 4.23 (dd, *J* = 7.8, 5.5 Hz, 1 H), 6.82 - 6.93 (m, 3 H), 7.29 (t, *J* = 8.0 Hz, 1 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 35.62, 54.24, 55.30, 113.46, 114.94, 122.08, 130.45, 135.80, 159.27, 171.68; LC/MS *m/z* (ESI$^+$) 196.20 [M + H]$^+$, t_R = 0.43 min; HRMS (ESI$^+$) observed 240.0622, calculated for $C_{10}H_{12}NO_3Na_2^+$ [M + 2Na - H]$^+$ 240.0608. ^{1}NMR data are consistent with literature values.[215]

2-Amino-3-(3-methylphenyl)propanoic acid. HCl



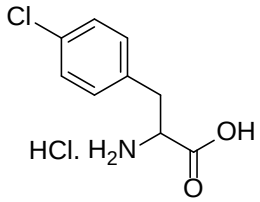
LC/MS *m/z* (ESI$^+$) 180.20 [M + H]$^+$, t_R = 0.43 min; HRMS (ESI$^+$) observed, calculated for $C_{10}H_{12}NO_2Na_2^+$ [M + 2Na - H]$^+$ 224.0659.

2-Amino-3-(4-fluorophenyl)propanoic acid. HCl



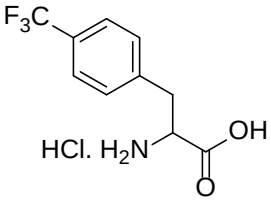
LC/MS *m/z* (ESI^+) 184.18 $[M + H]^+$, t_R = 0.43 min; HRMS (ESI^+) observed 228.0422, calculated for $C_9H_9FNO_2Na_2^+$ $[M + 2Na - H]^+$ 228.0408.

2-Amino-3-(4-chlorophenyl)propanoic acid. HCl



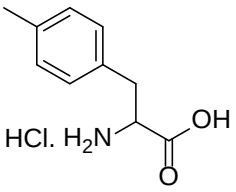
LC/MS m/z (ESI^+) 200.19 $[M + H]^+$, t_R = 0.63 min; HRMS (ESI^+) observed 244.0124, calculated for $C_9H_9ClNO_2Na_2^+$ $[M + 2Na - H]^+$ 244.0112.

2-Amino-3-(4-trifluoromethylphenyl)propanoic acid. HCl



LC/MS *m/z* (ESI^+) 234.19 $[M + H]^+$, t_R = 1.02 min; HRMS (ESI^+) observed 278.0392, calculated for $C_{10}H_9F_3NO_2Na_2^+$ $[M + 2Na - H]^+$ 278.0376.

2-Amino-3-(4-methylphenyl)propanoic acid. HCl



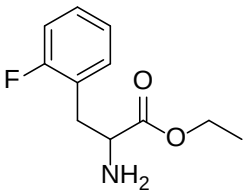
LC/MS *m/z* (ESI^+) 180.22 $[M + H]^+$, t_R = 0.43 min; HRMS (ESI^+) observed 224.0676, calculated for $C_{10}H_{12}NO_2Na_2^+$ $[M + 2Na - H]^+$ 224.0659.

Esterification of Phenylalanine Analogues

The amino acid HCl salt or mixture of amino acid HCl salt and ethyl ester were suspended in 2 mL of EtOH, and cooled to 4 °C. Thionyl chloride (1.4 equiv. based on total mass of starting material) was added dropwise, and the suspension allowed to stir at room temperature for 48 h. The solvent was removed *in vacuo* giving the desired product as a HCl salt that was used without further purification unless stated.

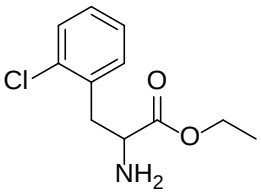
Carbon spectra where the number of peaks observed is less than the number of peaks expected due to overlap are labelled with an asterisk.

Ethyl 2-amino-3-(2-fluorophenyl)propanoate



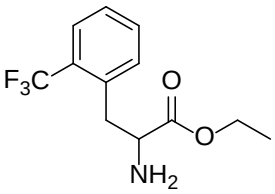
^{1}H NMR (400 MHz, D_2O) δ ppm 1.16 (t, *J* = 7.2 Hz, 3 H), 3.24 (dd, *J* = 14.7, 7.0 Hz, 1 H), 3.33 (dd, *J* = 14.7, 6.5 Hz, 1 H), 4.19 (q, *J* = 7.2 Hz, 2 H), 4.35 (app. t, *J* = 6.8 Hz, 1 H), 7.09 - 7.14 (m, 1 H), 7.14 - 7.18 (m, 1 H), 7.25 (td, *J* = 7.6, 1.7 Hz, 1 H), 7.34 (tdd, *J* = 7.8, 7.8, 5.6, 1.8 Hz, 1 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.05, 29.53 (d, *J* = 2.9 Hz), 53.13, 63.68, 115.65 (d, *J* = 21.3 Hz), 120.63 (d, *J* = 16.1 Hz), 124.91 (d, *J* = 3.7 Hz), 130.33 (d, *J* = 8.8 Hz), 131.82 (d, *J* = 4.4 Hz), 161.21 (d, *J* = 242.1 Hz), 169.40; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 212.31, t_R = 1.26 min; HRMS (ESI$^+$) observed 212.1093, calculated for $C_{11}H_{15}FNO_2^+$ $[M + H]^+$ 212.1087; ν_{max} cm^{-1} (neat) 2871 (C-H), 1741 (C=O), 1493 (phenyl).

Ethyl 2-amino-3-(2-chlorophenyl)propanoate



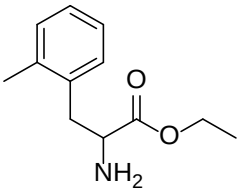
The resultant solid was dissolved in a 1:1 mixture of ethyl acetate and water. A saturated aqueous solution of $KHCO_3$ was added dropwise until the aqueous layer reached pH 10 as indicated by pH paper. The aqueous layer was extracted with 3 x 5 mL EtOAc. The combined organic layers were washed with brine and the solvent removed *in vacuo* to give the product as a free amine. ^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 1.02 (t, *J* = 7.1 Hz, 3 H), 2.89 (dd, *J* = 13.4, 7.2 Hz, 1 H), 2.95 (dd, *J* = 13.6, 7.5 Hz, 1 H), 3.56 (app. t, *J* = 7.3 Hz, 1 H), 3.95 (q, *J* = 7.1 Hz, 2 H), 7.20 - 7.30 (m, 3 H), 7.35 - 7.41 (m, 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 14.24, 54.36, 60.71, 127.47, 128.88, 129.58, 132.28, 133.67, 135.78, 175.28; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 228.22, t_R = 1.34 min; HRMS (ESI$^+$) observed 228.0797, calculated for $C_{11}H_{15}ClNO_2^+$ $[M + H]^+$ 228.0791; ν_{max} cm^{-1} (neat) 2922 (C-H), 1736 (C=O), 1474 (phenyl).

Ethyl 2-amino-3-(2-trifluoromethylphenyl)propanoate



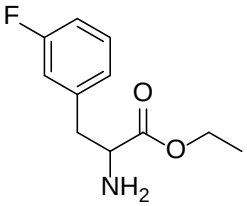
^{1}H NMR (400 MHz, CD_3OD) δ ppm 1.10 (t, *J* = 7.2 Hz, 3 H), 3.38 (dd, *J* = 14.3, 7.5 Hz, 1 H), 3.45 (dd, *J* = 14.4, 8.6 Hz, 1 H), 4.09 - 4.20 (m, 2 H), 4.29 (dd, *J* = 8.5, 7.5 Hz, 1 H), 7.52 - 7.59 (m, 2 H), 7.65 - 7.71 (m, 1 H), 7.78 (d, *J* = 7.6 Hz, 1 H); ^{13}C NMR (101 MHz, CD_3OD) δ ppm 12.59, 33.11, 53.51, 62.13, 126.18 (q, *J* = 5.9 Hz), 127.99, 128.51, 128.80, 131.95, 132.41, 132.93, 168.23; LC/MS *m/z* (ESI^+) $[M + H]^+$ 262.26, t_R = 1.42 min; HRMS (ESI^+) observed 262.1053, calculated for $C_{12}H_{15}F_3NO_2^+$ $[M + H]^+$ 262.1055; ν_{max} cm^{-1} (neat) 2923 (C-H), 1744 (C=O).

Ethyl 2-amino-3-(2-methylphenyl)propanoate



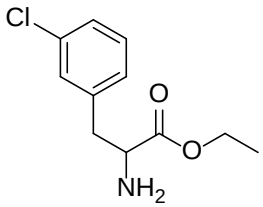
^{1}H NMR (400 MHz, D_2O) δ ppm 1.13 (t, *J* = 7.2 Hz, 3 H), 2.27 (s., 3 H), 3.12 (dd, *J* = 14.4, 8.6 Hz, 1 H), 3.31 (dd, *J* = 14.4, 7.1 Hz, 1 H), 4.17 (qd, *J* = 7.2, 1.2 Hz, 2 H), 4.24 (dd, *J* = 8.5, 7.2 Hz, 1 H), 7.10 - 7.26 (m, 4 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.04, 18.27, 33.54, 53.06, 63.58, 126.52, 128.23, 130.20, 130.93, 132.32, 137.22, 169.73; LC/MS *m/z* (ESI^+) $[M + H]^+$ 208.29, t_R = 1.30min; HRMS (ESI^+) observed 208.1346, calculated for $C_{12}H_{18}NO_2^+$ $[M + H]^+$ 208.1338; ν_{max} cm^{-1} (neat) 2873 (C-H), 1741 (C=O), 1494 (phenyl).

Ethyl 2-amino-3-(3-fluorophenyl)propanoate



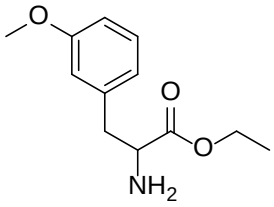
1H NMR (400 MHz, D_2O) δ ppm 1.18 (t, *J* = 7.2 Hz, 3 H), 3.19 (dd, *J* = 14.5, 7.3 Hz, 1 H), 3.28 (dd, *J* = 14.7, 6.2 Hz, 1 H), 4.22 (q, *J* = 7.2 Hz, 2 H), 4.34 (dd, *J* = 7.3, 6.3 Hz, 1 H), 6.99 (app. dt, *J* = 9.9, 2.0 Hz, 1 H), 7.02 - 7.08 (m, 2 H), 7.32 - 7.39 (m, 1 H); 13C NMR (101 MHz, D_2O) δ ppm 13.10, 35.26, 53.88, 63.66, 114.86 (d, *J* = 22.0 Hz), 116.14 (d, *J* = 21.3 Hz), 125.28 (d, *J* = 2.9 Hz), 130.94 (d, *J* = 8.1 Hz), 136.18 (d, *J* = 7.3 Hz), 162.76 (d, *J* = 246.5 Hz), 169.36; LC/MS m/z (ESI^+) $[M + H]^+$ 212.27, t_R = 1.27 min; HRMS (ESI^+) observed 212.1093, calculated for $C_{11}H_{15}FNO_2^+$ $[M + H]^+$ 212.1087; ν_{max} cm^{-1} (neat) 2982 (C-H), 1741 (C=O), 1489 (phenyl).

Ethyl 2-amino-3-(3-chlorophenyl)propanoate



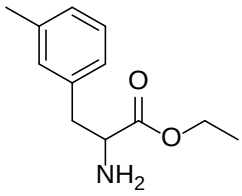
^{1}H NMR (400 MHz, D_2O) δ ppm 1.17 (t, *J* = 7.2 Hz, 3 H), 3.17 (dd, *J* = 14.4, 7.2 Hz, 1 H), 3.24 (dd, *J* = 14.7, 6.6 Hz, 1 H), 4.20 (q, *J* = 7.1 Hz, 1 H), 4.33 (app. t, *J* = 6.8 Hz, 1 H), 7.13 - 7.16 (m, 1 H), 7.26 (s., 1 H), 7.29 - 7.35 (m, 2 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.13, 35.24, 53.87, 63.66, 127.79, 128.05, 129.29, 130.61, 134.12, 135.85, 169.33; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 228.22, t_R = 1.37 min; HRMS (ESI$^+$) observed 228.0794, calculated for $C_{11}H_{15}ClNO_2^+$ $[M + H]^+$ 228.0791; ν_{max} cm^{-1} (neat) 2870 (C-H), 1739 (C=O), 1477 (phenyl).

Ethyl 2-amino-3-(3-methoxyphenyl)propanoate



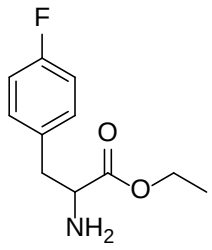
^{1}H NMR (400 MHz, CD_3OD) δ ppm 1.28 (t, *J* = 7.2 Hz, 3 H), 3.18 (dd, *J* = 14.2, 7.2 Hz, 1 H), 3.25 (dd, *J* = 14.2, 6.5 Hz, 1 H), 3.82 (s., 3 H), 4.28 (q, *J* = 7.1 Hz, 2 H), 4.33 (app. d, *J* = 6.8 Hz, 1 H), 6.83 - 6.88 (m, 2 H), 6.91 (ddd, *J* = 8.2, 2.2, 1.1 Hz, 1 H), 7.30 (t, *J* = 8.4 Hz, 1 H); ^{13}C NMR (101 MHz, CD_3OD) δ ppm 14.44, 37.61, 55.30, 55.87, 63.78, 114.49, 116.30, 122.73, 131.35, 136.87, 161.81, 170.14; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 224.29, t_R = 1.24 min; HRMS (ESI$^+$) observed 224.129, calculated for $C_{12}H_{18}NO_3^+$ $[M + H]^+$ 224.1287; ν_{max} cm^{-1} (neat) 2836 (C-H), 1739 (C=O), 1490 (phenyl).

Ethyl 2-amino-3-(3-methylphenyl)propanoate



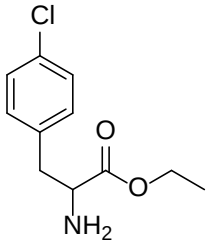
^{1}H NMR (400 MHz, D_2O) δ ppm 1.18 (t, *J* = 7.2 Hz, 3 H), 2.26 (s., 3 H), 3.15 (dd, *J* = 14.4, 7.2 Hz, 1 H), 3.22 (dd, *J* = 14.4, 6.2 Hz, 1 H), 4.21 (q, *J* = 7.1 Hz, 2 H), 4.30 (dd, *J* = 7.2, 6.3 Hz, 1 H), 7.02 (d, *J* = 7.3 Hz, 1 H), 7.05 (s., 1 H), 7.16 (d, *J* = 7.8 Hz, 1 H), 7.24 (d, *J* = 7.6 Hz, 1 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.11, 20.33, 35.56, 54.13, 63.57, 126.30, 128.65, 129.15, 129.98, 133.70, 139.38, 169.59; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 208.28, t_R = min 1.32; HRMS (ESI$^+$) observed 208.1340, calculated for $C_{12}H_{18}NO_2^+$ $[M + H]^+$ 208.1338; ν_{max} cm^{-1} (neat) 2924 (C-H), 1741 (C=O), 1489 (phenyl).

Ethyl 2-amino-3-(4-fluorophenyl)propanoate



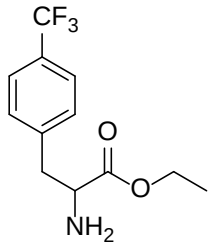
^{1}H NMR (400 MHz, D_2O) δ ppm 1.18 (t, *J* = 7.2 Hz, 3 H), 3.17 (dd, *J* = 14.5, 7.3 Hz, 1 H), 3.25 (dd, *J* = 14.5, 6.1 Hz, 1 H), 4.21 (q, *J* = 7.2 Hz, 2 H), 4.31 (dd, *J* = 7.2, 6.2 Hz, 1 H), 7.04 - 7.11 (m, 2 H), 7.20 - 7.25 (m, 2 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.10, 34.82, 54.07, 63.61, 115.88 (d, *J* = 22.0 Hz), 129.56 (d, *J* = 2.9 Hz), 131.17 (d, *J* = 8.1 Hz), 162.25 (d, *J* = 242.8 Hz), 169.48; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 212.27, t_R = 1.23 min; HRMS (ESI$^+$) observed 212.1090, calculated for $C_{11}H_{15}FNO_2^+$ $[M + H]^+$ 212.1087; ν_{max} cm^{-1} (neat) 2906 (C-H), 1740 (C=O), 1510 (phenyl).

Ethyl 2-amino-3-(4-chlorophenyl)propanoate



^{1}H NMR (400 MHz, D_2O) δ ppm 1.17 (t, *J* = 7.2 Hz, 3 H), 3.17 (dd, *J* = 14.9, 7.3 Hz, 1 H), 3.24 (dd, *J* = 15.0, 6.4 Hz, 1 H), 4.20 (q, *J* = 7.1 Hz, 1 H), 4.32 (t, *J* = 6.8 Hz, 1 H), 7.17 - 7.21 (m, 2 H), 7.33 - 7.38 (m, 2 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.09, 34.99, 53.91, 63.64, 129.09, 130.91, 132.40, 133.30, 169.40; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 2.28, t_R = 1.37 min; HRMS (ESI$^+$) observed 228.0797, calculated for $C_{11}H_{15}ClNO_2^+$ $[M + H]^+$ 228.0791; ν_{max} cm^{-1} (neat) 2870 (C-H), 1739 (C=O), 1492 (phenyl).

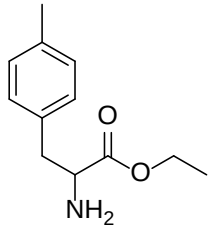
Ethyl 2-amino-3-(4-trifluoromethylphenyl)propanoate



The resultant solid was dissolved in a 1:1 mixture of ethyl acetate and water. The aqueous layer was acidified with 1 M HCl and the combined layers washed with 3 x 5 mL 0.2 M HCl. The combined aqueous layers were evaporated *in vacuo* to give the desired product as a HCl salt. ^{1}H NMR (400 MHz, CD_3OD) δ ppm 1.25 (t, *J* = 7.2 Hz, 3 H), 3.29 (dd, *J* = 14.3, 7.5 Hz, 1 H), 3.35 (dd, *J* = 14.4, 6.8 Hz, 1 H), 4.27 (q, *J* = 7.1 Hz, 2 H), 4.39 (app. t, *J* = 7.1 Hz, 1 H), 7.51 (d, *J* = 8.2 Hz, 2 H), 7.71 (d, *J* = 8.1 Hz, 2 H); ^{13}C NMR (101 MHz, CD_3OD) δ ppm 12.80, 35.78, 53.41, 62.38, 125.53 (q, *J* = 3.7 Hz), 129.85, 138.74, 168.35; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 262.27, t_R = 1.45 min; HRMS (ESI$^+$)

observed 262.1057, calculated for $C_{12}H_{15}F_3NO_2^+$ $[M + H]^+$ 262.1055; ν_{max} cm^{-1} (neat) 2985 (C-H), 1725 (C=O), (phenyl).

Ethyl 2-amino-3-(4-methylphenyl)propanoate



^{1}H NMR (400 MHz, D_2O) δ ppm 1.18 (t, *J* = 7.2 Hz, 3 H), 2.25 (s., 3 H) 3.14 (dd, *J* = 14.5, 7.2 Hz, 1 H), 3.22 (dd, *J* = 14.2, 6.1 Hz, 1 H), 4.21 (q, *J* = 7.2 Hz, 2 H), 4.29 (dd, *J* = 7.2, 6.3 Hz, 1 H), 7.11 (app. d, *J* = 8.4 Hz, 2 H), 7.19 (app. d, *J* = 8.3 Hz, 2 H); ^{13}C NMR (101 MHz, D_2O) δ ppm 13.10, 20.10, 35.21, 54.16, 63.57, 129.34, 129.75, 130.53, 138.28, 169.61; LC/MS *m/z* (ESI^+) $[M + H]^+$ 262.36, t_R = 1.33 min; HRMS (ESI^+) observed, 208.1350 calculated for $C_{12}H_{18}NO_2^+$ $[M + H]^+$ 208.1338; ν_{max} cm^{-1} (neat) 2829 (C-H), 1737 (C=O), 1500 (phenyl).

Reductive Alkylation of Ethyl Esters of Phenylalanine Analogues

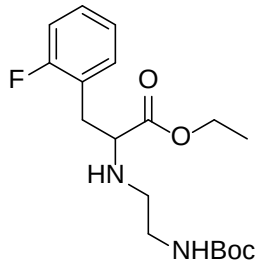
The amino acid ethyl ester was dissolved in a 5:1 solution of dry MeOH:Acetic Acid to give a solution of 0.5 M. 3 g of activated 4 Å molecular sieves were added. A 0.5 M solution of *tert*-butyl (2-oxoethyl)carbamate was prepared in dry MeOH and 1.33 equiv. added to the reaction mixture. A 0.6 M solution of sodium cyanoborohydride was prepared in dry MeOH and 1.4 equiv. added to the reaction mixture which was left for 20 h without stirring.

The reaction mixture was filtered and the molecular sieves washed with 10 mL MeOH. 10 mL aqueous sodium hydrogen carbonate/sodium carbonate (pH 9) was added and the mixture extracted with 3 x 10 mL DCM. The combined organic layers were washed with 10 mL brine and the solvent removed *in vacuo*.

Impurities were removed using preparative scale HPLC, however for many examples it was not possible to isolate pure product. These cases were carried through to the next step without purification. No NMR or IR data are given for these examples.

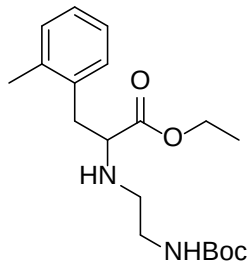
Carbon spectra where the number of peaks observed is less than the number of peaks expected due to overlap are labelled with an asterisk.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(2-fluorophenyl)propanoate



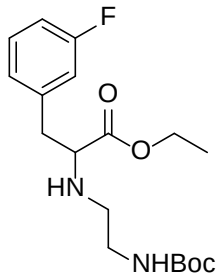
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.19 (t, *J* = 7.2 Hz, 3 H, $CH_2C\mathbf{H}_3$), 1.46 (s., 9 H, $C(C\mathbf{H}_3)_3$), 2.57 (ddd, *J* = 12.0, 7.0, 4.9 Hz, 1 H), 2.78 (ddd, *J* = 12.0, 7.3, 4.5 Hz, 1 H), 2.98 (ddd, *J* = 20.9, 13.4, 7.1 Hz, 2 H), 3.05 - 3.22 (m, 2 H), 3.51 (t, *J* = 7.2 Hz, 1 H) 4.13 (q, *J* = 7.1 Hz, 2 H, $C\mathbf{H}_2CH_3$) 4.79 - 4.90 (m, 1 H) 7.01 - 7.11 (m, 2 H), 7.17 - 7.27 (m, 2 H); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 14.10, 28.41, 33.1, 47.20, 60.87, 61.15, 115.26 (d, *J* = 22.0 Hz) 123.94 (d, *J* = 3.7 Hz) 128.56 (d, *J* = 8.1 Hz) 131.54 (d, *J* = 4.4 Hz) 155.99, 174.43*; LC/MS *m/z* (ESI^+) $[M + H]^+$ 355.39, t_R = 2.13 min; HRMS (ESI^+) observed 299.1396, calculated for $C_{14}H_{20}FN_2O_4^+$ $[M + 2H - {}^tBu]^+$ 299.1407.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(2-methylphenyl)propanoate



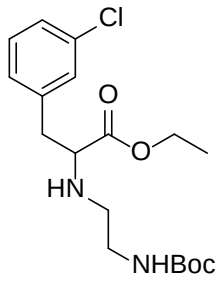
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.08 (t, *J* = 7.2 Hz, 3 H), 1.45 (s., 9 H), 2.36 (s., 3 H), 2.58 (m, 1 H), 2.64 - 2.71 (m, 1 H), 2.91 (dd, *J* = 13.6, 8.7 Hz, 1 H), 3.02 (dd, *J* = 13.6, 6.6 Hz, 1 H), 3.14 (app. t, *J* = 6.2 Hz, 2 H), 3.53 (dd, *J* = 8.6, 6.7 Hz, 1 H), 4.03 (m, 2 H), 7.08 - 7.18 (m, 4 H); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 14.11, 19.49, 28.42, 37.26, 47.33, 60.78, 125.87, 126.84, 129.89, 130.40, 136.39, 156.01, 174.89*; LC/MS *m/z* (ESI^+) $[M + H]^+$ 351.42, t_R = 2.13 min; HRMS (ESI^+) observed 295.1648, calculated for $C_{15}H_{23}N_2O_4^+$ $[M + 2H - {}^tBu]^+$ 295.1658.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(3-fluorophenyl)propanoate



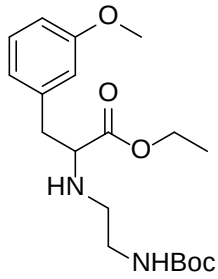
LC/MS *m/z* (ESI^+) $[M + H]^+$ 355.40, t_R = 2.12 min; HRMS (ESI^+) observed 299.1400, calculated for $C_{14}H_{20}FN_2O_4^+$ $[M + 2H - {}^tBu]^+$ 299.1407.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(3-chlorophenyl)propanoate



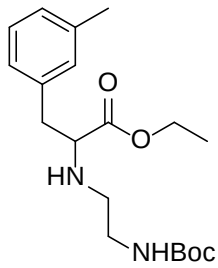
LC/MS *m/z* (ESI^+) $[M + H]^+$ 371.38, t_R = 2.21 min; HRMS (ESI^+) observed 315.1103, calculated for $C_{14}H_{20}ClN_2O_4^+$ $[M + 2H - {}^tBu]^+$ 315.1111.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(3-methoxyphenyl)propanoate



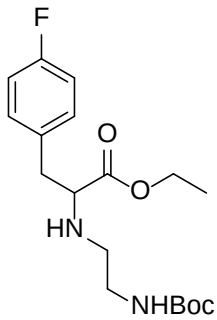
LC/MS *m/z* (ESI^+) $[M + H]^+$ 367.42, t_R = 2.01 min; HRMS (ESI^+) observed 311.1601, calculated for $C_{15}H_{23}N_2O_5^+$ $[M + 2H - {}^tBu]^+$ 311.1607.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(3-methylphenyl)propanoate



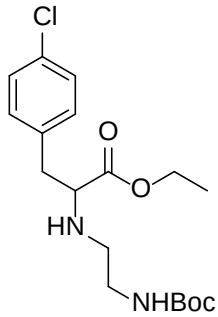
LC/MS *m/z* (ESI^+) $[M + H]^+$ 351.43, t_R = 2.16 min; HRMS (ESI^+) observed 295.1650, calculated for $C_{15}H_{23}N_2O_4^+$ $[M + 2H - {}^tBu]^+$ 295.1658.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(4-fluorophenyl)propanoate



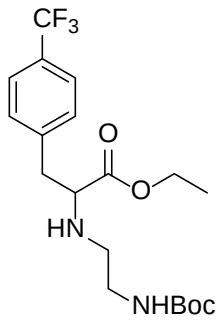
LC/MS *m/z* (ESI^+) $[M + H]^+$ 355.42, t_R = 2.00 min; HRMS (ESI^+) observed 299.1399, calculated for $C_{14}H_{20}FN_2O_4^+$ $[M + 2H - {}^tBu]^+$ 299.1407.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(4-chlorophenyl)propanoate



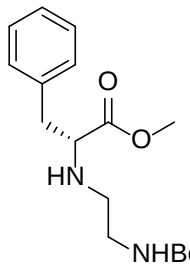
LC/MS *m/z* (ESI^+) $[M + H]^+$ 371.37, t_R = 2.17 min; HRMS (ESI^+) observed 315.1102, calculated for $C_{14}H_{20}ClN_2O_4^+$ $[M + 2H - {}^tBu]^+$ 315.1111.

Ethyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-(4-trifluoromethylphenyl)propanoate



LC/MS *m/z* (ESI^+) $[M + H]^+$ 405.41, t_R = 2.20 min; HRMS (ESI^+) observed 349.1363, calculated for $C_{15}H_{20}F_3N_2O_4^+$ $[M + 2H - {}^tBu]^+$ 349.1375.

Compound 89: Methyl 2-((2-((tert-butoxycarbonyl)amino)ethyl)amino)-3-phenylpropanoate



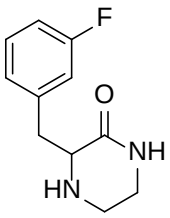
^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 1.46 (s., 9 H), 2.50 - 2.62 (m, 1 H), 2.76 (ddd, *J* = 11.9, 6.8, 4.7 Hz, 1 H), 2.90 (dd, *J* = 13.6, 7.6 Hz, 1 H), 2.99 (dd, *J* = 13.6, 6.1 Hz, 1 H), 3.05 - 3.22 (m, 2 H), 3.50 (dd, *J* = 7.4, 6.4 Hz, 1 H), 3.69 (s., 3 H), 4.82 (br. s., 1 H), 7.20 (d, *J* = 6.8 Hz, 2 H), 7.22 - 7.28 (m, 1 H), 7.29 - 7.35 (m, 2 H); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 28.35, 39.53, 47.74, 52.09, 62.13, 127.05, 128.59, 129.27, 156.42, 160.42*; LC/MS *m/z* (ESI^+) $[M + H]^+$ 323.26, t_R = 1.12 min; HRMS (ESI^+) observed 267.1349, calculated for $C_{13}H_{22}N_2O_4^+$ $[M + 2H - {}^tBu]^+$ 267.1345. ^{1}NMR data are consistent with literature values.[203]

Preparation of Ketopiperazines

Ethyl 2-amino-3-phenylpropanoates prepared in the preceding step were dissolved in DCM to a concentration of 0.35 M. Triethylsilane (2.5 equiv.) was added, followed by trifluoroacetic acid (13 equiv.). The reaction mixture was stirred at room temperature in a sealed vial for 1 h. The solvent was removed *in vacuo* and the resulting gum triturated with diethyl ether. The resulting solid was washed once with diethyl ether, dissolved in 3 mL DCM and 300 µL triethylamine. The resulting mixture was stirred at 60 °C in a sealed vial for 48 h. Solvent was removed *in vacuo* and the resultant solid dissolved in 2 mL DCM and 2 mL H_2O. Additional H_2O was added until the aqueous phase was clear, and the biphasic mixture was extracted with 3 x 3 mL DCM. The combined organic phase was dried *in vacuo* and the desired product isolated using preparative scale HPLC.

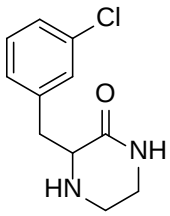
Carbon spectra where the number of peaks observed is less than the number of peaks expected due to overlap are labelled with an asterisk.

Compound 109: 3-(3-Fluorobenzyl)piperazin-2-one



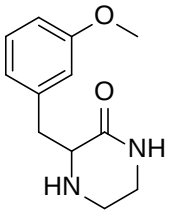
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.70 (ddd, *J* = 13.0, 9.0, 4.4 Hz, 1 H), 2.79 (dd, *J* = 13.9, 9.1 Hz, 1 H), 2.91 (dt, *J* = 12.7, 3.9 Hz, 1 H), 3.02 - 3.20 (m, 3 H), 3.44 (dd, *J* = 8.9, 3.7 Hz, 1 H), 6.97 - 7.04 (m, 1 H), 7.06 - 7.11 (m, 2 H), 7.26 - 7.34 (m, 1 H), 7.65 (br. s., 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 37.35, 41.43, 42.56, 59.79, 113.16 (d, *J* = 21.3 Hz), 116.50 (d, *J* = 22.0 Hz), 125.96 (d, *J* = 2.2 Hz), 130.23 (d, *J* = 8.1 Hz), 162.48 (d, *J* = 244.3 Hz)*; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 209.26, t_R = 1.02 min; HRMS (ESI$^+$) observed 209.1082, calculated for $C_{11}H_{14}FN_2O^+$ $[M + H]^+$ 209.1090; ν_{max} cm^{-1} (neat) 3233 (N-H), 1656 (C=O).

Compound 110: 3-(3-Chlrorobenzyl)piperazin-2-one



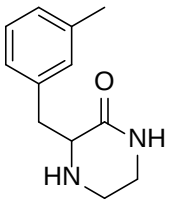
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.69 (ddd, J = 13.3, 8.9, 4.5 Hz, 1 H), 2.78 (dd, J = 13.9, 8.9 Hz, 1 H), 2.90 (dt, J = 12.7, 4.0 Hz, 1 H), 3.01 - 3.18 (m, 3 H), 3.41 (dd, J = 8.9, 3.7 Hz, 1 H), 7.19 - 7.23 (m, 1 H), 7.23 - 7.26 (m, 1 H,) 7.27 - 7.30 (m, 1 H), 7.30 - 7.33 (m, 1 H), 7.63 (br. s., 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 37.31, 41.46, 42.68, 59.84, 126.35, 128.59, 129.73, 130.23, 133.02, 142.53, 170.66; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 225.25, t_R = 1.13 min; HRMS (ESI$^+$) observed 225.0801, calculated for $C_{11}H_1ClN_2O^+$ $[M + H]^+$ 225.0795; ν_{max} cm^{-1} (neat) 3247 (N-H), 2934 (C-H), 2871 (C-H), 1659 (C=O).

Compound 111: 3-(3-Methoxyobenzyl)piperazin-2-one



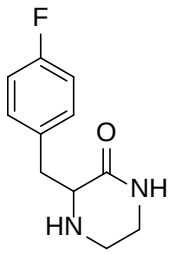
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.64 - 2.73 (m, 2 H), 2.90 (dt, J = 12.6, 3.9 Hz, 1 H), 3.01 - 3.18 (m, 3 H), 3.39 (dd, J = 9.4, 3.5 Hz, 1 H), 3.73 (s., 3 H), 6.68 - 6.84 (m, 3 H), 7.19 (t, J = 8.1 Hz, 1 H), 7.62 (br. s., 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 37.84, 41.49, 42.66, 55.32, 60.11, 111.87, 115.40, 122.00, 129.53, 141.35, 159.57, 170.81; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 221.30, t_R = 1.08 min; HRMS (ESI$^+$) observed 221.1282, calculated for $C_{12}H_{17}N_2O_2^+$ $[M + H]^+$ 221.1290; ν_{max} cm^{-1} (neat) 3295 (N-H), 2937 (C-H), 1660 (C=O).

Compound 112: 3-(3-Methylbenzyl)piperazin-2-one



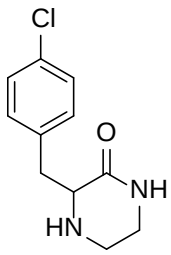
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.28 (s., 3 H), 2.62 - 2.71 (m, 2 H), 2.90 (dt, J = 12.6, 3.8 Hz, 1 H), 3.02 - 3.19 (m, 4 H), 6.97 - 7.07 (m, 3 H), 7.12 - 7.20 (m, 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 21.51, 37.77, 41.48, 42.69, 60.23, 126.79, 127.08, 128.47, 130.44, 137.50, 139.71, 170.88; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 205.30, t_R = 1.12 min; HRMS (ESI$^+$) observed 205.1328, calculated for $C_{12}H_{17}FN_2O^+$ $[M + H]^+$ 205.1341; ν_{max} cm^{-1} (neat) 3269 (N-H), 2916 (C-H), 2848 (C-H), 1664 (C=O).

Compound 113: 3-(4-Fluorobenzyl)piperazin-2-one



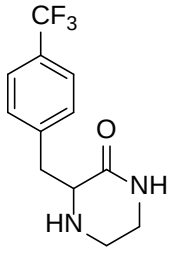
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.68 (ddd, *J* = 13.1, 9.2, 4.4 Hz, 1 H), 2.75 (dd, *J* = 13.8, 9.0 Hz, 1 H), 2.89 (dt, *J* = 12.7, 4.0 Hz, 1 H), 3.01 - 3.15 (m, 3 H), 3.37 (dd, *J* = 9.0, 3.7 Hz, 1 H), 7.03 - 7.12 (m, 2 H), 7.23 - 7.30 (m, 2 H), 7.61 (br. s., 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 36.88, 41.46, 42.69, 60.08, 115.10 (d, *J* = 19.8 Hz), 131.58 (d, *J* = 8.1 Hz), 135.88 (d, *J* = 2.9 Hz), 161.26 (d, *J* = 241.4 Hz), 170.77; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 209.26, t_R = 1.08 min; HRMS (ESI$^+$) observed 209.1081, calculated for $C_{11}H_{14}FN_2O^+$ $[M + H]^+$ 209.1090; ν_{max} cm^{-1} (neat) 3240 (N-H), 2934 (C-H), 2872 (C-H), 1656 (C=O).

Compound 114: 3-(4-Chlrorobenzyl)piperazin-2-one



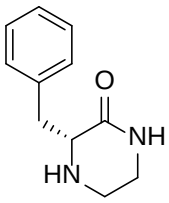
^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.68 (ddd, *J* = 12.9, 9.0, 4.4 Hz, 1 H), 2.76 (dd, *J* = 13.8, 9.2 Hz, 1 H), 2.88 (dt, *J* = 12.6, 3.9 Hz, 1 H), 3.01 - 3.18 (m, 3 H), 3.38 (dd, *J* = 8.9, 3.7 Hz, 1 H), 7.24 - 7.28 (m, 2 H), 7.29 - 7.34 (m, 2 H), 7.61 (br. s., 1 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 37.04, 41.47, 42.69, 59.96, 128.35, 131.03, 131.74, 138.89, 170.70; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 225.24, t_R = 1.13 min; HRMS (ESI$^+$) observed 225.0794, calculated for $C_{11}H_{14}ClN_2O^+$ $[M + H]^+$ 225.0795; ν_{max} cm^{-1} (neat) 3235 (N-H), 2934 (C-H), 2872 (C-H), 1661 (C=O).

Compound 115: 3-(4-Trifluoromethylbenzyl)piperazin-2-one



^{1}H NMR (400 MHz, DMSO-d_6) δ ppm 2.69 (ddd, *J* = 13.1, 8.9, 4.3 Hz, 1 H), 2.82 - 2.93 (m, 2 H), 3.02 - 3.13 (m, 2 H), 3.16 - 3.23 (m, 1 H), 3.44 (dd, *J* = 8.9, 3.7 Hz, 1 H), 7.47 (d, *J* = 8.1 Hz, 2 H), 7.62 (d, *J* = 8.1 Hz, 2 H); ^{13}C NMR (101 MHz, DMSO-d_6) δ ppm 37.54, 41.44, 42.68, 59.85, 125.22, 130.67, 145.01, 153.76, 170.63*; LC/MS *m/z* (ESI$^+$) $[M + H]^+$ 259.26, t_R = 1.22 min; HRMS (ESI$^+$) observed 259.1051, calculated for $C_{12}H_{14}F_3N_2O^+$ $[M + H]^+$ 259.1058; ν_{max} cm^{-1} (neat) 3243 (N-H), 3019 (C-H), 1665 (C=O).

Compound 90: 3-Benzylpiperazin-2-one



^{1}H NMR (400 MHz, $CDCl_3$) δ ppm 2.83 - 2.96 (m, 2 H), 3.09 (dt, *J* = 12.4, 3.5 Hz, 1 H), 3.27 (dq, *J* = 11.4, 3.5 Hz, 1 H), 3.39 (dd, *J* = 10.7, 4.5 Hz, 1 H), 3.45 (dd, *J* = 13.8, 3.4 Hz, 1 H), 3.64 (dd, *J* = 9.9, 3.4 Hz, 1 H), 6.01 (br. s., 1 H), 7.23-7.36 (m, 4 H); ^{13}C NMR (101 MHz, $CDCl_3$) δ ppm 37.80, 41.71, 43.18, 60.46, 126.72, 128.72, 129.37, 138.24*; LC/MS *m/z* (ESI^+) $[M + H]^+$ 191.25, t_R = 0.81 min; HRMS (ESI^+) observed 191.1176, calculated for $C_{11}H_{15}N_2O^+$ $[M + H]^+$ 191.1184; ν_{max} cm^{-1} (neat) 3243 (N-H), 3019 (C-H), 1665 (C=O). ^{1}NMR data are consistent with literature values.[203]

Appendices for Chapter 2 - PHD Family Analysis

Appendix 2.1 - PHDs with Known Structures used for Initial HMM Model

PDB ID	Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
1XWH	AIRE	AIRE(1)	4557291	295	342
3RSN	ASH2L	ASH2L	157412280	103	160
1F62	BAZ1B	BAZ1B	14670392	1184	1232
2RI7	FALZ	FALZ(2)	38788260	2724	2773
2L43	BRD1	BRD1(1)	11321642	213	263
1MM2	CHD4	CHD4(2)	51599156	448	494
2L5U	CHD4	CHD4(1)	51599156	369	416
1WEM	DIDO1	DIDO1	18375617	268	320
2KWO	DPF3	DPF3(1)	60459281	260	317
2KWO	DPF3	DPF3(2)	60459281	318	356
4BBQ	FBXL11	FBXL11	16306580	616	677
2QIC	ING1	ING1	38201661	352	400
2G6Q	ING2	ING2	4504695	211	259
1X4I	ING3	ING3	38201655	359	407
1WEN	ING4	ING4	38201670	194	243
3C6W	ING5	ING5	18644730	185	234
1WEV	INTS12	INTS12	21361851	161	214
2KGI	JARID1A	JARID1A(3)	110618244	1606	1660
3KV6	KIAA1718	KIAA1718	90093355	38	86
2YSM	MLL3	MLL3(2)	91718902	340	389
2YSM	MLL3	MLL3(3)	91718902	340	389
2LN0	MYST3	MYST3(1)	150378463	206	263
3V43	MYST3	MYST3(2)	150378463	264	311
3O7A	PHF13	PHF13	167466270	234	279
3KQI	PHF2	PHF2	117190342	6	55
1WEQ	PHF7	PHF7(2)	21361543	252	301
3KV4	PHF8	PHF8	32698700	6	54
2VPG	PYGO1	PYGO1	30911103	342	397
2XB1	PYGO2	PYGO2	23510333	330	383
2V86	RAG2	RAG2	151301080	416	482
2K16	TAF3	TAF3	151301171	864	914
3O36	TRIM24	TRIM24	47419909	791	837
1FP0	TRIM28	TRIM28	5032179	625	670
3ASK	UHRF1	UHRF1	115430233	330	378
2E6R	JARID1D	JARID1D(1)	33356560	315	361
2YT5	MTF2	MTF2(1)	6678764	103	155
2KYU	MLL	MLL(3)	56550039	1570	1626
2LV9	MLL5	MLL5	33636768	119	164

PDB ID	Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
2YQL	PHF21A	PHF21A	156546894	487	534
2E6S	UHRF2	UHRF2	23312364	344	393

Appendix 2.2 - PHDs Identified by HMMER Search

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
AIRE	AIRE(1)	4557291	295	342
ASH1L	ASH1L	110349788	2577	2629
ATRX	ATRX	20336205	180	238
BAZ1A	BAZ1A	32967603	1149	1196
BAZ1B	BAZ1B	14670392	1184	1232
BAZ2A	BAZ2A	91176325	1676	1724
BAZ2B	BAZ2B	94681063	1931	1979
BRD1	BRD1	11321642	213	263
BRPF1	BRPF1	51173720	272	322
BRPF3	BRPF3	148727368	211	261
CHD3	CHD3(1)	52630322	378	425
CHD3	CHD3(2)	52630322	455	501
CHD4	CHD4(1)	51599156	369	416
CHD4	CHD4(2)	51599156	448	494
CHD5	CHD5(1)	24308089	342	389
CHD5	CHD5(2)	24308089	415	461
CXXC1	CXXC1	156142180	27	75
DIDO1	DIDO1	18375617	268	320
DNMT3A	DNMT3A	12751473	528	587
DNMT3B	DNMT3B	5901940	466	536
DNMT3L03	DNMT3L03	28872778	93	150
DPF1	DPF1(1)	205830430	299	353
DPF1	DPF1(2)	205830430	354	401
DPF1B	DPF1B(2)	205830432	310	367
DPF2	DPF2(1)	5454004	270	328
DPF2	DPF2(2)	5454004	327	375
DPF3B	DPF3B(1)	60459281	260	317
DPF3B	DPF3B(2)	60459281	318	356
FALZ	FALZ(1)	38788260	390	435
FALZ	FALZ(2)	38788260	2724	2773
FBXL10	FBXL10	54112380	627	693
FBXL11	FBXL11	16306580	616	677
FBXL19	FBXL19	157168349	84	150
ING1	ING1	38201661	352	400
ING2	ING2	4504695	211	259
ING3	ING3	38201655	359	407

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
ING4	ING4	38201670	194	243
ING5	ING5	18644730	185	234
INTS12	INTS12	21361851	161	214
JARID1A	JARID1A(1)	110618244	294	341
JARID1A	JARID1A(2)	110618244	1161	1217
JARID1A	JARID1A(3)	110618244	1606	1660
JARID1B	JARID1B(1)	57242796	310	357
JARID1B	JARID1B(2)	57242796	1176	1223
JARID1B	JARID1B(3)	57242796	1483	1536
JARID1C	JARID1C(1)	109255243	324	372
JARID1C	JARID1C(2)	109255243	1185	1249
JARID1D	JARID1D(1)	33356560	315	361
JARID1D	JARID1D(2)	33356560	1172	1236
JMJD2A	JMJD2A(1)	98986459	704	783
JMJD2A	JMJD2A(2)	98986459	826	888
JMJD2B	JMJD2B(1)	45504380	726	805
JMJD2B	JMJD2B(2)	45504380	847	910
JMJD2C	JMJD2C(1)	109255247	684	763
JMJD2C	JMJD2C(2)	109255247	806	868
KIAA1718	KIAA1718	90093355	38	86
LOC93349	LOC93349	134133279	404	447
MLL2	MLL2(1)	148762969	226	274
MLL2	MLL2(2)	148762969	275	321
MLL2	MLL2(3)	148762969	1377	1428
MLL2	MLL2(4)	148762969	1429	1475
MLL2	MLL2(5)	148762969	1501	1562
MLL3	MLL3(1)	91718902	340	389
MLL3	MLL3(2)	91718902	390	436
MLL3	MLL3(3)	91718902	957	1008
MLL3	MLL3(4)	91718902	1007	1055
MLL3	MLL3(5)	91718902	1081	1142
MLL4	MLL4(1)	7662046	1200	1250
MLL4	MLL4(2)	7662046	1251	1302
MLL4	MLL4(3)	7662046	1338	1394
MLL5	MLL5	33636768	119	164
MLL	MLL(1)	56550039	1428	1484
MLL	MLL(2)	56550039	1481	1532
MLL	MLL(3)	56550039	1570	1626
MLLT10	MLLT10	4757726	24	73
MLLT6	MLLT6	57222568	8	56
MTF2	MTF2	6678764	103	155
MYST3	MYST3(1)	150378463	206	263
MYST3	MYST3(2)	150378463	264	311

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
MYST4	MYST4(1)	100816397	213	270
MYST4	MYST4(2)	100816397	271	318
NSD1	NSD1(1)	19923586	1545	1597
NSD1	NSD1(2)	19923586	1707	1749
NSD1	NSD1(3)	19923586	2118	2163
PHF10	PHF10(1)	194328734	377	434
PHF10	PHF10(2)	194328734	435	480
PHF11	PHF11	94681065	109	158
PHF12	PHF12(1)	30842829	56	103
PHF12	PHF12(2)	30842829	272	319
PHF13	PHF13	167466270	234	279
PHF14	PHF14(1)	55769548	318	378
PHF14	PHF14(2)	55769548	727	778
PHF15	PHF15	40556370	198	247
PHF16	PHF16	7662006	199	248
PHF17	PHF17	19923609	202	251
PHF19	PHF19	58331161	97	148
PHF1	PHF1	4505777	88	140
PHF20	PHF20	18034775	654	698
PHF20L1	PHF20L1	111120331	681	727
PHF21A	PHF21A	156546894	487	534
PHF21B	PHF21B	19923937	351	398
PHF23	PHF23	116268091	341	385
PHF2	PHF2	117190342	6	55
PHF3	PHF3	7662018	718	770
PHF8	PHF8	32698700	6	54
PHRF1	PHRF1	221139764	183	231
PRKCBP1	PRKCBP1	34335262	108	152
PYGO1	PYGO1	30911103	342	397
PYGO2	PYGO2	23510333	330	383
RAG2	RAG2	151301080	416	482
RSF1	RSF1	38788333	891	940
SHPRH	SHPRH	27436873	661	707
SP100	SP100	122939208	701	746
SP110	SP110	190343006	533	579
SP140	SP140	217330601	691	734
TAF3	TAF3	151301171	864	914
TCF19	TCF19	117414152	295	340
TIF1	TIF1	47419909	791	837
TRIM28	TRIM28	5032179	625	670
TRIM33	TRIM33	74027249	886	932
TRIM66	TRIM66	209977097	969	1015
UBR7	UBR7	154426322	134	188

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
UHRF1	UHRF1	115430233	330	378
UHRF2	UHRF2	23312364	344	393
WHSC1	WHSC1(1)	19913348	666	711
WHSC1	WHSC1(2)	19913348	831	873
WHSC1	WHSC1(3)	19913348	1239	1284
WHSC1L1	WHSC1L1(1)	13699811	700	746
WHSC1L1	WHSC1L1(2)	13699811	1320	1366
ZMYND11	ZMYND11	238814385	98	146

Appendix 2.3 - Complete List of PHD Sequences

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
AIRE	AIRE(1)	4557291	295	342
AIRE	AIRE(2)	4557291	434	475
ASH1L	ASH1L	110349788	2577	2629
ASH2L	ASH2L	157412280	103	160
ASXL1	ASXL1	29570782	1503	1540
ASXL2	ASXL2	153792780	1397	1434
ASXL3	ASXL3	149944526	2210	2247
ATRX	ATRX	20336205	180	238
BAZ1A	BAZ1A	32967603	1149	1196
BAZ1B	BAZ1B	14670392	1184	1232
BAZ2A	BAZ2A	91176325	1676	1724
BAZ2B	BAZ2B	94681063	1931	1979
BRD1	BRD1(1)	11321642	213	263
BRD1	BRD1(2)	11321642	326	388
BRPF1	BRPF1(1)	51173720	272	322
BRPF1	BRPF1(2)	51173720	385	447
BRPF3	BRPF3(1)	148727368	211	261
BRPF3	BRPF3(2)	148727368	324	386
CHD3	CHD3(1)	52630322	378	425
CHD3	CHD3(2)	52630322	455	501
CHD4	CHD4(1)	51599156	369	416
CHD4	CHD4(2)	51599156	448	494
CHD5	CHD5(1)	24308089	342	389
CHD5	CHD5(2)	24308089	415	461
CREBBP	CREBBP	119943104	1235	1314
CXXC1	CXXC1	156142180	27	75
DIDO1	DIDO1	18375617	268	320
DNMT3A	DNMT3A	12751473	528	587
DNMT3B	DNMT3B	5901940	466	536
DNMT3L03	DNMT3L03	28872778	93	150
DPF1	DPF1(1)	205830430	299	353

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
DPF1A	DPF1(2)	205830430	354	401
DPF1B	DPF1B(2)	205830432	310	367
DPF2	DPF2(1)	5454004	270	328
DPF2	DPF2(2)	5454004	327	375
DPF3B	DPF3B(1)	60459281	260	317
DPF3B	DPF3B(2)	60459281	318	356
EP300	EP300	50345997	1198	1278
FALZ	FALZ(1)	38788260	390	435
FALZ	FALZ(2)	38788260	2724	2773
FBXL10	FBXL10	54112380	627	693
FBXL11	FBXL11	16306580	616	677
FBXL19	FBXL19	157168349	84	150
ING1	ING1	38201661	352	400
ING2	ING2	4504695	211	259
ING3	ING3	38201655	359	407
ING4	ING4	38201670	194	243
ING5	ING5	18644730	185	234
INTS12	INTS12	21361851	161	214
JARID1A	JARID1A(1)	110618244	294	341
JARID1A	JARID1A(2)	110618244	1161	1217
JARID1A	JARID1A(3)	110618244	1606	1660
JARID1B	JARID1B(1)	57242796	310	357
JARID1B	JARID1B(2)	57242796	1176	1223
JARID1B	JARID1B(3)	57242796	1483	1536
JARID1C	JARID1C(1)	109255243	324	372
JARID1C	JARID1C(2)	109255243	1185	1249
JARID1D	JARID1D(1)	33356560	315	361
JARID1D	JARID1D(2)	33356560	1172	1236
JMJD2A	JMJD2A(1)	98986459	704	783
JMJD2A	JMJD2A(2)	98986459	826	888
JMJD2B	JMJD2B(1)	45504380	726	805
JMJD2B	JMJD2B(2)	45504380	847	910
JMJD2C	JMJD2C(1)	109255247	684	763
JMJD2C	JMJD2C(2)	109255247	806	868
KIAA1045	KIAA1045	149944593	129	190
KIAA1333	KIAA1333(1)	33620749	78	132
KIAA1333	KIAA1333(2)	33620749	229	290
KIAA1718B	KIAA1718B	90093355	38	86
LOC93349	LOC93349	134133279	404	447
MLL2B	MLL2B(1)	148762969	169	222
MLL2B	MLL2B(2)	148762969	226	274
MLL2B	MLL2B(3)	148762969	275	321
MLL2B	MLL2B(4)	148762969	1377	1428

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
MLL2B	MLL2B(5)	148762969	1429	1475
MLL2B	MLL2B(6)	148762969	1501	1562
MLL2B	MLL2B(7)	148762969	5089	5141
MLL3B	MLL3B(1)	91718902	284	331
MLL3B	MLL3B(2)	91718902	340	389
MLL3B	MLL3B(3)	91718902	390	436
MLL3B	MLL3B(4)	91718902	466	518
MLL3B	MLL3B(5)	91718902	957	1008
MLL3B	MLL3B(6)	91718902	1007	1055
MLL3B	MLL3B(7)	91718902	1081	1142
MLL3B	MLL3B(8)	91718902	4461	4507
MLL4	MLL4(1)	7662046	1200	1250
MLL4	MLL4(2)	7662046	1251	1302
MLL4	MLL4(3)	7662046	1338	1394
MLL4	MLL4(4)	7662046	1638	1689
MLL5	MLL5	33636768	119	164
MLL	MLL(1)	56550039	1428	1484
MLL	MLL(2)	56550039	1481	1532
MLL	MLL(3)	56550039	1570	1626
MLL	MLL(4)	56550039	1932	1978
MLLT10	MLLT10(1)	4757726	24	73
MLLT10	MLLT10(2)	4757726	134	210
MLLT6	MLLT6(1)	57222568	8	56
MLLT6	MLLT6(2)	57222568	125	185
MTF2	MTF2(1)	6678764	103	155
MTF2	MTF2(2)	6678764	203	253
MYST3	MYST3(1)	150378463	206	263
MYST3	MYST3(2)	150378463	264	311
MYST4	MYST4(1)	100816397	213	270
MYST4	MYST4(2)	100816397	271	318
NSD1	NSD1(1)	19923586	1545	1597
NSD1	NSD1(2)	19923586	1592	1639
NSD1	NSD1(3)	19923586	1640	1693
NSD1	NSD1(4)	19923586	1707	1749
NSD1	NSD1(5)	19923586	2118	2163
PHF10	PHF10(1)	194328734	377	434
PHF10	PHF10(2)	194328734	435	480
PHF11	PHF11	94681065	109	158
PHF12	PHF12(1)	30842829	56	103
PHF12	PHF12(2)	30842829	272	319
PHF13	PHF13	167466270	234	279
PHF14	PHF14(1)	55769548	318	378
PHF14	PHF14(2)	55769548	439	502

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
PHF14	PHF14(3)	55769548	727	778
PHF14	PHF14(4)	55769548	870	919
PHF15	PHF15(1)	40556370	198	247
PHF15	PHF15(2)	40556370	310	364
PHF16	PHF16(1)	7662006	199	248
PHF16	PHF16(2)	7662006	306	370
PHF17	PHF17(1)	19923609	202	251
PHF17	PHF17(2)	19923609	317	373
PHF19	PHF19(1)	58331161	97	148
PHF19	PHF19(2)	58331161	197	247
PHF1	PHF1(1)	4505777	88	140
PHF1	PHF1(2)	4505777	188	238
PHF20	PHF20	18034775	654	698
PHF20L1	PHF20L1	111120331	681	727
PHF21A	PHF21A	156546894	487	534
PHF21B	PHF21B	19923937	351	398
PHF23	PHF23	116268091	341	385
PHF2	PHF2	117190342	6	55
PHF3	PHF3	7662018	718	770
PHF6	PHF6(1)	28557677	81	132
PHF6	PHF6(2)	28557677	279	330
PHF7	PHF7(1)	21361543	97	145
PHF7	PHF7(2)	21361543	252	301
PHF8	PHF8	32698700	6	54
PHRF1	PHRF1	221139764	183	231
PRKCBP1	PRKCBP1	34335262	108	152
PYGO1	PYGO1	30911103	342	397
PYGO2	PYGO2	23510333	330	383
RAG2	RAG2	151301080	416	482
RAI1	RAI1	40807477	1856	1903
RSF1	RSF1	38788333	891	940
SHPRH	SHPRH	27436873	661	707
SP100	SP100	122939208	701	746
SP110	SP110	190343006	533	579
SP140	SP140	217330601	691	734
TAF3	TAF3	151301171	864	914
TCF19	TCF19	117414152	295	340
TCF20	TCF20	31652244	1886	1933
TIF1	TIF1	47419909	791	837
TRIM28	TRIM28	5032179	625	670
TRIM33	TRIM33	74027249	886	932
TRIM66	TRIM66(1)	209977097	4	69
TRIM66	TRIM66(2)	209977097	969	1015

Gene ID	PHD Name	Protein GI	Sequence Start	Sequence End
UBR7	UBR7	154426322	134	188
UHRF1	UHRF1	115430233	330	378
UHRF2	UHRF2	23312364	344	393
WHSC1	WHSC1(1)	19913348	666	711
WHSC1	WHSC1(2)	19913348	717	762
WHSC1	WHSC1(3)	19913348	761	821
WHSC1	WHSC1(4)	19913348	831	877
WHSC1	WHSC1(5)	19913348	1239	1284
WHSC1L1	WHSC1L1(1)	13699811	700	746
WHSC1L1	WHSC1L1(2)	13699811	751	798
WHSC1L1	WHSC1L1(3)	13699811	799	851
WHSC1L1	WHSC1L1(4)	13699811	1320	1366
ZMYND11	ZMYND11	238814385	98	146

Appendix 2.4 - RING Domains

Gene ID	RING Name	Protein GI	Sequence Start	Sequence End
AMFR	AMFR	21071001	339	380
BFAR	BFAR	7706091	30	76
BRAP	BRAP	1.88E+08	263	305
BRCA1B	BRCA1B	2.38E+08	15	68
CBL	CBL	52426745	380	423
CBLB	CBLB	54112420	372	415
CBLCB	CBLCB	1.96E+08	304	347
CHFRB	CHFRB	2.39E+08	255	302
COMMD3BMI1	COMMD3BMI1	3.23E+08	153	202
DTX1	DTX1	41352718	408	473
DTX2	DTX2	1.57E+08	409	474
DTX3	DTX3	30425428	159	203
DTX3L	DTX3L	19923717	557	602
DTX4	DTX4	1.48E+08	406	469
DZIP3	DZIP3	7662244	1143	1189
LNX1B	LNX1B	1.88E+08	34	80
LNX2	LNX2	24025688	42	90
LOC120824	LOC120824	3.31E+08	8	57
LOC283116	LOC283116	3.31E+08	8	58
LOC642446	LOC642446	2.57E+08	12	58
LOC646754	LOC646754	3.31E+08	9	58
LOC646862	LOC646862	3.04E+08	60	113
LOC653192	LOC653192	2.57E+08	12	58
LOC729384	LOC729384	1.58E+08	12	57
LOC93312	LOC93312	1.48E+08	468	509
LONFR3	LONFR3A(1)	37622896	154	197

Gene ID	RING Name	Protein GI	Sequence Start	Sequence End
LONFR3	LONFR3A(2)	37622896	423	465
LONRF1	LONRF1A(1)	87080813	121	160
LONRF1	LONRF1A(2)	87080813	476	519
LONRF2	LONRF2	1.49E+08	446	489
LTN1	LTN1	2.32E+08	1759	1810
MEX3B	MEX3B	47716512	517	560
MEX3C	MEX3C	1.48E+08	605	650
MGRN1	MGRN1	3.35E+08	276	319
MID1B	MID1B	15451852	2	61
MID2	MID2	2.24E+08	22	80
MKRN1	MKRN1	2.23E+08	277	336
MKRN2	MKRN2	32880199	235	292
MKRN3	MKRN3	5032243	309	366
MYCBP2	MYCBP2	2.91E+08	4425	4481
NHLRC1	NHLRC1	40255283	24	73
OR5R1	OR5R1	38045931	91	137
PCGF1	PCGF1	1.09E+08	39	88
PCGF2	PCGF2	6005964	10	59
PCGF3	PCGF3	31742478	11	57
PCGF5	PCGF5	33300663	12	59
PCGF6	PCGF6	37655165	127	173
PDZRN3	PDZRN3	57529737	12	54
PDZRN4B	PDZRN4B	2.57E+08	10	54
PEX10	PEX10	4505715	270	313
PHRF1	PHRF1	2.21E+08	106	150
PJA1	PJA1	41281725	592	638
PJA2	PJA2	1.57E+08	631	677
PXMP3	PXMP3	1.21E+08	242	286
RAD18	RAD18	2.57E+08	17	64
RAG1	RAG1	4557841	292	334
RAPSN	RAPSN	15619013	360	404
RBBP6B	RBBP6B	33620716	250	301
RBCK1	RBCK1	1.45E+08	279	326
RC3H1	RC3H1	73695473	9	55
RC3H2	RC3H2	1.56E+08	9	55
RCHY1	RCHY1	58331195	143	188
RFPL4B	RFPL4B	1.54E+08	3	54
RFWD2B	RFWD2B	21359963	132	174
RFWD3	RFWD3	71143112	285	333
RING1	RING1	51479192	42	90
RLIM2	RLIM2	34452684	569	613
RNF103	RNF103	3.12E+08	616	660
RNF10	RNF10	34452681	224	268

Gene ID	RING Name	Protein GI	Sequence Start	Sequence End
RNF111	RNF111	3.95E+08	940	986
RNF113	RNF113	5902158	261	302
RNF113B	RNF113B	30578416	255	296
RNF114	RNF114	8923898	27	69
RNF115	RNF115	33859668	221	271
RNF11	RNF11	7657520	98	143
RNF121	RNF121	21361732	223	279
RNF125	RNF125	37595555	34	77
RNF126	RNF126	37622894	223	272
RNF128B	RNF128B	37588873	275	320
RNF130	RNF130	29788758	263	306
RNF133	RNF133	21040269	254	299
RNF135C	RNF135C	2.97E+08	16	64
RNF138	RNF138	21361539	13	59
RNF139	RNF139	21314654	545	589
RNF13	RNF13	6005864	238	285
RNF141	RNF141	21361493	146	194
RNF145	RNF145	3.14E+08	552	593
RNF146	RNF146	33636758	35	77
RNF148	RNF148	37675277	255	301
RNF149	RNF149	2.84E+08	267	311
RNF150	RNF150	58331204	276	321
RNF151	RNF151	87241872	13	60
RNF152	RNF152	27734873	9	55
RNF157	RNF157	58743365	275	318
RNF165	RNF165	3.78E+08	99	146
RNF166	RNF166	30520320	27	73
RNF167	RNF167	14149702	228	275
RNF168	RNF168	31377566	11	57
RNF169	RNF169	1.49E+08	65	108
RNF170	RNF170	2.38E+08	84	133
RNF175	RNF175	27734859	224	280
RNF180	RNF180	1.66E+08	427	476
RNF181	RNF181	7706039	70	118
RNF182	RNF182	2.59E+08	14	68
RNF183	RNF183	1.53E+08	10	62
RNF185	RNF185	31542783	35	81
RNF186	RNF186	9506663	37	87
RNF187	RNF187	2.56E+08	8	54
RNF208	RNF208	1.19E+08	139	192
RNF213B	RNF213B	3.66E+08	3996	4036
RNF215	RNF215	63025220	323	367
RNF219	RNF219	88759348	17	58

Gene ID	RING Name	Protein GI	Sequence Start	Sequence End
RNF220	RNF220	46397375	512	554
RNF222	RNF222	2.26E+08	12	66
RNF223	RNF223	3.27E+08	44	103
RNF224	RNF224	2.98E+08	21	73
RNF24C	RNF24C	6857791	76	121
RNF2	RNF2	6005747	45	93
RNF32	RNF32	13569903	290	354
RNF38	RNF38	37577185	377	422
RNF39	RNF39	2.97E+08	80	134
RNF41	RNF41	37588861	12	58
RNF43	RNF43	56711322	271	313
RNF44	RNF44	7662486	375	422
RNF4	RNF4	2.97E+08	128	180
RNF5	RNF5	5902054	24	69
RNF6	RNF6	5174653	631	676
RNF7	RNF7	3.19E+08	48	89
RNF8	RNF8	4504867	395	444
RNFT1	RNFT1	1.09E+08	372	415
SCAF11	SCAF11	1.18E+08	39	84
SH3MD4	SH3MD4	1.5E+08	52	98
SH3RF1	SH3RF1	51988887	6	54
SH3RF2	SH3RF2	2.22E+08	6	54
SMARCA3B	SMARCA3B	21071052	757	804
SPRYD5	SPRYD5	2.1E+08	9	57
SYVN1	SYVN1	27436927	289	332
TMEM118B	TMEM118B	1.58E+08	382	423
TOPORS	TOPORS	3.07E+08	35	79
TRAF2	TRAF2	22027612	26	73
TRAF3	TRAF3	22027616	43	95
TRAF4	TRAF4	22027622	9	58
TRAF5	TRAF5	11321603	35	86
TRAF6	TRAF6	22027630	61	111
TRAIP	TRAIP	40807469	6	53
TRIM10	TRIM10	1.57E+08	8	62
TRIM11	TRIM11	21630277	8	58
TRIM13	TRIM13	55953112	6	63
TRIM15	TRIM15	1.49E+08	15	62
TRIM17C	TRIM17C	1.98E+08	8	67
TRIM21	TRIM21	15208660	9	57
TRIM22	TRIM22	3.14E+08	8	61
TRIM23	TRIM23	4502197	26	77
TRIM25	TRIM25	68160937	6	55
TRIM26B	TRIM26B	3.39E+08	9	58

Gene ID	RING Name	Protein GI	Sequence Start	Sequence End
TRIM27	TRIM27	5730009	9	58
TRIM2	TRIM2	1.94E+08	48	93
TRIM31	TRIM31	62865604	6	59
TRIM32B	TRIM32B	1.54E+08	11	67
TRIM35	TRIM35	94536782	15	62
TRIM37	TRIM37	15147333	14	57
TRIM38	TRIM38	5454014	8	64
TRIM39	TRIM39	25777696	20	71
TRIM3	TRIM3	32454739	20	65
TRIM40	TRIM40	20162564	9	58
TRIM43	TRIM43	20270353	8	57
TRIM47	TRIM47	54792146	8	60
TRIM48	TRIM48	2.02E+08	28	74
TRIM49	TRIM49	9966829	8	58
TRIM4	TRIM4	15011941	4	54
TRIM50	TRIM50	30023818	9	59
TRIM54	TRIM54	78482626	18	84
TRIM55	TRIM55	34878852	18	84
TRIM56	TRIM56	30794216	19	63
TRIM58	TRIM58	1.12E+08	9	61
TRIM59	TRIM59	27436877	3	61
TRIM5	TRIM5	2.83E+08	7	61
TRIM60	TRIM60	22749269	8	57
TRIM61	TRIM61	60099474	9	57
TRIM62	TRIM62	2.17E+08	5	55
TRIM63	TRIM63	19924163	20	82
TRIM64	TRIM64	2.1E+08	8	58
TRIM65	TRIM65	3.71E+08	6	52
TRIM68	TRIM68	37622899	8	62
TRIM69	TRIM69	88999601	33	82
TRIM6	TRIM6	51477692	36	90
TRIM6TRIM34	TRIM6TRIM34	51477690	362	416
TRIM72	TRIM72	2.7E+08	8	58
TRIM73	TRIM73	65285121	9	57
TRIM74	TRIM74	38524612	9	57
TRIM77	TRIM77	2.26E+08	10	57
TRIM7	TRIM7	16076875	26	84
TRIM8	TRIM8	1.49E+08	9	58
TRIML1	TRIML1	31542779	8	57
TTC3B	TTC3B	49640009	1956	1997
UHRF1C	UHRF1C	1.15E+08	715	765
UHRF2B	UHRF2B	23312364	723	774
VPS11	VPS11	17978477	819	860

Gene ID	RING Name	Protein GI	Sequence Start	Sequence End
ZNF179	ZNF179	2.65E+08	56	99
ZNRF1	ZNRF1	14150005	183	222
ZNRF2	ZNRF2	23821044	198	237
ZNRF3B	ZNRF3B	1.5E+08	191	236
ZNRF4	ZNRF4	1.5E+08	307	354
ZSWIM2	ZSWIM2A(2)	71043932	340	386
ZSWIM2	ZSWIM2A(1)	71043932	139	200

Appendices for Chapter 3 - Computational Assessment of the Ligandability of PHDs and Other Epigenetic Reader Domains

Appendix 3.1 - PHD Structures Used

Gene ID	PDB ID
AIRE(1)	1XWH
AIRE(1)	2KE1
AIRE(2)	2KFT
ATRX	2JM1
ATRX	2LBM
ATRX	2LD1
ATRX	3QL9
ATRX	3QLA
ATRX	3QLC
ATRX	3QLN
BAZ1B	1F62
BRD1	2KU3
BRD1	2L43
CHD4(1)	2L5U
CHD4(2)	1MM2
CHD4(2)	2L75
DIDO1	1WEM
DNMT3A	3A1A
DNMT3A	3A1B
DNMT3L03	2PV0
DNMT3L03	2PVC
DPF3B(1)	2KWJ
DPF3B(1)	2KWK
DPF3B(1)	2KWN
DPF3B(1)	2KWO
FALZ(2)	2F6J
FALZ(2)	2F6N
FALZ(2)	2FSA
FALZ(2)	2FUI
FALZ(2)	2FUU
FALZ(2)	2RI7
FALZ(2)	3QZV
ING1	2QIC
ING2	1WES
ING2	2G6Q
ING3	1X4I
ING4	1WEN

Gene ID	PDB ID
ING4	1WEU
ING4	2K1J
ING4	2PNX
ING4	2VNF
ING5	3C6W
JARID1A(3)	2KGG
JARID1A(3)	2KGI
JARID1A(3)	3GL6
JARID1D(1)	2E6R
KIAA1718	3KV5
KIAA1718	3KV6
MLL3(1)	2YSM
MLL5	2LV9
MLL(3)	2KU7
MLL(3)	2KYU
MLL(3)	3LQH
MLL(3)	3LQI
MLL(3)	3LQJ
MTF2A	2YT5
MYST3(1)	2LN0
MYST3(1)	3V43
PHF13	3O70
PHF13	3O7A
PHF21A	2PUY
PHF21A	2YQL
PHF2	3KQI
PHF8	3KV4
PYGO1	2DX8
PYGO1	2VP7
PYGO1	2VPB
PYGO1	2VPD
PYGO1	2VPE
PYGO1	2VPG
PYGO1	2YYR
PYGO2	2XB1
RAG2	2JWO
RAG2	2V83
RAG2	2V85
RAG2	2V86
RAG2	2V87
RAG2	2V88
RAG2	2V89
TAF3	2K16

Gene ID	PDB ID
TAF3	2K17
TIF1A	3O33
TIF1A	3O35
TIF1A	3O36
TIF1A	3O37
TRIM28	1FP0
TRIM28	2RO1
TRIM33	3U5M
TRIM33	3U5N
TRIM33	3U5O
TRIM33	3U5P
UHRF1	2LGG
UHRF1	2LGK
UHRF1	2LGL
UHRF1	3ASK
UHRF1	3ASL
UHRF1	3SHB
UHRF1	3SOU
UHRF1	3SOW
UHRF1	3SOX
UHRF1	3T6R
UHRF1	3ZVY
UHRF1	4GY5

Appendix 3.2 - Tudor Domain Structures Used

Gene ID	PDB ID
CCDC101	3MEW
CCDC101	3MEA
CCDC101	3MET
CCDC101	3MEU
CCDC101	3ME9
CCDC101	3LX7
JMJC2C(1&2)	2XDP
JMJD2A(1&2)	2QQS
JMJD2A(1&2)	2QQR
JMJD2A(1&2)	2GFA
JMJD2A(1&2)	2GF7
MTF2	2EQJ
PHF19	4BD3
PHF1	2M0O
PHF1	4HCZ
PHF1	2E5P
PHF20(1)	3SD4
PHF20(1)	3Q1J
RNF17(2)	2EQK
SETDB1(1&2)	3DLM
SMN1	4A4E
SMN1	4A4G
SMN1	1MHN
SMN1	1G5V
SMNDC1	4A4F
SMNDC1	4A4H
TDRD3	3PMT
TDRD3	3PNW
TDRD3	2LTO
TDRD3	3S6W
TDRD3	2D9T
TDRKH	2DIQ
TDRKH	3FDR
TP53BP1(1&2)	2G3R
TP53BP1(1&2)	2IG0
TP53BP1(1&2)	3LGL
TP53BP1(1&2)	3L1D
TP53BP1(1&2)	2LVM
TP53BP1(1&2)	1XNI
TP53BP1(1&2)	3LH0
TP53BP1(1&2)	1SSF
UHRF1(1&2)	4GY5

Gene ID	**PDB ID**
UHRF1(1&2)	3ASK
UHRF1(1&2)	2L3R
UHRF1(1&2)	3DB3
UHRF1(1&2)	3DB4
ZGPAT	4II1
SPIN1	2NS2
SPIN1	4H75

Appendix 3.3 - Multi-domain Structures Used

Gene Name	Domain Architecture	PDB ID
ASH2L	PHD-Set1	3RSN
CCDC101	Tudor-Tudor	3MEW
CCDC101	Tudor-Tudor	3MEA
CCDC101	Tudor-Tudor	3MET
CCDC101	Tudor-Tudor	3MEU
CCDC101	Tudor-Tudor	3ME9
CCDC101	Tudor-Tudor	3LX7
CHD1	Chromo-Chromo	2B2T
CHD1	Chromo-Chromo	2B2U
CHD1	Chromo-Chromo	2B2V
CHD1	Chromo-Chromo	2B2W
CHD1	Chromo-Chromo	2B2Y
CHD1	Chromo-Chromo	4NW2
CHD1	Chromo-Chromo	4O42
CREBBP	Bromo-PHD	4N4F
DPF3B	PHD-PHD	2KWJ
DPF3B	PHD-PHD	2KWK
DPF3B	PHD-PHD	2KWN
DPF3B	PHD-PHD	2KWO
EP300	Bromo-RING-PHD-HAT	4BHW
FALZ	PHD-Bromo	2F6J
FALZ	PHD-Bromo	2F6N
FALZ	PHD-Bromo	2FSA
FALZ	PHD-Bromo	2RI7
FALZ	PHD-Bromo	3QZV
JMJC2A	Tudor-Tudor	2QQS
JMJC2A	Tudor-Tudor	2QQR
JMJC2A	Tudor-Tudor	2GFA
JMJC2A	Tudor-Tudor	2GF7
JMJC2C	Tudor-Tudor	2XDP
KIAA1718	PHD-JmjC	3KV5
KIAA1718	PHD-JmjC	3KV6
L3MBTL	MBT-MBT	3OQ5
L3MBTL	MBT-MBT	1OYX
L3MBTL	MBT-MBT	1OZ2
L3MBTL	MBT-MBT	1OZ3
L3MBTL	MBT-MBT	2PQW
L3MBTL	MBT-MBT	2RHI
L3MBTL	MBT-MBT	2RHX
L3MBTL	MBT-MBT	2RJC
L3MBTL	MBT-MBT	2RJD
L3MBTL	MBT-MBT	2RJE

Gene Name	Domain Architecture	PDB ID
L3MBTL	MBT-MBT	2RJF
L3MBTL	MBT-MBT	3P8H
L3MBTL	MBT-MBT	3UWN
L3MBTL	MBT-MBT	2RHU
L3MBTL	MBT-MBT	2RHY
L3MBTL	MBT-MBT	2RHZ
L3MBTL	MBT-MBT	2RI2
L3MBTL	MBT-MBT	2RI3
L3MBTL	MBT-MBT	2RI5
L3MBTL2	MBT-MBT	3CEY
L3MBTL2	MBT-MBT	3F70
L3MBTL3	MBT-MBT	3UT1
L3MBTL3	MBT-MBT	4FL6
L3MBTL3	MBT-MBT	4L59
MBTD1	MBT-MBT	3FEO
MBTD1	MBT-MBT	4C5I
MLL3B	PHD-PHD	2YSM
MLL	PHD-Bromo	3LQH
MLL	PHD-Bromo	3LQI
MLL	PHD-Bromo	3LQJ
MYST3	PHD-PHD	2LN0
MYST3	PHD-PHD	3V43
MYST3	PHD-PHD	4LKA
MYST3	PHD-PHD	4LK9
MYST3	PHD-PHD	4LJN
MYST3	PHD-PHD	4LLB
PHF8A	PHD-JmjC	3KV4
PRKCBP1	PHD-Bromo-PWWP	4COS
SCMH1	MBT-MBT	2P0K
SCMH2	MBT-MBT	2BIV
SCMH2	MBT-MBT	1OI1
SCMH2	MBT-MBT	2VYT
SCMH2	MBT-MBT	4EDU
SETDB1	Tudor-Tudor	3DLM
SP100	PHD-Bromo	4PTB
SPIN1	Tudor-Tudor	2NS2
SPIN1	Tudor-Tudor	4H75
TAF1	Bromo-Bromo	1EQF
TIF1A	PHD-Bromo	3O33
TIF1A	PHD-Bromo	3O34
TIF1A	PHD-Bromo	3O35
TIF1A	PHD-Bromo	3O36
TIF1A	PHD-Bromo	3O37

Gene Name	Domain Architecture	PDB ID
TP53BP1	Tudor-Tudor	2G3R
TP53BP1	Tudor-Tudor	2IG0
TP53BP	Tudor-Tudor	3LGL
TP53BP1	Tudor-Tudor	2LVM
TRIM28	PHD-Bromo	2RO1
TRIM33	PHD-Bromo	3U5M
TRIM33	PHD-Bromo	3U5N
TRIM33	PHD-Bromo	3U5O
TRIM33	PHD-Bromo	3U5P
UHRF1	Tandem-tudor-PHD	3ASK
UHRF1	Tandem-tudor-PHD	4GY5
WHSC1L1	PHD-PHD-like	4GND
WHSC1L1	PHD-PHD-like	4GNE
WHSC1L1	PHD-PHD-like	4GNF
WHSC1L1	PHD-PHD-like	4GNG
ZMYND11	Bromo-PWWP	4N4G
ZMYND11	Bromo-PWWP	4N4H
ZMYND11	Bromo-PWWP	4N4I

Appendices for Chapter 4 - Assessing the ligandability of tandem PHDs

Appendix 4.1 – BLI Screening of Peptide Partners for DPF2

Key

Acetyl lysine	Δ
Monomethyl lysine	Φ
Dimethyl lysine	Π
Trimethyl lysine	Θ
Phospho serine	Σ
Phospho threonine	Ω
Monomethyl arginine	Ξ
Asymmetric dimethyl arginine	Ψ
Aminohexanoic acid	Spacer
N-terminal acetylation	Ac-

Histone	Residues	Sequence	Shift in wavelength (nm)
		Control 1	0.0687
Histone H3	1-21	ARTKQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.7539
		AΞTKQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.6508
		AΨTKQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.7149
		AΨΩKQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.1617
		AΨΩΘQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.2558
		AΨΤΘQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.4008
		ARΩKQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.3785
		ARΩΘQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.2768
		ARTΦQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.7777
		ARTΠQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.7638
		ARTΘQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.5928
		ARTΘQTARΔSTGGKAPRKQLA *-spacer-Biotin*	0.6731
		ARTΘQTARΘSTGGKAPRKQLA *-spacer-Biotin*	0.4402
		ARTKQTARΔSTGGKAPRKQLA *-spacer-Biotin*	0.0052
		ARTKQTARΔΣTGGKAPRKQLA *-spacer-Biotin*	0.6757
		ARTKQTARΔSΩGGKAPRKQLA *-spacer-Biotin*	0.9845
		ARTKQTARΔΣΩGGKAPRKQLA *-spacer-Biotin*	0.6369
		ARTKQTARΦSTGGKAPRKQLA *-spacer-Biotin*	0.905
		ARTKQTARΠSTGGKAPRKQLA *-spacer-Biotin*	0.767
		ARTKQTARΘSTGGKAPRKQLA *-spacer-Biotin*	0.624
		ARTKQTARΘSΩGGKAPRKQLA *-spacer-Biotin*	0.9738

Histone	Residues	Sequence	Shift in wavelength (nM)
Histone 3	1-21	ARTKQTARΘΣTGGKAPRKQLA *-spacer-Biotin*	0.8648
		ARTKQTARΘΣΩGGKAPRKQLA *-spacer-Biotin*	0.7938
		ARTKQTARKΣTGGKAPRKQLA *-spacer-Biotin*	0.7969
		ARTKQTARKSΩGGKAPRKQLA *-spacer-Biotin*	0.9246
		ARTKQTARKΣΩGGKAPRKQLA *-spacer-Biotin*	0.5936
	4-24	Ac-KQTARKΣΩGGΔAPRKQLATKA *-spacer-Biotin*	0.0106
		Ac-KQTARKΣΩGGΘAPRKQLATKA *-spacer-Biotin*	0.017
		Ac-KQTARKSΩGGΔAPRKQLATKA *-spacer-Biotin*	0.0576
		Ac-KQTARKSΩGGΘAPRKQLATKA *-spacer-Biotin*	0.0879
		Ac-KQTARΔSTGGΔAPRKQLATKA *-spacer-Biotin*	0.088
	8-28	Ac-RKSTGGΔAPRKQLATKAARKS *-spacer-Biotin*	0.3663
		Ac-RKSTGGΔAPΨKQLATKAARKS *-spacer-Biotin*	0.1225
		Ac-RKSTGGΦAPRKQLATKAARKS *-spacer-Biotin*	0.3835
		Ac-RKSTGGΠAPRKQLATKAARKS *-spacer-Biotin*	0.2907
		Ac-RKSTGGΘAPRKQLATKAARKS *-spacer-Biotin*	0.2306
		Ac-RKSTGGΘAPΨKQLATKAARKS *-spacer-Biotin*	-0.0084
		Ac-RKSTGGKAPΞKQLATKAARKS *-spacer-Biotin*	0.5753
		Ac-RKSTGGKAPΨKQLATKAARKS *-spacer-Biotin*	0.384
		Ac-RKSTGGΔAPRΔQLATKAARKS *-spacer-Biotin*	0.3173
	12-32	Ac-GGKAPΨΔQLATKAARKSAPAT *-spacer-Biotin*	0.0381
		Ac-GGKAPRΔQLATKAARKSAPAT *-spacer-Biotin*	0.0749
		Ac-GGKAPRΔQLATΘAARKSAPAT *-spacer-Biotin*	0.0464
		Ac-GGKAPRΔQLATΔAARKSAPAT *-spacer-Biotin*	0.0344
	18-38	Ac-KQLATΔAARKSAPATGGVKKP *-spacer-Biotin*	0.0631
		Ac-KQLATΔAAΨKSAPATGGVKKP *-spacer-Biotin*	0.0171
		Ac-KQLATΔAAΨΔSAPATGGVKKP *-spacer-Biotin*	0.0041
		Ac-KQLATΦAARKSAPATGGVKKP *-spacer-Biotin*	0.106
		Ac-KQLATΠAARKSAPATGGVKKP *-spacer-Biotin*	0.101
		Ac-KQLATΘAARKSAPATGGVKKP *-spacer-Biotin*	0.1199
		Ac-KQLATΘAAΨKSAPATGGVKKP *-spacer-Biotin*	0.0359
		Ac-KQLATΘAAΨΔSAPATGGVKKP *-spacer-Biotin*	0.0068
	20-40	Ac-LATKAARKSAPATGGVKKPHR *-spacer-Biotin*	0.2157
		Ac-LATKAAΞKSAPATGGVKKPHR *-spacer-Biotin*	0.2513
		Ac-LATKAAΨKSAPATGGVKKPHR *-spacer-Biotin*	0.1182
		Ac-LATKAAΨΔSAPATGGVKKPHR *-spacer-Biotin*	0.0609
		Ac-LATKAAΨΔΣAPATGGVKKPHR *-spacer-Biotin*	0.0265
		Ac-LATKAAΨKΣAPATGGVKKPHR *-spacer-Biotin*	0.0409
		Ac-LATKAARΔSAPATGGVKKPHR *-spacer-Biotin*	0.0911
		Ac-LATKAARΔΣAPATGGVKKPHR *-spacer-Biotin*	0.0153
		Ac-LATKAARΦSAPATGGVKKPHR *-spacer-Biotin*	0.1425

Histone	Residues	Sequence	Shift in wavelength (nM)
Histone 3	20-40	Ac-LATKAARΠSAPATGGVKKPHR *-spacer-Biotin*	0.3438
		Ac-LATKAARΘSAPATGGVKKPHR *-spacer-Biotin*	0.2803
		Ac-LATKAAΨΘSAPATGGVKKPHR *-spacer-Biotin*	0.1177
		Ac-LATKAARΘΣAPATGGVKKPHR *-spacer-Biotin*	0.0284
		Ac-LATKAAΨΘΣAPATGGVKKPHR *-spacer-Biotin*	0.0335
		Ac-LATKAARKΣAPATGGVKKPHR *-spacer-Biotin*	0.0597
Histone 4	1-21	SGRGKGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.3729
		ΣGRGKGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.3066
		SGΞGKGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.3145
		SGΨGKGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.2784
		SGΨGΔGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.1622
		ΣGΨGΔGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.1463
		SGRGΔGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.37
		SGRGΔGGΔGLGΔGGAKRHRKV *-spacer-Biotin*	0.0004
		SGRGΔGGΔGLGKGGAKRHRKV *-spacer-Biotin*	0.2382
		SGRGKGGΔGLGKGGAKRHRKV *-spacer-Biotin*	0.3858
		SGRGKGGΔGLGΔGGAKRHRKV *-spacer-Biotin*	0.2976
		SGRGKGGKGLGΦGGAKRHRKV *-spacer-Biotin*	0.3076
		SGRGKGGKGLGΠGGAKRHRKV *-spacer-Biotin*	0.2924
		SGRGKGGKGLGΘGGAKRHRKV *-spacer-Biotin*	0.2843
		SGRGKGGΔGLGΦGGAKRHRKV *-spacer-Biotin*	0.2936
		SGRGKGGΔGLGΘGGAKRHRKV *-spacer-Biotin*	0.2524
	6-26	Ac-GGKGLGΦGGAΔRHRKVLRDNI *-spacer-Biotin*	-0.015
		Ac-GGKGLGΘGGAΔRHRKVLRDNI *-spacer-Biotin*	0.1523
		Ac-GGKGLGKGGAΔRHRKVLRDNI *-spacer-Biotin*	0.1596
	11-31	Ac-GKGGAKRHRKVLRDNIQGITK *-spacer-Biotin*	0.0843
		Ac-GKGGAKRHRΔVLRDNIQGITK *-spacer-Biotin*	0.1406
		Ac-GKGGAΔRHRΔVLRDNIQGITK *-spacer-Biotin*	0.0848
		Ac-GKGGAΔRHRΦVLRDNIQGITK *-spacer-Biotin*	0.0312
		Ac-GKGGAΔRHRΘVLRDNIQGITK *-spacer-Biotin*	0.0812
		Ac-GKGGAKRHRΦVLRDNIQGITK *-spacer-Biotin*	0.2116
		Ac-GKGGAKRHRΠVLRDNIQGITK *-spacer-Biotin*	0.1688
		Ac-GKGGAKRHRΘVLRDNIQGITK *-spacer-Biotin*	0.1469
		Control 2	0.0077

Histone	Residues	Sequence	Shift in wavelength (nM)
		Control 3	0
Histone 4	1-21	SGRGKGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.7566
		SGRGΔGGKGLGKGGAKRHRKV *-spacer-Biotin*	0.5385
		SGRGKGGΔGLGKGGAKRHRKV *-spacer-Biotin*	0.6713
		SGRGKGGKGLGΔGGAKRHRKV *-spacer-Biotin*	0.4928
		SGRGKGGKGLGKGGAΔRHRKV *-spacer-Biotin*	0.4898
		SGRGΔGGΔGLGKGGAKRHRKV *-spacer-Biotin*	0.4703
		SGRGΔGGKGLGΔGGAKRHRKV *-spacer-Biotin*	0.4216
		SGRGΔGGKGLGKGGAΔRHRKV *-spacer-Biotin*	0.4149
		SGRGKGGΔGLGΔGGAKRHRKV *-spacer-Biotin*	0.4853
		SGRGKGGΔGLGKGGAΔRHRKV *-spacer-Biotin*	0.4607
		SGRGKGGKGLGΔGGAΔRHRKV *-spacer-Biotin*	0.3505
		SGRGΔGGΔGLGΔGGAKRHRKV *-spacer-Biotin*	0.5527
		SGRGKGGΔGLGΔGGAΔRHRKV *-spacer-Biotin*	0.5523
		SGRGΔGGKGLGΔGGAΔRHRKV *-spacer-Biotin*	0.4593
		SGRGΔGGΔGLGKGGAΔRHRKV *-spacer-Biotin*	0.5012
		SGRGΔGGΔGLGΔGGAΔRHRKV *-spacer-Biotin*	0.3271
	9-29	Ac-GLGKGGAKRHRKVLRDNIQGI *-spacer-Biotin*	0.2231
		Ac-GLGKGGAKRHRΔVLRDNIQGI *-spacer-Biotin*	0.1951
		Ac-GLGKGGAΔRHRKVLRDNIQGI *-spacer-Biotin*	0.1666
		Ac-GLGΔGGAΔRHRKVLRDNIQGI *-spacer-Biotin*	0.1222
		Ac-GLGKGGAΔRHRΔVLRDNIQGI *-spacer-Biotin*	0.0544
		Ac-GLGΔGGAΔRHRΔVLRDNIQGI *-spacer-Biotin*	-0.0086
Histone 2A	1-21	SGRGKQGGKARAKAKTRSSRA *-spacer-Biotin*	0.5094
		SGRGΔQGGKARAKAKTRSSRA *-spacer-Biotin*	0.6783
		SGRGKQGGΔARAKAKTRSSRA *-spacer-Biotin*	0.744
		SGRGKQGGKARAΔAKTRSSRA *-spacer-Biotin*	0.6117
		SGRGKQGGKARAKAΔTRSSRA *-spacer-Biotin*	0.5603
		SGRGΔQGGΔARAKAKTRSSRA *-spacer-Biotin*	0.5393
		SGRGΔQGGKARAΔAKTRSSRA *-spacer-Biotin*	0.4645
		SGRGΔQGGKARAKAΔTRSSRA *-spacer-Biotin*	0.3473
		SGRGKQGGΔARAΔAKTRSSRA *-spacer-Biotin*	0.4119
		SGRGKQGGΔARAKAΔTRSSRA *-spacer-Biotin*	0.3853
		SGRGKQGGKARAΔAΔTRSSRA *-spacer-Biotin*	0.3521
		SGRGΔQGGΔARAΔAKTRSSRA *-spacer-Biotin*	0.3236
		SGRGΔQGGKARAΔAΔTRSSRA *-spacer-Biotin*	0.2684
		SGRGKQGGΔARAΔAΔTRSSRA *-spacer-Biotin*	0.3394
		SGRGΔQGGΔARAKAΔTRSSRA *-spacer-Biotin*	0.4629
		SGRGΔQGGΔARAΔAΔTRSSRA *-spacer-Biotin*	0.2459
Histone 2B	1-21	PEPAKSAPAPKKGSKKAVTKA *-spacer-Biotin*	0.2527
		PEPAΔSAPAPKKGSKKAVTKA *-spacer-Biotin*	0.0716
		PEPAKSAPAPΔKGSKKAVTKA *-spacer-Biotin*	0.0648

Histone	Residues	Sequence	Shift in wavelength (nM)
Histone 2B	1-21	PEPAKSAPAPKΔGSKKAVTKA *-spacer-Biotin*	0.0928
		PEPAKSAPAPKKGSΔKAVTKA *-spacer-Biotin*	0.0591
		PEPAΔSAPAPΔKGSKΔAVTKA *-spacer-Biotin*	-0.0252
		PEPAΔSAPAPKΔGSKKAVTKA *-spacer-Biotin*	0.0278
		PEPAΔSAPAPKKGSΔKAVTKA *-spacer-Biotin*	-0.0036
		PEPAΔSAPAPKKGSKΔAVTKA *-spacer-Biotin*	0.0421
		PEPAKSAPAPGSKKAVTKA *-spacer-Biotin*	-0.0085
		PEPAKSAPAPΔKGSΔKAVTKA *-spacer-Biotin*	-0.003
		PEPAKSAPAPΔKGSKΔAVTKA *-spacer-Biotin*	-0.0037
		PEPAΔSAPAPΔΔGSKKAVTKA *-spacer-Biotin*	-0.0069
		PEPAΔSAPAPKKGSΔΔAVTKA *-spacer-Biotin*	-0.0302
		PEPAKSAPAPΔΔGSΔΔAVTKA *-spacer-Biotin*	-0.0347
		PEPAKSAPAPΔKGSΔΔAVTKA *-spacer-Biotin*	-0.0343
		PEPAKSAPAPKΔGSΔΔAVTKA *-spacer-Biotin*	-0.0455
		PEPAKSAPAPΔΔGSΔKAVTKA *-spacer-Biotin*	-0.0264
		PEPAKSAPAPΔΔGSKΔAVTKA *-spacer-Biotin*	0.0041
		PEPAΔSAPAPΔΔGSΔΔAVTKA *-spacer-Biotin*	-0.0612
	13-33	Ac-GSKKAVTKAQKKDGKKRKRSR *-spacer-Biotin*	0.9105
		Ac-GSKKAVTΔAQKKDGKKRKRSR *-spacer-Biotin*	1.0681
		Ac-GSKKAVTKAQΔKDGKKRKRSR *-spacer-Biotin*	1.0752
		Ac-GSKKAVTKAQKΔDGKKRKRSR *-spacer-Biotin*	0.9499
		Ac-GSKKAVTΔAQΔΔDGKKRKRSR *-spacer-Biotin*	0.7628
H3	1-21	ARTKQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.6969
		ARTΔQTARKSTGGKAPRKQLA *-spacer-Biotin*	0.2993
		ARTKQTARΔSTGGKAPRKQLA *-spacer-Biotin*	0.496
		ARTKQTARKSTGGΔAPRKQLA *-spacer-Biotin*	0.6536
		ARTΔQTARΔSTGGKAPRKQLA *-spacer-Biotin*	0.2429
		ARTΔQTARKSTGGΔAPRKQLA *-spacer-Biotin*	0.3065
		ARTKQTARΔSTGGΔAPRKQLA *-spacer-Biotin*	0.6154
		ARTΔQTARΔSTGGΔAPRKQLA *-spacer-Biotin*	0.3286
	11-31	Ac-TGGKAPRKQLATKAARKSAPA *-spacer-Biotin*	0.1311
		Ac-TGGΔAPRKQLATKAARKSAPA *-spacer-Biotin*	0.1835
		Ac-TGGKAPRΔQLATKAARKSAPA *-spacer-Biotin*	0.1081
		Ac-TGGKAPRKQLATΔAARKSAPA *-spacer-Biotin*	0.0669
		Ac-TGGKAPRKQLATKAARΔSAPA *-spacer-Biotin*	0.068
		Ac-TGGΔAPRΔQLATKAARKSAPA *-spacer-Biotin*	0.1382
		Ac-TGGΔAPRKQLATΔAARKSAPA *-spacer-Biotin*	0.0436
		Ac-TGGΔAPRKQLATKAARΔSAPA *-spacer-Biotin*	0.0482
		Ac-TGGKAPRΔQLATΔAARKSAPA *-spacer-Biotin*	0.023
		Ac-TGGKAPRΔQLATKAARΔSAPA *-spacer-Biotin*	0.0495
		Ac-TGGKAPRKQLATΔAARΔSAPA *-spacer-Biotin*	0.024
		Ac-TGGΔAPRKQLATΔAARΔSAPA *-spacer-Biotin*	0.0263

Histone	Residues	Sequence	Shift in wavelength (nM)
H3	11-31	Ac-TGGKAPRΔQLATΔAARΔSAPA *-spacer-Biotin*	-0.008
		Ac-TGGΔAPRΔQLATKAARΔSAPA *-spacer-Biotin*	-0.0084
		Ac-TGGΔAPRΔQLATΔAARΔSAPA *-spacer-Biotin*	-0.0201
	23-43	Ac-KAARKSAPATGGVKKPHRYRP *-spacer-Biotin*	0.4517
		Ac-KAARKSAPATGGVΔKPHRYRP *-spacer-Biotin*	0.3467
		Ac-KAARKSAPATGGVKΔPHRYRP *-spacer-Biotin*	0.2835
		Ac-KAARKSAPATGGVΔΔPHRYRP *-spacer-Biotin*	0.1125
		Ac-KAARΔSAPATGGVKKPHRYRP *-spacer-Biotin*	0.3221
		Ac-KAARΔSAPATGGVΔKPHRYRP *-spacer-Biotin*	0.1612
		Ac-KAARΔSAPATGGVKΔPHRYRP *-spacer-Biotin*	0.159
		Ac-KAARΔSAPATGGVΔΔPHRYRP *-spacer-Biotin*	0.0554
		Control 4	0.0112

Appendix 4.2 – X-ray Crystal Structure Statistics for DPF3b

DATA COLLECTION

X-ray source	DIAMOND-I04-1
Resolution	30.7-1.7
Space group	P3
Cell dimensions [Å]	a = 56.1 Å b = 56.1 Å c = 153.6 Å
Unique reflections	60852
Completeness [%]	99.1
R_{merge} [%]	16.4
R_{pim} [%]	8.3
$CC_{1/2}$	0.990
I/σI	7.8

REFINEMENT

Resolution range [Å]	30.1-1.7
Number of reflections	58206
No. of atoms (protein/sulphate/citrate/ chloride/MPD/nitrate/ EG/sugar/H_2O/tartrate/ acetate)	3424/0/0/0/0/ 0/0/0/71/0/0
B factors [$Å^2$]	11.369
R_{factor} [%)]	22.5
R_{free} [%]	25.9
r.m.s.d. bonds [Å]	0.032
r.m.s.d. angles [deg]	2.998
Ramachandran statistics	
Favoured [%]	89.68%
Disallowed [%]	2.52%

References

(1) Lukk, M.; Kapushesky, M.; Nikkilä, J.; Parkinson, H.; Goncalves, A.; Huber, W.; Ukkonen, E.; Brazma, A. A global map of human gene expression. *Nature Biotechnology*, 2010, *28*, 322–324.

(2) Waddington, C. H. *Int. J. Epidemiol.* **2012**, *41* (1), 10–13.

(3) Kornberg, R. D.; Lorch, Y. *Cell* **1999**, *98* (3), 285–294.

(4) Jenuwein, T.; Allis, C. D. *Science* **2001**, *293* (5532), 1074–1080.

(5) Arrowsmith, C. H.; Bountra, C.; Fish, P. V; Lee, K.; Schapira, M. *Nat. Rev. Drug Discov.* **2012**, *11* (5), 384–400.

(6) Kornberg, R. D. *Science* **1974**, *184* (139), 868–871.

(7) Harp, J. M.; Hanson, B. L.; Timm, D. E.; Bunick, G. J. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2000**, *56* (12), 1513–1534.

(8) Richmond, R. K.; Sargent, D. F.; Richmond, T. J.; Luger, K.; Ma, A. W. **1997**, *7*, 251–260.

(9) Clapier, C. R.; Cairns, B. R. *Annu. Rev. Biochem.* **2009**, *78*, 273–304.

(10) Lewis, J.; Bird, A. *FEBS Lett.* **1991**, *285* (2), 155–159.

(11) Ooi, S. K. T.; Qiu, C.; Bernstein, E.; Li, K.; Jia, D.; Yang, Z.; Erdjument-Bromage, H.; Tempst, P.; Lin, S.-P.; Allis, C. D.; Cheng, X.; Bestor, T. H. *Nature* **2007**, *448* (7154), 714–717.

(12) Viré, E.; Brenner, C.; Deplus, R.; Blanchon, L.; Fraga, M.; Didelot, C.; Morey, L.; Van Eynde, A.; Bernard, D.; Vanderwinden, J.-M.; Bollen, M.; Esteller, M.; Di Croce, L.; de Launoit, Y.; Fuks, F. *Nature* **2006**, *439* (7078), 871–874.

(13) Schlesinger, Y.; Straussman, R.; Keshet, I.; Farkash, S.; Hecht, M.; Zimmerman, J.; Eden, E.; Yakhini, Z.; Ben-Shushan, E.; Reubinoff, B. E.; Bergman, Y.; Simon, I.; Cedar, H. *Nat. Genet.* **2007**, *39* (2), 232–236.

(14) Woodcock, D. M.; Crowther, P. J.; Diver, W. *Biochem. Biophys. Res. Commun.* **1987**, *145* (2), 888–894.

(15) Lister, R.; Pelizzola, M.; Dowen, R. H.; Hawkins, R. D.; Hon, G.; Tonti-Filippini, J.; Nery, J. R.; Lee, L.; Ye, Z.; Ngo, Q.-M.; Edsall, L.; Antosiewicz-Bourget, J.; Stewart, R.; Ruotti, V.; Millar, a H.; Thomson, J. a; Ren, B.; Ecker, J. R. *Nature* **2009**, *462* (7271), 315–322.

(16) Scourzic, L.; Mouly, E.; Bernard, O. A. *Genome Med.* **2015**, *7* (1), 1–16.

(17) Bannister, A. J.; Kouzarides, T. *Cell Res.* **2011**, *21* (3), 381–395.

(18) Lee, K. K.; Workman, J. L. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (4), 284–295.

(19) Zhang, X.; Wen, H.; Shi, X. *Acta Biochim. Biophys. Sin. (Shanghai).* **2012**, *44* (1), 14–27.

(20) Di Lorenzo, A.; Bedford, M. T. *FEBS Lett.* **2011**, *585* (13), 2024–2031.

(21) Berndsen, C. E.; Denu, J. M. *Curr. Opin. Struct. Biol.* **2008**, *18* (6), 682–689.

(22) Structural Genomics Consortium. KAT Tree http://apps.thesgc.org/resources/phylogenetic_trees/?domain=KAT#options (accessed Sep 30, 2015).

(23) Lombardi, P. M.; Cole, K. E.; Dowling, D. P.; Christianson, D. W. *Curr. Opin. Struct. Biol.* **2011**, *21* (6), 735–743.

(24) FDA. Belinostat http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm403960.htm (accessed Sep 8, 2015).

(25) U.S. Food and Drug Administration, Panobinostat http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm435339.htm (accessed Jul 26, 2015).

(26) Qiao, Z.; Ren, S.; Li, W.; Wang, X.; He, M.; Guo, Y.; Sun, L.; He, Y.; Ge, Y.; Yu, Q. *Biochem. Biophys. Res. Commun.* **2013**, *434* (1), 95–101.

(27) Filippakopoulos, P.; Picaud, S.; Mangos, M.; Keates, T.; Lambert, J.-P.; Barsyte-Lovejoy, D.; Felletar, I.; Volkmer, R.; Müller, S.; Pawson, T.; Gingras, A.-C.; Arrowsmith, C. H.; Knapp, S. *Cell* **2012**, *149* (1), 214–231.

(28) Filippakopoulos, P.; Knapp, S. *Nat. Rev. Drug Discov.* **2014**, *13* (5), 337–356.

(29) Gallenkamp, D.; Gelato, K. a; Haendler, B.; Weinmann, H. *ChemMedChem* **2014**, *9* (3), 438–464.

(30) Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I.; Philpott, M.; Munro, S.; McKeown, M. R.; Wang, Y.; Christie, A. L.; West, N.; Cameron, M. J.; Schwartz, B.; Heightman, T. D.; La Thangue, N.; French, C. a; Wiest, O.; Kung, A. L.; Knapp, S.; Bradner, J. E. *Nature* **2010**, *468* (7327), 1067–1073.

(31) Hay, D. A.; Fedorov, O.; Martin, S.; Singleton, D. C.; Tallant, C.; Wells, C.; Picaud, S.; Philpott, M.; Monteiro, O. P.; Rogers, C. M.; Conway, S. J.; Rooney, T. P. C.; Tumber, A.; Yapp, C.; Filippakopoulos, P.; Bunnage, M. E.; Müller, S.; Knapp, S.; Schofield, C. J.; Brennan, P. E. *J. Am. Chem. Soc.* **2014**, *136* (26), 9308–9319.

(32) Barski, A.; Cuddapah, S.; Cui, K.; Roh, T.-Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K. *Cell* **2007**, *129* (4), 823–837.

(33) Feng, Q.; Wang, H.; Ng, H. H.; Erdjument-Bromage, H.; Tempst, P.; Struhl, K.; Zhang, Y. *Curr. Biol.* **2002**, *12* (12), 1052–1058.

(34) Kipp, D. R.; Quinn, C. M.; Fortin, P. D. *Biochemistry* **2013**, *52* (39), 6866–6878.

(35) Wu, H.; Min, J.; Lunin, V. V.; Antoshenko, T.; Dombrovski, L.; Zeng, H.; Allali-Hassani, A.; Campagna-Slater, V.; Vedadi, M.; Arrowsmith, C. H.; Plotnikov, A. N.; Schapira, M. *PLoS One* **2010**, *5* (1).

(36) Trievel, R. C.; Flynn, E. M.; Houtz, R. L.; Hurley, J. H. *Nat. Struct. Biol.* **2003**, *10* (7), 545–552.

(37) Shi, Y.; Lan, F.; Matson, C.; Mulligan, P.; Whetstine, J. R.; Cole, P. A.; Casero, R. A.; Shi, Y. *Cell* **2004**, *119* (7), 941–953.

(38) Sánchez-Fernández, E. M.; Tarhonskaya, H.; Al-Qahtani, K.; Hopkinson, R. J.; McCullagh, J. S. O.; Schofield, C. J.; Flashman, E. *Biochem. J.* **2013**, *449* (2), 491–496.

(39) Seward, D. J.; Cubberley, G.; Kim, S.; Schonewald, M.; Zhang, L.; Tripet, B.; Bentley, D. L. *Nat. Struct. Mol. Biol.* **2007**, *14* (3), 240–242.

(40) Chang, B.; Chen, Y.; Zhao, Y.; Bruick, R. K. *Science* **2007**, *318* (5849), 444–447.

(41) Musselman, C. A.; Lalonde, M.-E.; Côté, J.; Kutateladze, T. G. *Nat. Struct. Mol. Biol.* **2012**, *19* (12), 1218–1227.

(42) Wagner, T.; Robaa, D.; Sippl, W.; Jung, M. *ChemMedChem* **2014**, *9* (3), 466–483.

(43) Budhidarmo, R.; Nakatani, Y.; Day, C. L. *Trends Biochem. Sci.* **2012**, *37* (2), 58–65.

(44) Xi, Q.; Wang, Z.; Zaromytidou, A.-I.; Zhang, X. H.-F.; Chow-Tsang, L.-F.; Liu, J. X.; Kim, H.; Barlas, A.; Manova-Todorova, K.; Kaartinen, V.; Studer, L.; Mark, W.; Patel, D. J.; Massagué, J. *Cell* **2011**, *147* (7), 1511–1524.

(45) Dhayalan, A.; Tamas, R.; Bock, I.; Tattermusch, A.; Dimitrova, E.; Kudithipudi, S.; Ragozin, S.; Jeltsch, A. *Hum. Mol. Genet.* **2011**, *20* (11), 2195–2203.

(46) Rothbart, S. B.; Krajewski, K.; Nady, N.; Tempel, W.; Xue, S.; Badeaux, A. I.; Barsyte-Lovejoy, D.; Martinez, J. Y.; Bedford, M. T.; Fuchs, S. M.; Arrowsmith, C. H.; Strahl, B. D. *Nat. Struct. Mol. Biol.* **2012**, *19* (11), 1155–1160.

(47) Baker, L. A.; Allis, C. D.; Wang, G. G. *Mutat. Res.* **2008**, *647* (1-2), 3–12.

(48) Gellert, M. *Annu. Rev. Biochem.* **2002**, *71* (D), 101–132.

(49) Marrella, V.; Poliani, P. L.; Sobacchi, C.; Grassi, F.; Villa, A. *Trends Immunol.* **2008**, *29* (3), 133–140.

(50) Ramón-Maiques, S.; Kuo, A. J.; Carney, D.; Matthews, A. G. W.; Oettinger, M. a; Gozani, O.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (48), 18993–18998.

(51) Peña, P. V; Davrazou, F.; Shi, X.; Walter, K. L.; Verkhusha, V. V; Gozani, O.; Zhao, R.; Kutateladze, T. G. *Nature* **2006**, *442* (7098), 100–103.

(52) Campos, E. I.; Martinka, M.; Mitchell, D. L.; Dai, D. L.; Li, G. *Int. J. Oncol.* **2004**, *25* (1), 73–80.

(53) Moore, M. A. S.; Chung, K. Y.; Plasilova, M.; Schuringa, J. J.; Shieh, J. H.; Zhou, P.; Morrone, G. In *Annals of the New York Academy of Sciences*; 2007; Vol. 1106, pp 114–142.

(54) Kasper, L. H.; Brindle, P. K.; Schnabel, C. A.; Pritchard, C. E.; Cleary, M. L.; van Deursen, J. M. *Mol. Cell. Biol.* **1999**, *19* (1), 764–776.

(55) Van Zutven, L. J. C. M.; Önen, E.; Velthuizen, S. C. J. M.; Van Drunen, E.; Von Bergh, A. R. H.; Van Den Heuvel-Eibrink, M. M.; Veronese, A.; Mecucci, C.; Negrini, M.; De Greef, G. E.; Beverloo, H. B. *Genes Chromosom. Cancer* **2006**, *45* (5), 437–446.

(56) Reader, J. C.; Meekins, J. S.; Gojo, I.; Ning, Y. *Leukemia* **2007**, *21* (4), 842–844.

(57) Rosati, R.; La Starza, R.; Veronese, A.; Aventin, A.; Schwienbacher, C.; Vallespi, T.; Negrini, M.; Martelli, M. F.; Mecucci, C. *Blood* **2002**, *99* (10), 3857–3860.

(58) Tatton-Brown, K.; Douglas, J.; Coleman, K.; Baujat, G.; Cole, T. R. P.; Das, S.; Horn, D.; Hughes, H. E.; Temple, I. K.; Faravelli, F.; Waggoner, D.; Turkmen, S.; Cormier-Daire, V.; Irrthum, A.; Rahman, N. *Am. J. Hum. Genet.* **2005**, *77* (2), 193–204.

(59) Argentaro, A.; Yang, J.-C.; Chapman, L.; Kowalczyk, M. S.; Gibbons, R. J.; Higgs, D. R.; Neuhaus, D.; Rhodes, D. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (29), 11939–11944.

(60) Gécz, J.; Turner, G.; Nelson, J.; Partington, M. *Eur. J. Hum. Genet.* **2006**, *14* (12), 1233–1237.

(61) Van Vlierberghe, P.; Palomero, T.; Khiabanian, H.; Van der Meulen, J.; Castillo, M.; Van Roy, N.; De Moerloose, B.; Philippé, J.; González-García, S.; Toribio, M. L.; Taghon, T.; Zuurbier, L.; Cauwelier, B.; Harrison, C. J.; Schwab, C.; Pisecker, M.; Strehl, S.; Langerak, A. W.; Gecz, J.; Sonneveld, E.; Pieters, R.; Paietta, E.; Rowe, J. M.; Wiernik, P. H.; Benoit, Y.; Soulier, J.; Poppe, B.; Yao, X.; Cordon-Cardo, C.; Meijerink, J.; Rabadan, R.; Speleman, F.; Ferrando, A. *Nat. Genet.* **2010**, *42* (4), 338–342.

(62) Ikari, J.; Inamine, a; Yamamoto, T.; Watanabe-Takano, H.; Yoshida, N.; Fujimura, L.; Taniguchi, T.; Sakamoto, a; Hatano, M.; Tatsumi, K.; Tokuhisa, T.; Arima, M. *Allergy* **2014**, *69* (2), 223–230.

(63) Paul, S.; Kuo, A.; Schalch, T.; Vogel, H.; Joshua-Tor, L.; McCombie, W. R.; Gozani, O.; Hammell, M.; Mills, A. A. *Cell Rep.* **2013**, *3* (1), 92–102.

(64) Bunnage, M. E.; Chekler, E. L. P.; Jones, L. H. *Nat. Chem. Biol.* **2013**, *9* (4), 195–199.

(65) Frye, S. V. *Nat. Chem. Biol.* **2010**, *6* (3), 159–161.

(66) SGC | Epigenetics Probes Collection http://www.thesgc.org/chemical-probes/epigenetics (accessed Aug 8, 2014).

(67) Edwards, A. M.; Arrowsmith, C. H.; Bountra, C.; Bunnage, M. E.; Feldmann, M.; Knight, J. C.; Patel, D. D.; Prinos, P.; Taylor, M. D.; Sundström, M.; Barker, P.; Barsyte, D.; Bengtson, M. H.; Bell, C.; Bowness, P.; Boycott, K. M.; Buser-Doepner, C.; Carpenter, C. L.; Carr, A. J.; Clark, K.; Das, A. M.; Dhanak, D.; Dirks, P.; Ellis, J.; Fantin, V. R.; Flores, C.; Fon, E. A.; Frail,

D. E.; Gileadi, O.; O'Hagan, R. C.; Howe, T.; Isaac, J. T. R.; Jabado, N.; Jakobsson, P.-J.; Klareskog, L.; Knapp, S.; Lee, W. H.; Lima-Fernandes, E.; Lundberg, I. E.; Marshall, J.; Massirer, K. B.; MacKenzie, A. E.; Maruyama, T.; Mueller-Fahrnow, A.; Muthuswamy, S.; Nanchahal, J.; O'Brien, C.; Oppermann, U.; Ostermann, N.; Petrecca, K.; Pollock, B. G.; Poupon, V.; Prinjha, R. K.; Rosenberg, S. H.; Rouleau, G.; Skingle, M.; Slutsky, A. S.; Smith, G. A. M.; Verhelle, D.; Widmer, H.; Young, L. T. *Nat. Rev. Drug Discov.* **2015**, *14* (3), 149–150.

(68) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. *Nat. Rev. Drug Discov.* **2010**, *9* (3), 203–214.

(69) Lowe, D. Chemical Probes Versus Drugs http://pipeline.corante.com/archives/2013/04/01/chemical_probes_versus_drugs.php (accessed Sep 9, 2015).

(70) Anighoro, A.; Bajorath, J.; Rastelli, G. *J. Med. Chem.* **2014**, *57*, 7874–7887.

(71) James, L. I.; Barsyte-Lovejoy, D.; Zhong, N.; Krichevsky, L.; Korboukh, V. K.; Herold, J. M.; Macnevin, C. J.; Norris, J. L.; Sagum, C. a; Tempel, W.; Marcon, E.; Guo, H.; Gao, C.; Huang, X.-P.; Duan, S.; Emili, A.; Greenblatt, J. F.; Kireev, D. B.; Jin, J.; Janzen, W. P.; Brown, P. J.; Bedford, M. T.; Arrowsmith, C. H.; Frye, S. V. *Nat. Chem. Biol.* **2013**, *9* (3), 184–191.

(72) Simhadri, C.; Daze, K. D.; Douglas, S. F.; Quon, T. T. H.; Dev, A.; Gignac, M. C.; Peng, F.; Heller, M.; Boulanger, M. J.; Wulff, J. E.; Hof, F. *J. Med. Chem.* **2014**, *7*.

(73) Wagner, E. K.; Nath, N.; Flemming, R.; Feltenberger, J. B.; Denu, J. M. *Biochemistry* **2012**.

(74) Miller, T. C. R.; Rutherford, T. J.; Birchall, K.; Chugh, J.; Fiedler, M.; Bienz, M. *ACS Chem. Biol.* **2014**, *9* (12), 2864–2874.

(75) Herold, J. M.; Wigle, T. J.; Norris, J. L.; Lam, R.; Korboukh, V. K.; Gao, C.; Ingerman, L. a; Kireev, D. B.; Senisterra, G.; Vedadi, M.; Tripathy, A.; Brown, P. J.; Arrowsmith, C. H.; Jin, J.; Janzen, W. P.; Frye, S. V. *J. Med. Chem.* **2011**, *54* (7), 2504–2511.

(76) Herold, J. M.; James, L. I.; Korboukh, V. K.; Gao, C.; Coil, K. E.; Bua, D. J.; Norris, J. L.; Kireev, D. B.; Brown, P. J.; Jin, J.; Janzen, W. P.; Gozani, O.; Frye, S. V. *Medchemcomm* **2012**, *3* (1), 45.

(77) James, L. I.; Korboukh, V. K.; Krichevsky, L.; Baughman, B. M.; Herold, J. M.; Norris, J. L.; Jin, J.; Kireev, D. B.; Janzen, W. P.; Arrowsmith, C. H.; Frye, S. V. *J. Med. Chem.* **2013**, *56* (18), 7358–7371.

(78) Santiago, C.; Nguyen, K.; Schapira, M. *J. Comput. Aided. Mol. Des.* **2011**, *25* (12), 1171–1178.

(79) Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; Farrell, C. M.; Hart, J.; Landrum, M. J.; McGarvey, K. M.; Murphy, M. R.; O'Leary, N. A.; Pujar, S.; Rajput, B.; Rangwala, S. H.; Riddick, L. D.; Shkeda, A.; Sun, H.; Tamez, P.; Tully, R. E.; Wallin, C.; Webb, D.; Weber, J.; Wu, W.; Dicuccio, M.; Kitts, P.; Maglott, D. R.; Murphy, T. D.; Ostell, J. M. *Nucleic Acids Res.* **2014**, *42* (D1), 756–763.

(80) Brice, M. D.; Rodgers, J. R.; Kennard, O. *Methods Biochem. Anal.* **2003**, *44*, 181–198.

(81) Li, Y.; Li, H. *Acta Biochim. Biophys. Sin. (Shanghai).* **2012**, *44* (1), 28–39.

(82) Musselman, C.; Kutateladze, T. *Mol. Interv.* **2009**, *9* (6), 314–323.

(83) Musselman, C. A.; Kutateladze, T. G. *Nucleic Acids Res.* **2011**, *39* (21), 9061–9071.

(84) Sanchez, R.; Zhou, M.-M. *Trends Biochem. Sci.* **2011**, *36* (7), 364–372.

(85) Eddy, S. R. *PLoS Comput. Biol.* **2011**, *7* (10), e1002195.

(86) Benson, D. A.; Karsch-Mizrachi, I.; Clark, K.; Lipman, D. J.; Ostell, J.; Sayers, E. W. *Nucleic Acids Res.* **2012**, *40* (D1), 1–6.

(87) Magrane, M.; Consortium, U. P. *Database* **2011**, *2011*, 1–13.

(88) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. *Nucleic Acids Res.* **2003**, *31* (1), 365–370.

(89) NCBI. GenBank, RefSeq, TPA and UniProt: What's in a Name? http://www.ncbi.nlm.nih.gov/projects/RefSeq/GenBankvsRefSeq.pdf (accessed Aug 24, 2015).

(90) Finn, R. D.; Clements, J.; Eddy, S. R. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W29–W37.

(91) Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215* (3), 403–410.

(92) Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Dj, L. *Nucleic acids Res* **1997**, *25* (17), 3389–3402.

(93) He, C.; Li, F.; Zhang, J.; Wu, J.; Shi, Y. *J. Biol. Chem.* **2013**, *288* (7), 4692–4703.

(94) Delvecchio, M.; Gaucher, J.; Aguilar-Gurrieri, C.; Ortega, E.; Panne, D. *Nat. Struct. Mol. Biol.* **2013**, *20* (9), 1040–1046.

(95) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.; Finn, R. D. *Nucleic Acids Res.* **2012**, *40* (Database issue), D290–D301.

(96) Letunic, I.; Doerks, T.; Bork, P. *Nucleic Acids Res.* **2012**, *40* (Database issue), D302–D305.

(97) Thompson, J. D.; Linard, B.; Lecompte, O.; Poch, O. *PLoS One* **2011**, *6* (3).

(98) EMBL-EBI. Bioinformatics Tools for Multiple Sequence Alignment http://www.ebi.ac.uk/Tools/msa/ (accessed Aug 24, 2015).

(99) Baum, D. *Nat. Educ.* **2009**, *1* (1), 1–5.

(100) Harrison, C. J.; Langdale, J. A. *Plant J.* **2006**, *45* (4), 561–572.

(101) Steel, M.; Penny, D. *Mol. Biol. Evol.* **2000**, *17* (6), 839–850.

(102) Edwards, A. W. F. *Syst. Biol.* **1996**, *45* (1), 79–91.

(103) Hall, B. G. *Mol. Biol. Evol.* **2013**, *30* (5), 1229–1235.

(104) Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. *Mol. Biol. Evol.* **2013**, *30* (12), 2725–2729.

(105) Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. *Mol. Biol. Evol.* **2011**, *28* (10), 2731–2739.

(106) Whelan, S.; Goldman, N. *Mol. Biol. Evol.* **2001**, *18* (5), 691–699.

(107) Chen, Y.; Wan, B.; Wang, K. C.; Cao, F.; Yang, Y.; Protacio, A.; Dou, Y.; Chang, H. Y.; Lei, M. *EMBO Rep.* **2011**, *12* (8), 797–803.

(108) Peña, P. V.; Hom, R. A.; Hung, T.; Lin, H.; Kuo, A. J.; Wong, R. P. C.; Subach, O. M.; Champagne, K. S.; Zhao, R.; Verkhusha, V. V.; Li, G.; Gozani, O.; Kutateladze, T. G. *J. Mol. Biol.* **2008**, *380* (2), 303–312.

(109) Palacios, A.; Muñoz, I. G.; Pantoja-Uceda, D.; Marcaida, M. J.; Torres, D.; Martín-García, J. M.; Luque, I.; Montoya, G.; Blanco, F. J. *J. Biol. Chem.* **2008**, *283* (23), 15956–15964.

(110) Champagne, K. S.; Saksouk, N.; Peña, P. V.; Johnson, K.; Ullah, M.; Yang, X. J.; Côté, J.; Kutateladze, T. G. *Proteins Struct. Funct. Genet.* **2008**, *72* (4), 1371–1376.

(111) Horton, J. R.; Upadhyay, A. K.; Qi, H. H.; Zhang, X.; Shi, Y.; Cheng, X. *Nat. Struct. Mol. Biol.* **2010**, *17* (1), 38–43.

(112) Ruthenburg, A. J.; Allis, C. D.; Wysocka, J. *Mol. Cell* **2007**, *25* (1), 15–30.

(113) Otani, J.; Nankumo, T.; Arita, K.; Inamoto, S.; Ariyoshi, M.; Shirakawa, M. *EMBO Rep.* **2009**, *10* (11), 1235–1241.

(114) Iwase, S.; Xiang, B.; Ghosh, S.; Ren, T.; Lewis, P. W.; Cochrane, J. C.; Allis, C. D.; Picketts, D. J.; Patel, D. J.; Li, H.; Shi, Y. *Nat. Struct. Mol. Biol.* **2011**, *18* (7), 769–776.

(115) Chakravarty, S.; Zeng, L.; Zhou, M. M. *Structure* **2009**, *17* (5), 670–679.

(116) Dreveny, I.; Deeves, S. E.; Fulton, J.; Yue, B.; Messmer, M.; Bhattacharya, A.; Collins, H. M.; Heery, D. M. *Nucleic Acids Res.* **2014**, *42* (2), 822–835.

(117) Arita, K.; Isogai, S.; Oda, T.; Unoki, M.; Sugita, K.; Sekiyama, N.; Kuwata, K.; Hamamoto, R.; Tochio, H.; Sato, M.; Ariyoshi, M.; Shirakawa, M. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, 1–6.

(118) Cheng, J.; Yang, Y.; Fang, J.; Xiao, J.; Zhu, T.; Chen, F.; Wang, P.; Li, Z.; Yang, H.; Xu, Y. *J. Biol. Chem.* **2013**, *288* (2), 1329–1339.

(119) Fiedler, M.; Sánchez-Barrena, M. J.; Nekrasov, M.; Mieszczanek, J.; Rybin, V.; Müller, J.; Evans, P.; Bienz, M. *Mol. Cell* **2008**, *30* (4), 507–518.

(120) Lemak, A.; Yee, A.; Wu, H.; Yap, D.; Zeng, H.; Dombrovski, L.; Houliston, S.; Aparicio, S.; Arrowsmith, C. H. *PLoS One* **2013**, *8* (10), e77020.

(121) Qin, S.; Jin, L.; Zhang, J.; Liu, L.; Ji, P.; Wu, M.; Wu, J.; Shi, Y. *J. Biol. Chem.* **2011**, *286* (42), 36944–36955.

(122) Liu, L.; Qin, S.; Zhang, J.; Ji, P.; Shi, Y.; Wu, J. *J. Struct. Biol.* **2012**, *180* (1), 165–173.

(123) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. *BMC Bioinformatics* **2009**, *10*, 168.

(124) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. *Nat. Protoc.* **2015**, *10* (5), 733–755.

(125) An, J.; Totrov, M.; Abagyan, R. *Mol. Cell. Proteomics* **2005**, *4* (6), 752–761.

(126) Halgren, T. *Chem. Biol. Drug Des.* **2007**, *69* (2), 146–148.

(127) Schmidtke, P.; Barril, X. *J. Med. Chem.* **2010**, *53* (15), 5858–5867.

(128) Vidler, L. R.; Brown, N.; Knapp, S.; Hoelder, S. *J. Med. Chem.* **2012**, *55* (17), 7346–7359.

(129) Campagna-Slater, V.; Mok, M. W.; Nguyen, K. T.; Feher, M.; Najmanovich, R.; Schapira, M. *J. Chem. Inf. Model.* **2011**, *51* (3), 612–623.

(130) Halgren, T. *J. Chem. Inf. Model.* **2009**, *49* (2), 377–389.

(131) Schrödinger. Schrödinger Suite 2013 Protein Preparation Wizard.

(132) Zeng, L.; Zhang, Q.; Li, S.; Plotnikov, A. N.; Walsh, M. J.; Zhou, M.-M. *Nature* **2010**, *466* (7303), 258–262.

(133) Qiu, Y.; Liu, L.; Zhao, C.; Han, C. *Genes Dev.* **2012**, *26*, 1376–1391.

(134) Chen, C.; Nott, T. J.; Jin, J.; Pawson, T. *Nat. Rev. Mol. Cell Biol.* **2011**, *12* (10), 629–642.

(135) Chew, T. G.; Peaston, A.; Lim, A. K.; Lorthongpanich, C.; Knowles, B. B.; Solter, D. *PLoS One* **2013**, *8* (7), e69764.

(136) Yang, N.; Wang, W.; Wang, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (44), 17954–17959.

(137) Li, H.; Ilin, S.; Wang, W.; Duncan, E. M.; Wysocka, J.; Allis, C. D.; Patel, D. J. *Nature* **2006**, *442* (7098), 91–95.

(138) Liu, L.; Zhen, X. T.; Denton, E.; Marsden, B. D.; Schapira, M. *Bioinformatics* **2012**, *28* (16), 2205–2206.

(139) Wen, H.; Li, Y.; Xi, Y.; Jiang, S.; Stratton, S.; Peng, D.; Tanaka, K.; Ren, Y.; Xia, Z.; Wu, J.; Li, B.; Barton, M. C.; Li, W.; Li, H.; Shi, X. *Nature* **2014**, *508* (7495), 263–268.

(140) Simhadri, C.; Daze, K. D.; Douglas, S. F.; Quon, T. T. H.; Dev, A.; Gignac, M. C.; Peng, F.; Heller, M.; Boulanger, M. J.; Wulff, J. E.; Hof, F. *J. Med. Chem.* **2014**, *7*.

(141) Oliver, S. S.; Musselman, C. A.; Srinivasan, R.; Svaren, J. P.; Kutateladze, T. G.; Denu, J. M. *Biochemistry* **2012**, *51* (33), 6534–6544.

(142) Lessard, J.; Wu, J. I.; Ranish, J. a; Wan, M.; Winslow, M. M.; Staahl, B. T.; Wu, H.; Aebersold, R.; Graef, I. a; Crabtree, G. R. *Neuron* **2007**, *55* (2), 201–215.

(143) Ali, M.; Yan, K.; Lalonde, M.-E.; Degerny, C.; Rothbart, S. B.; Strahl, B. D.; Côté, J.; Yang, X.-J.; Kutateladze, T. G. *J. Mol. Biol.* **2012**, *424* (5), 328–338.

(144) Dreveny, I.; Deeves, S. E.; Fulton, J.; Yue, B.; Messmer, M.; Bhattacharya, A.; Collins, H. M.; Heery, D. M. *Nucleic Acids Res.* **2013**, 1–14.

(145) Chestkov, A.; Baka, I.; Kost, M. *Genomics* **1996**, *177*, 174–177.

(146) Perez-Campo, F. M.; Costa, G.; Lie-A-Ling, M.; Kouskoff, V.; Lacaud, G. *Immunology* **2013**, *139* (2), 161–165.

(147) Yang, X.-J.; Ullah, M. *Oncogene* **2007**, *26* (37), 5408–5419.

(148) Wei, M.; Liu, B.; Su, L.; Li, J.; Zhang, J.; Yu, Y.; Yan, M.; Yang, Z.; Chen, X.; Liu, J.; Lv, X.; Nie, H.; Zhang, Q.; Zheng, Z.; Yu, B.; Ji, J.; Zhang, J.; Zhu, Z.; Gu, Q. *Mol. Cancer Ther.* **2010**, *9* (6), 1764–1774.

(149) Li, C.; Nie, H.; Wang, M.; Su, L.; Li, J.; Yu, B.; Wei, M.; Ju, J.; Yu, Y.; Yan, M.; Gu, Q.; Zhu, Z.; Liu, B. *Cancer Lett.* **2012**, *320* (2), 189–197.

(150) Baell, J. B.; Holloway, G. A. *J. Med. Chem.* **2010**, *53* (7), 2719–2740.

(151) PerkinElmer. AlphaLISA and AlphaScreen No-Wash Assays http://www.perkinelmer.co.uk/Resources/TechnicalResources/ApplicationSupportKnowledgebase/AlphaLISA-AlphaScreen-no-washassays/AlphaLISA-AlphaScreen-no-wash-assays.xhtml (accessed Sep 28, 2015).

(152) Wells, J. A.; McClendon, C. L. *Nature* **2007**, *450* (7172), 1001–1009.

(153) Zinzalla, G.; Thurston, D. E. *Future Med. Chem.* **2009**, *1* (1), 65–93.

(154) Scott, D. E.; Ehebauer, M. T.; Pukala, T.; Marsh, M.; Blundell, T. L.; Venkitaraman, A. R.; Abell, C.; Hyvönen, M. *Chembiochem* **2013**, *14* (3), 332–342.

(155) Surade, S.; Blundell, T. L. *Chem. Biol.* **2012**, *19* (1), 42–50.

(156) Murray, C. W.; Verdonk, M. L.; Rees, D. C. *Trends Pharmacol. Sci.* **2012**, *33* (5), 224–232.

(157) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. *Drug Discov. Today* **2003**, *8* (19), 876–877.

(158) Jhoti, H.; Williams, G.; Rees, D. C.; Murray, C. W. *Nat. Rev. Drug Discov.* **2013**, *12* (8), 644–645.

(159) Turnbull, A. P.; Boyd, S. M.; Walse, B. *Res. Reports Biochem.* **2014**, *4*, 13–26.

(160) Fink, T.; Reymond, J. *J. Med. Chem.* **2007**, *47*, 342–353.

(161) Morley, A. D.; Pugliese, A.; Birchall, K.; Bower, J.; Brennan, P.; Brown, N.; Chapman, T.; Drysdale, M.; Gilbert, I. H.; Hoelder, S.; Jordan, A.; Ley, S. V; Merritt, A.; Miller, D.; Swarbrick, M. E.; Wyatt, P. G. *Drug Discov. Today* **2013**, *00* (00), 1–7.

(162) Ripphausen, P.; Nisius, B.; Bajorath, J. *Drug Discov. Today* **2011**, *16* (9-10), 372–376.

(163) Friesner, R. a; Banks, J. L.; Murphy, R. B.; Halgren, T. a; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(164) Halgren, T. a; Murphy, R. B.; Friesner, R. a; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.* **2004**, *47* (7), 1750–1759.

(165) Friesner, R. a; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. a; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49* (21), 6177–6196.

(166) Zhao, Q.; Qin, L.; Jiang, F.; Wu, B.; Yue, W.; Xu, F.; Rong, Z.; Yuan, H.; Xie, X.; Gao, Y.; Bai, C.; Bartlam, M.; Pei, X.; Rao, Z. *J. Biol. Chem.* **2007**, *282* (1), 647–656.

(167) Walters, W. P.; Namchuk, M. *Nat. Rev. Drug Discov.* **2003**, *2* (4), 259–266.

(168) Schrödinger. LigPrep, 2013.

(169) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. *Nucleic Acids Res.* **2014**, *42* (D1).

(170) Fahrni, C. J.; O'Halloran, T. V. *J. Am. Chem. Soc.* **1999**, *121* (49), 11448–11458.

(171) Marley, J.; Lu, M.; Bracken, C. *J. Biomol. NMR* **2001**, *20* (1), 71–75.

(172) Cortecnet. Spectra 9-N, 98% 15N http://www.cortecnet.com/stable-isotopes/growth-media/spectra-9-n-98-15n.html (accessed Aug 17, 2015).

(173) Chaikuad, A.; Petros, A. M.; Fedorov, O.; Xu, J.; Knapp, S. *Medchemcomm* **2014**, *5*, 1843–1848.

(174) Esposito, D.; Chatterjee, D. K. *Curr. Opin. Biotechnol.* **2006**, *17* (4), 353–358.

(175) Wheeler, M. J.; Russi, S.; Bowler, M. G.; Bowler, M. W. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2012**, *68* (1), 111–114.

(176) ESRF. Calculation of equilibrium relative humidity (RH) http://www.esrf.eu/UsersAndScience/Experiments/MX/How_to_use_our_beamlines/forms (accessed Aug 19, 2015).

(177) Salem, M.; Mauguen, Y.; Prangé, T. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **2010**, *66* (Pt 3), 225–228.

(178) Yeates, T. O. *Methods Enzymol.* **1997**, *276* (1974), 344–358.

(179) Practical-Fragments. Poll results: affiliation, fragment-finding methods, and library size http://practicalfragments.blogspot.co.uk/2014/01/poll-results-affiliation-fragment.html (accessed Aug 17, 2015).

(180) Johansson, C.; Tumber, A.; Che, K.; Cain, P.; Nowak, R.; Gileadi, C.; Oppermann, U. *Epigenomics* **2014**, *6* (1), 89–120.

(181) Liu, W.; Tanasa, B.; Tyurina, O. V; Zhou, T. Y.; Gassmann, R.; Liu, W. T.; Ohgi, K. a; Benner, C.; Garcia-Bassets, I.; Aggarwal, A. K.; Desai, A.; Dorrestein, P. C.; Glass, C. K.; Rosenfeld, M. G. *Nature* **2010**, *466* (7305), 508–512.

(182) Feng, W.; Yonezawa, M.; Ye, J.; Jenuwein, T.; Grummt, I. *Nat. Struct. Mol. Biol.* **2010**, *17* (4), 445–450.

(183) Laumonnier, F.; Holbert, S.; Ronce, N.; Faravelli, F.; Lenzner, S.; Schwartz, C. E.; Lespinasse, J.; Van Esch, H.; Lacombe, D.; Goizet, C.; Phan-Dinh Tuy, F.; van Bokhoven, H.; Fryns, J.-P.; Chelly, J.; Ropers, H.-H.; Moraine, C.; Hamel, B. C. J.; Briault, S. *J. Med. Genet.* **2005**, *42* (10), 780–786.

(184) Loenarz, C.; Ge, W.; Coleman, M. L.; Rose, N. R.; Cooper, C. D. O.; Klose, R. J.; Ratcliffe, P. J.; Schofield, C. J. *Hum. Mol. Genet.* **2010**, *19* (2), 217–222.

(185) Rose, N. R.; Woon, E. C. Y.; Tumber, A.; Walport, L. J.; Chowdhury, R.; Li, X. S.; King, O. N. F.; Lejeune, C.; Ng, S. S.; Krojer, T.; Chan, M. C.; Rydzik, A. M.; Hopkinson, R. J.; Che, K. H.; Daniel, M.; Strain-Damerell, C.; Gileadi, C.; Kochan, G.; Leung, I. K. H.; Dunford, J.; Yeoh, K. K.; Ratcliffe, P. J.; Burgess-Brown, N.; von Delft, F.; Muller, S.; Marsden, B.; Brennan, P. E.; McDonough, M. a; Oppermann, U.; Klose, R. J.; Schofield, C. J.; Kawamura, A. *J. Med. Chem.* **2012**, *55* (14), 6639–6643.

(186) Yue, W. W.; Hozjan, V.; Ge, W.; Loenarz, C.; Cooper, C. D. O.; Schofield, C. J.; Kavanagh, K. L.; Oppermann, U.; McDonough, M. A. *FEBS Lett.* **2010**, *584* (4), 825–830.

(187) Højfeldt, J. W.; Agger, K.; Helin, K. *Nat. Rev. Drug Discov.* **2013**, *12* (12), 917–930.

(188) Thinnes, C. C.; England, K. S.; Kawamura, A.; Chowdhury, R.; Schofield, C. J.; Hopkinson, R. J. *Biochim. Biophys. Acta - Gene Regul. Mech.* **2014**, *1839*, 1416–1432.

(189) SGC. KDOAM25 - A Chemical Probe for JARID1 http://www.thesgc.org/node/10586 (accessed Aug 22, 2015).

(190) Hopkinson, R. J.; Tumber, A.; Yapp, C.; Chowdhury, R.; Aik, W.; Che, K. H.; Li, X. S.; Kristensen, J. B. L.; King, O. N. F.; Chan, M. C.; Yeoh, K. K.; Choi, H.; Walport, L. J.; Thinnes, C. C.; Bush, J. T.; Lejeune, C.; Rydzik, A. M.; Rose, N. R.; Bagg, E. A.; McDonough, M. A.; Krojer, T. J.; Yue, W. W.; Ng, S. S.; Olsen, L.; Brennan, P. E.; Oppermann, U.; Müller, S.; Klose, R. J.; Ratcliffe, P. J.; Schofield, C. J.; Kawamura, A. *Chem. Sci.* **2013**, *4* (8), 3110.

(191) Kruidenier, L.; Chung, C.; Cheng, Z.; Liddle, J.; Che, K.; Joberty, G.; Bantscheff, M.; Bountra, C.; Bridges, A.; Diallo, H.; Eberhard, D.; Hutchinson, S.; Jones, E.; Katso, R.; Leveridge, M.; Mander, P. K.; Mosley, J.; Ramirez-Molina, C.; Rowland, P.; Schofield, C. J.; Sheppard, R. J.; Smith, J. E.; Swales, C.; Tanner, R.; Thomas, P.; Tumber, A.; Drewes, G.; Oppermann, U.; Patel, D. J.; Lee, K.; Wilson, D. M. *Nature* **2012**, *488* (7411), 404–408.

(192) Rose, N. R.; Ng, S. S.; Mecinović, J.; Liénard, B. M. R.; Bello, S. H.; Sun, Z.; McDonough, M. a; Oppermann, U.; Schofield, C. J. *J. Med. Chem.* **2008**, *51* (22), 7053–7056.

(193) Cunliffe, C. J.; Franklin, T. J.; Hales, N. J.; Hill, G. B. *J. Med. Chem.* **1992**, *35* (14), 2652–2658.

(194) Rose, N. R.; Woon, E. C. Y.; Kingham, G. L.; King, O. N. F.; Mecinović, J.; Clifton, I. J.; Ng, S. S.; Talib-Hardy, J.; Oppermann, U.; McDonough, M. A.; Schofield, C. J. *J. Med. Chem.* **2010**, *53* (4), 1810–1818.

(195) Mackeen, M. M.; Kramer, H. B.; Chang, K.-H.; Hopkinson, R. J.; Schofield, C. J.; Kessler, B. M. *J. Proteome Res.* **2010**, *9* (8), 4082–4092.

(196) Lohse, B.; Nielsen, A. L.; Kristensen, J. B. L.; Helgstrand, C.; Cloos, P. a C.; Olsen, L.; Gajhede, M.; Clausen, R. P.; Kristensen, J. L. *Angew. Chemie - Int. Ed.* **2011**, *50* (39), 9100–9103.

(197) Woon, E. C. Y.; Tumber, A.; Kawamura, A.; Hillringhaus, L.; Ge, W.; Rose, N. R.; Ma, J. H. Y.; Chan, M. C.; Walport, L. J.; Che, K. H.; Ng, S. S.; Marsden, B. D.; Oppermann, U.; McDonough, M. A.; Schofield, C. J. *Angew. Chemie - Int. Ed.* **2012**, *51* (7), 1631–1634.

(198) Upadhyay, A. K.; Rotili, D.; Han, J. W.; Hu, R.; Chang, Y.; Labella, D.; Zhang, X.; Yoon, Y.-S.; Mai, A.; Cheng, X. *J. Mol. Biol.* **2012**, *416* (3), 319–327.

(199) Suzuki, T.; Ozasa, H.; Itoh, Y.; Zhan, P.; Sawada, H.; Mino, K.; Walport, L.; Ohkubo, R.; Kawamura, A.; Yonezawa, M.; Tsukada, Y.; Tumber, A.; Nakagawa, H.; Hasegawa, M.; Sasaki, R.; Mizukami, T.; Schofield, C. J.; Miyata, N. *J. Med. Chem.* **2013**, *56* (18), 7222–7231.

(200) Kawamura, A.; Tumber, A.; Rose, N. R.; King, O. N. F.; Daniel, M.; Oppermann, U.; Heightman, T. D.; Schofield, C. *Anal. Biochem.* **2010**, *404* (1), 86–93.

(201) PDB. Protein Data Bank - 1WEP Solution structure of PHD domain in protein AA017385 (accessed Aug 27, 2015).

(202) McLean, L. R.; Zhang, Y.; Li, H.; Li, Z.; Lukasczyk, U.; Choi, Y. M.; Han, Z.; Prisco, J.; Fordham, J.; Tsay, J. T.; Reiling, S.; Vaz, R. J.; Li, Y. *Bioorganic Med. Chem. Lett.* **2009**, *19* (23), 6717–6720.

(203) Herlev, T. K. H.; Malov, B. P.; Knud Erik Andersen, S. Compounds With Growth Hormone Releasing Properties. 5919777, 1999.

(204) Mehta, A.; Jaouhari, R.; Benson, T. J.; Douglas, K. T. *Tetrahedron Lett.* **1992**, *33* (37), 5441–5444.

(205) Gerhart, F.; Higgins, W.; Tardif, C.; Ducep, J. B. *J. Med. Chem.* **1990**, *33* (8), 2157–2162.

(206) Delon, L.; Laurent, P.; Blancou, H. *J. Fluor. Chem.* **2005**, *126* (11-12), 1487–1492.

(207) Billington, A.; Clayton, N. M.; Giblin, G. M. P.; Healy, M. P. Benzo [f] Isoindol-2-ylphenyl Acetic Acid Derivatives as EP4 Receptor Agonists. WO 2007088190 A1, 2007.

(208) Trost, B. M.; Bunt, R. C.; Lemoine, R. C.; Calkins, T. L. *J. Am. Chem. Soc.* **2000**, *122* (25), 5968–5976.

(209) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner*; Springer, 2007.

(210) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. *J. Comput. Aided. Mol. Des.* **2007**, *21* (12), 681–691.

(211) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *J. Appl. Crystallogr.* **2007**, *40* (4), 658–674.

(212) Faidallah, H. M.; Khan, K. A. *J. Fluor. Chem.* **2012**, *142*, 96–104.

(213) Gloge, A.; Zoń, J.; Kövári, A.; Poppe, L.; Rétey, J. *Chemistry* **2000**, *6* (18), 3386–3390.

(214) Schiller, P. W.; Weltrowska, G.; Nguyen, T. M. D.; Lemieux, C.; Chung, N. N.; Marsden, B. J.; Wilkes, B. C. *J. Med. Chem.* **1991**, *34* (10), 3125–3132.

(215) Arava, V. R.; Amasa, S. R.; Goud Bhatthula, B. K.; Kompella, L. S.; Matta, V. P.; Subha, M. C. S. *Synth. Commun.* **2013**, *43* (21), 2892–2897.