Conformational selection mechanism of ASHH2 methyltransferase CW domain recognising H3K4me1 histone modification

Maxim Bril'kov

UNIVERSITY OF BERGEN

Conformational selection mechanism of ASHH2 methyltransferase CW domain recognising H3K4me1 histone modification

Maxim Bril'kov

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 26.10.2020

Year: 2020

Title: Conformational selection mechanism of ASHH2 methyltransferase CW domain recognising H3K4me1 histone modification

Name: Maxim Bril'kov

Print: Skipnes Kommunikasjon / University of Bergen

«There is no royal road to science, and only those who do not dread the fatiguing climb of its steep paths have a chance of gaining its luminous summits.»

Karl Marx

Scientific environment

The work presented in this dissertation was carried out at the Faculty of Mathematics and Natural Sciences at The Department of Biological Sciences (formerly The Department of Molecular Biology) between April 2015 and June 2019. For the whole period the project was supervised by Professor Øyvind Halskau and co-supervised by Dr. Olena Dobrovolska and Professor Rein Aasland. The project was indirectly supported through the Norwegian NMR Platform, NNP (Norvegian Research Council infrastructure grant 226244/F50). Access to PhD-level courses was supported by The Norwegian Biochemical Society, BioStruct and BioCat research schools (various travel grants to Bril'kov) and the Molecular and Computational Biology Research School.

Acknowledgements

First of all, I want to thank my supervisor Øyvind Halskau. I have learnt a lot from you and always felt motivated after leaving your office. Thanks to my co-supervisors Olena Dobrovolska and Rein Aasland for discussions around the thesis work and for giving ideas on approaching the problems.

Thanks to people at the department who maintained the labs functional, especially Diana Turcu for the help with instruments and experimental set ups. Thanks to the staff at the Norwegian NMR Platform (NNP) for supporting the project.

My special thanks to my colleagues and fellow students with whom I could discuss the project issues and always get advises on how to solve them. Big thanks to Martin Jakubec, Morten Govasli and Øyvind Strømland. My biggest thanks here is to Kirill Jefimov who showed me how to work hard and how to balance this hard work with proper procrastination. Thank you for all those endless chromatin talks, hikes and trips.

Maxim Bril'kov

Bergen, July 4th, 2019

Table of Contents

List of Publications

Paper I ^{1}H, ^{13}C, and ^{15}N resonance assignment of CW domain of the N-methyltransferase ASHH2 free and bound to the mono-, di- and trimethylated histone H3 tail peptides

Dobrovolska O., **Bril'kov M**., Ødegård-Fougner Ø., Aasland R., & Halskau Ø.

Biomolecular NMR Assignments, 12:215-220, 2018

Paper II   The *Arabidopsis* (ASHH2) CW domain binds monomethylated K4 of the histone H3 tail through conformational selection

Dobrovolska O., **Bril'kov M.**, Madeleine N., Ødegård-Fougner Ø., Strømland Ø., Martin S.R., Marco V.D., Christodoulou E., Teigen K., Isaksson J., Underhaug J., Reuter N., Aalen R.B., Aasland R., & Halskau Ø.

FEBS J, published online ahead of print, 2020 Feb 21 DOI: https://doi.org/10.1111/febs.15256

Paper III   Binding specificity of ASHH2 CW-domain towards H3K4me1 ligand is coupled to its structural stability through its α1-helix

Bril'kov M.S., Dobrovolska O., Ødegård-Fougner Ø., Strømland Ø., Aasland R., & Halskau Ø.

Manuscript, 2020

Selected abbreviations

PTM, post translational modification;

H3K4me1, Histone H3 trimethylated at Lys-4;

H3K36me3, Histone H3 trimethylated at the Lys-36;

ITC, isothermal titration calorimetry;

MD, molecular dynamic simulation;

NMR, nuclear magnetic resonance;

HSQC, heteronuclear single quantum coherence;

CSP, chemical shift perturbation;

NOE, Nuclear Overhauser Effect;

CPMG, Carr-Purcell Meiboom-Gill Relaxation Dispersion;

SEC, size-exclusion chromatography;

MALS, multi-angle light scattering.

Abstract

Eukaryotic DNA is complexed with several different proteins and exists in the form of chromatin. Chromatin itself is a very dynamic and fluctuating structure that is regulated in accordance with the function of a cell and the different signails it receives. The basic unit of chromatin structure is a nucleosome, which consists of histone core proteins H2A, H2B, H3 and H4, which form an octamer, and the DNA winds around this core. These histone proteins are characterized by the presence of unstructured C- and N-terminal tails, which protrude out of each nucleosome. These tails are prone to different chemical modifications called post-translational modifications (PTM). Different PTM patterns have different effect on associated genes and lead either to gene activation or gene silencing. The pattern of PTMs is referred to as the histone code. Maintenance of the code is promoted by different enzymes that are able to "read" and "edit" these modifications, contributing to gene regulation.

ASHH2 methyltransferase is a plant (*Arabidopsis thaliana*) protein that involved in regulation of more than 1000 genes. It has several domains that function in reader-editor combination. By its CW domain ASHH2 selectively recognize monomethylated modification at the 4th lysine of H3 histone tail (H3K4me1) and by its SET methyltransferase domain it transfers a methyl group to H3K36 position and activates associated genes by H3K36me2/me3 modifications.

The CW domain is also present in other protein families with different functions related to chromatin remodelling and gene regulation. But the CW from those families was shown to be more selective towards H3K4me3 modification.

As for now, there is no consensus in what determines and selectivity of the CW domains and how it is regulated. The overall objective of the thesis is to explore the question of selectivity mechanism of the CW domain from ASHH2 methyltransferase, which is unique in preference towards H3K4me1 modification. Using comparative structural approach supplemented with biophysical analysis of interacting proteins, the work aimed to solve the structure of CW domain in complex with H3K4me1 peptide,

assess the protein's internal dynamics and characterize the interaction mechanism thermodynamically.

Paper I is an analysis and description paper documenting NMR spectra assignments of CW in free state and bound to peptides, which form the foundation of the NMR work in paper II and paper III.

Paper II summarizes the results of solved structure and dynamics evaluation. The main conclusion is that CW recognizes its ligand by a conformational selection mechanism, as it appears as a very flexible protein, even when it is bound to the ligand. Upon ligand binding, CW undergoes compaction and stabilization. Stabilization of the complex structure is mediated by two unstructured coils flanking the C-terminal α1-helix. Molecular dynamic simulation, performed in comparison with the previously published X-ray structure, suggested also β-augmentation by the ligand.

In Paper III, biophysical aspects of the interaction mechanism and the domain's specificity were investigated, and the results showed that the interaction with the ligand is enthalpy driven and that the specificity towards H3K4me1 arises from different balances of enthalpy-entropy contributions when the domain interacts with H3K4me1/2/3 peptides. Mutation analysis led to the conclusion that the C-terminal α1-helix maintains the fold of the domain through residues I915 and L919 oriented towards the hydrophobic tryptophan binding pocket. Point mutations resulted in the fold disturbances that affect the interaction in a way that it was reduced or lost, rather than in an overall change in the specificity of the domain. As a conclusion, selectivity was linked to stability of the fold maintained by α1-helix, and energetically favourable conformation of the CW-H3K4me1 complex, versus especially the H3K4me3 complex.

1. INTRODUCTION

1.1 DNA and chromatin organization

DNA is a central biomolecule that carries and ensures the inheritance of genetic information. Within eukaryotic cells, DNA exists as an association with proteins that help to organize it into structures called chromosomes. Such organization to a compact form is necessary, primarily due to limited space in the nucleus. Second, it provides stability to the DNA molecule, protecting it from damage. And, lastly, a chromosome is an efficient way to transmit genetic information during cell division [1-3].

Prior to the formation of chromosomes, DNA is organized and compacted via different intermediate structures with help of associated special proteins. These proteins and DNA complex at a given genomic region is referred to as chromatin. The basic unit of chromatin is a nucleosome, which represents "a reel" with DNA wound around an octamer of eight core histone proteins. The octamer is formed by two copies of each of H2A, H2B, H3 and H4 proteins. This level of organization is often referred to as "beads on string". There is also the H1 histone protein, which is involved in next level of chromatin organization, the so-called chromatin fiber, by sealing the nucleosomes from outside, and referred to as a linker histone [3-5]. Different levels of chromatin organization and nucleosome structure are depicted in Figure 1 and 2. Beside histone proteins, there are numerous proteins that are involved in not only maintenance of chromatin structure, but also DNA-binding proteins that regulate replication, transcription and recombination, and are involved in repair mechanisms [6-9].

During the cell cycle, a cell utilizes different sets of genes by turning them on and off and regulating their expression, depending on the function of a cell or external signals and stimuli the cell receives [9,10]. Such alteration in the use of genetic information requires extensive control of compartmentalization, structuration and dynamical reorganization of chromatin in the nucleus, condensing the genomic regions which are not in use, forming heterochromatin, and allowing active transcription of genes located in the disordered regions of the genome, named euchromatin [2,5,11].

1.2 General mechanism of transcriptional regulation in eukaryotes

Transcription in eukaryotes is characterized by advanced interplay of different genomic regions. One region where the transcription machinery is getting assembled is called the promoter. Regions of individual binding sites for regulatory proteins comprise regulatory sequences and are often grouped in units called enhancers. Regulatory sequences can be located at a great distance from promoters and because of that DNA has to form a loop, bringing them together. That requires other regulatory features such as insulators and boundary elements to ensure that a given enhancer regulates only one gene out of several genes in its vicinity [12-16].

Regulation of transcription is done by proteins that bind to specific DNA sequences and switch the genes on and off. Thus, transcriptional regulators can be divided into activators and repressors.

Eukaryotic activators have a DNA-binding site and activating regions that activate transcription by recruiting protein complexes to genes. That can be recruitment of transcription factors (the so-called Mediator complex), recruitment of nucleosome modifiers to "clear" the promoters that they encave, and recruitment of factors that stimulate Polymerase II initiation and elongation [17-20].

As the eukaryotic DNA is present in the form of nucleosomes, regulatory sequences can be obstructed from recognition by regulatory proteins and transcription machinery. "Clearance" of these regions is carried out by the enzymes that either modify histone proteins in the core of a nucleosome and change how tight DNA is associated with them, or by the enzymes that are able to "remodel" the nucleosomes by utilizing energy from ATP-hydrolysis and displace nucleosomes exposing the regulatory sequences [21-24].

Transcriptional repressors work in the way opposite to the activators. Firstly, they can compete directly with activators for a binding site, blocking the binding of an activator. Repressors can also inhibit an activator or directly inhibit the activity of the whole transcription machinery. Finally, they can recruit histone modifiers to compact the

chromatin or remove modifications that are recognized by transcription machinery [25-27].

Beside the modification of histone proteins to regulate transcription, the DNA molecule itself can be methylated and silence genes by inhibiting the binding of the transcription complex or its activators [28-32]. Methylation and demethylation of the DNA is mediated by enzymes called DNA methyltransferases and DNA demethylases [33-35].
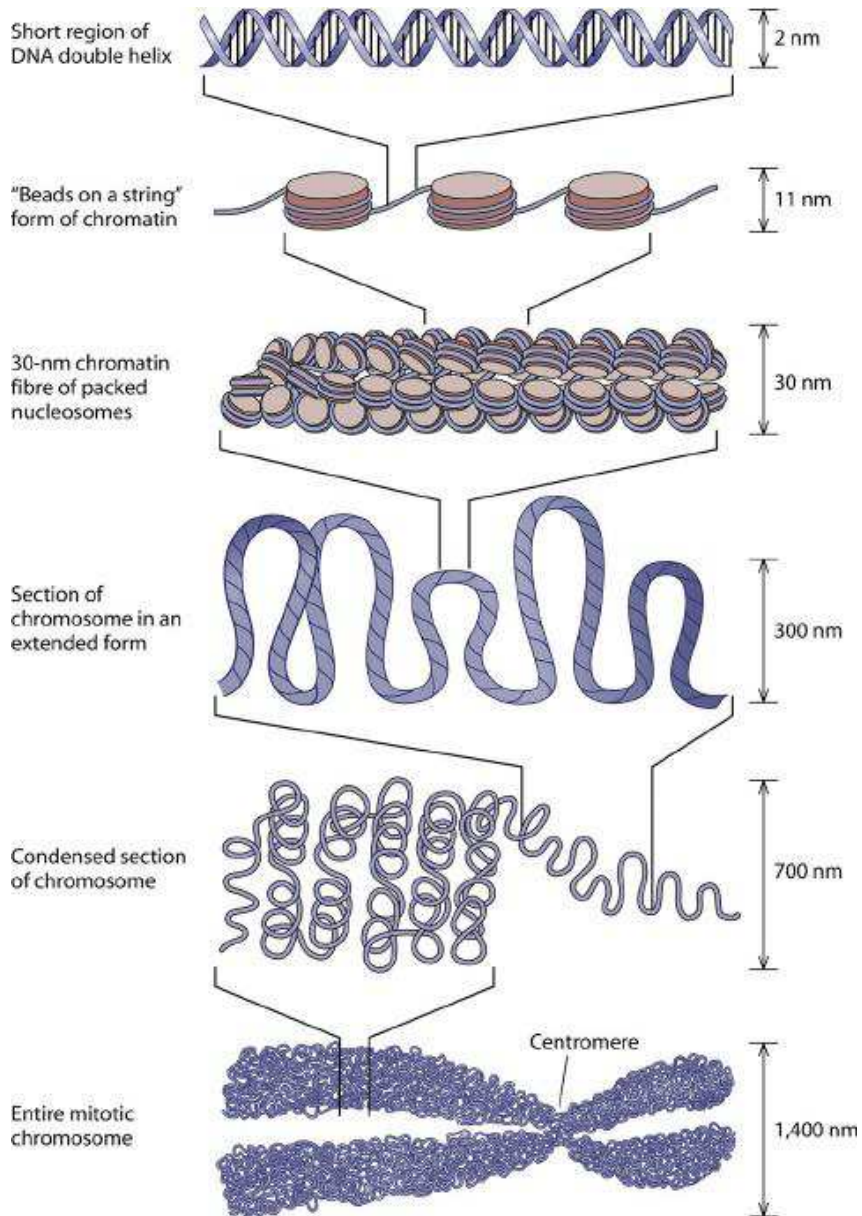
Figure 1. Levels of chromatin organization. The DNA helix is wrapped around a histone octamer and form nucleosomes that further fold into a structure referred to as the 30-nm fiber. This fiber structure is organized into higher order structures with formation of chromosomes at the end. Figure from Jansen *et al*. [3].

1.3 Epigenetic gene regulation

The core histones are characterized by structural disorder of their C and N-terminal tails, which direct the DNA winding around the histone octamer upon formation of a nucleosome [4,5]. These unstructured tails protrude out of a nucleosome, that makes them prone to different enzymatically controlled covalent modifications, named post-translational modifications (PTM). These modifications can be methylation, acetylation, phosphorylation, ubiquitination and more, and they have different effects on the associated chromatin. For example, acetylation of lysines at H3 histone at positions 9 (H3K9) or 14 (H3K14) will lead to activation of associated genes by promoting the formation of euchromatin. Methylation at H3K4 and H3K36 is also associated with active expression, while methylation at H3K9 and H3K27 are the marks of a repressed state of chromatin with tightly packed DNA. Patterns of PTMs and their associated functions are conserved among different species of organisms and can be inherited. These patterns of modification are often referred to as the histone code, and the inherited patterns of gene expression are called epigenetic regulation [36-40]. Histones with modification positions are shown in Figure 2.

Maintenance of the histone code is based on the variety of proteins which are able to recognize and modify the PTM state of the chromatin, playing an important role in regulation of gene expression. The functionality of the histone code maintenance is divided between protein domains that can "read" and "edit" the histone modifications. Thus, there are "reader" domains, such as bromodomains, chromodomains or PHD fingers, which can differentiate and selectively recognize certain PTMs on histone tails [41-44]. "Editors" are represented by different transferases, demethylases, deacetylases, etc. domains, which are able to transfer a modification to a histone tail residue or remove it [45-48]. An example of methyltransferases is the SET-domain family that transfers methyl groups to lysines of histone and non-histone proteins [49-52]. Domains which are able to transfer or remove acetyl groups from histones are called histone acetyltransferses (HATs) and histone deacetylases (HADs) [53-55]. These "reader" and "editor" domains usually work in combinatorial mode, where a "reader" domain recognizes a certain PTM mark and brings a functional "editor" to

alter the state of a histone protein allowing regulation of associated genes. Composition and presence of such "readers" and "editors" is specific to a protein family and conditions a protein's function [36,56-58].
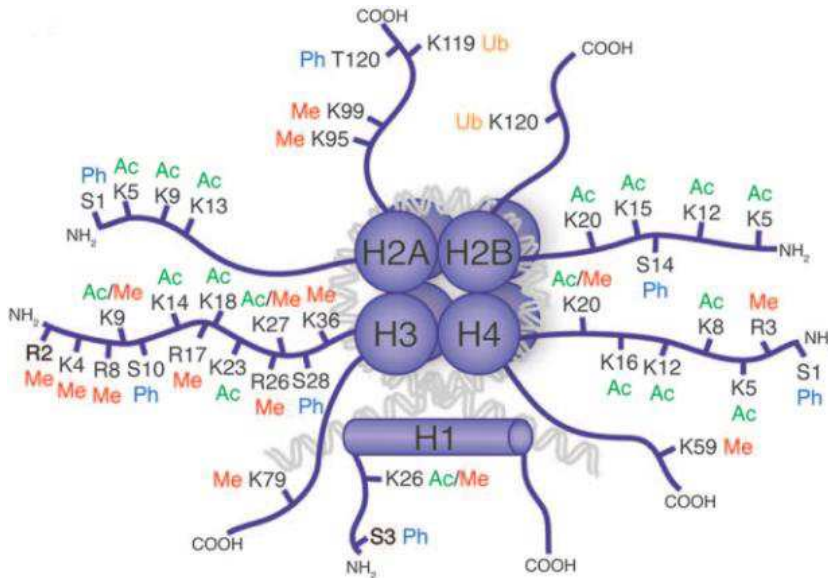
Figure 2. Nucleosome structure and post-translational modification of N and C-terminal histone tails. H2A, H2B, H3 and H4 – core histones; H1 – linker histone; Ac – acetylation, Me – methylation, Ph – phosphorylation, Ub – ubiquitination modifications. Figure from Tollervey *et al.* [40].

1.4 PTM "readers": structural aspects, function and selectivity

For a histone modifying enzyme to function with a certain level of precision, its "reader" domain needs to be selective to a specific histone tail modification, to be able to position the protein at the required location. The specificity of the interaction arises from the structure of the "reader" and its mechanisms of interaction and recognition of specific PTM. The examples of such domains are Bromodomains, Chromodomains, PHD fingers and CW.

1.4.1 Bromodomains

Bromodomains mediate recruitment of enzymes, factors or associated chromatin remodeling complexes by the recognition of acetylation modifications. They are found

in many multidomain proteins which constitute chromatin remodeling complexes and transcriptional regulators. Bromodomain containing proteins are represented in humans by eight protein families with different functions [59,60]. Among them, for example, is the ASH1L methyltransferase which acts as a transcription factor [48]. p300 and CREBBP are acetyltransferases which are able to modify histone proteins and regulate chromatin structure, and play a role in DNA replication, damage signaling and repair processes [61,62]. BRD1, BRD2, BRD3, BRD4, BRDT are examples of transcription factors which recruit transcription elongation complexes, and also associated with assembling and mediating the SWI/SNF chromatin remodeling complexes [59].

All bromodomains share conserved structural organization of four α-helixes (Figure 3A), designates as αA, αB, αC and αZ, linked by loops with various lengths and amino acid sequences. The loop between αZ and αA helixes is called the ZA-loop and the loop between αB and αC is called the BC-loop. Structural studies explain how bromodomains are able to recognize specific acetylation marks. A hydrophobic binding pocket is formed by the extension of the ZA and BC loops of the cavity formed by four α-helixes. The acetylated side chain of lysine residue of a peptide is inserted into the binding pocket, forming an extensive interaction network of hydrogen bonds and Van der Waals interaction [46,48,59,63-65].

1.4.2 Chromodomains

Chromodomains were the first methylated histones recognition domains to be characterized and are also constituents of multidomain proteins with particular functions and domain organization. Chromodomains are shared among different protein families with functions related to activation or silencing of gene expression. Proteins with these domains can be divided into three subclasses: proteins which have chromodomains at their N-terminal and C-terminal chromo shadow domains; proteins that have single chromodomain and proteins with paired or tandem chromodomains [66,67].

Heterochromatin protein 1 (HP1) is an example of the first of the three classes. HP1 was originally characterized in Drosophila and then identified in animals and yeasts but not in plants. Proteins in the HP1 family have a chromodomain at their N-terminal and a C-terminal chromo shadow domain, connected by a linker region of variable length [68]. HP1 proteins are able to recognize H3K9 methylated marks by its chromodomains, and are involved in gene silencing and organization of heterochromatin [43,69].

The chromodomain of Polycomb proteins has specificity to methylation marks at H3K27 [70]. The Polycomb group proteins constitute polycomb repressive complexes (PRC1 and PRC2), which are involved in gene silencing and regulation of high-order chromatin organization during cell differentiation and hematopoiesis [71-73].

In contrast, MOF and MSL3 are the proteins of the male-specific lethal (MSL) group of proteins and are able to bind non-coding roX2 RNA by their chromodomains, upon formation of a X-chromosome-associated dosage compensation complex (DCC). In Drosophila, to compensate for the lack of the second X chromosome in males, the DCC is functioning to increase transcription from the male X chromosome [74-76].

The conserved structure of the chromodomains is characterized by formation of tree-stranded anti-parallel β-sheets and a C-terminal α-helix (Figure 3B). Binding mechanism for these domains is conditioned by the formation of an aromatic binding pocket in which a methylated lysine residue is inserted. The rest of the peptide sequence binds as a β-strand in the conserved groove and promotes the formation of so-called β-sandwich structure. In contrast to acetylation, methylation modification does not neutralize the positive charge of the modified residue. This positive charge on methylated lysine is in π-cation interaction with tryptophan side chain in the aromatic cage [77,78].
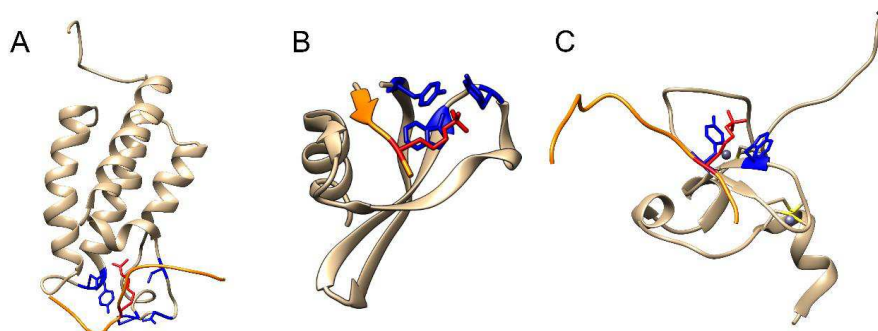
Figure 3. Representative structures of: A – Bromodomain (bromodomain of PCAF p300/CBP-associated factor in complex with H3K9ac, PDB: 2RNW [79]); B – Chromodomain (HP1 chromodomain in complex with H3K9me1, PDB: 1KNE [80]; and C – PHD-finger (BPTF PHD finger in complex with H3K4me3, PDB: 2FUU [81]). Residues in the binding pockets are highlighted in blue; bound peptide colored in orange with residue that is being recognized colored in red.

1.4.3 PHD fingers

Plant homeodomain (PHD) fingers are another well-studied example of epigenetic readers with selectivity towards methylated and unmodified H3 histone. Just like the other "reader" domains, PHD fingers are shared among different protein families with different functions related to activation or repression of transcription. PHD fingers of ING protein family selectively recognize H3K4me3 modification, bringing acetyltransferase or deacetylase domains to modify chromatin [41,42,82,83].

PHD fingers are found not only in proteins with certain catalytic activities, but also can act as scaffold proteins for macromolecular chromatin remodeling complexes. The CHD4 ATPase is a subunit of the NURD nucleosome remodeling complex. CHD4 has two PHD fingers, and one of them was shown to recognize H3K9ac and H3K9me modifications [84,85]. The BPTF transcription factor is a subunit of the NURF nucleosome remodeling factor, functioning by H3K4me3 recognition by its PHD finger [81]. The DPF3b protein functions in association with the BAF chromatin remodeling complex and was shown to specifically recognize acetylation modifications of H3 histones at 14th lysine (H3K14ac) [86-88].

PHD fingers show low sequence similarity but have a highly conserved fold. Their structure is maintained by two Zn2+ ions coordinated by conserved cysteine residues. Two anti-parallel β-sheets in the fold of the domains scaffold the aromatic binding pocket and promote interaction with histone peptides (Figure 3C). Mechanism of recognition and specificity is shaped by conserved structural features, forming binding sites. The side chain of a methylated lysine residue of a histone peptide is inserted in the aromatic binding pocket formed by aromatic sidechains. The aromatic binding cage might be formed by two to four residues which promote the formation of π-cation, hydrophobic and Van der Waals interactions. Bound peptides form an additional anti-parallel β-strand by β-augmentation. In contrast, PHD finger domains, which are able to recognize unmodified H3K4 peptides, lack the aromatic binding pocket. Interaction with unmodified peptides is mediated by a cluster of acidic residues, forming hydrogen bonds and salt bridges with the inserted side chain of a lysine residue. In the case of a binding pocket, which is selective to acetylation modification (H3K14ac), it also consists of aromatic and charged residues [81,83-86,88].

1.4.4 CW domains

The CW domain family was initially described by Perry and Zhao, 2003 as a four-cysteine zinc-finger motif, and shown to be shared among vertebrates, vertebrate-infecting parasites and higher plants. Like Bromodomains, Chromodomains and the PHD-fingers, the CW appears in different protein families. The name of the domain is derived from its conserved cysteines and tryptophans at specific positions.

Proteins that contain the CW domain were identified to be involved in epigenetic regulation and chromatin remodeling, where CW selectively recognizes methylation modifications at the histone H3 lysine 4 (Figure 4A). Among them, for example, is plant ASHH2 methyltransferase, which regulates gene expression by Histone H3 trimethylation at Lys-36 (H3K36me3) [89,90]. ZmMBD101 protein, which belongs to the MBD protein family, is involved in maintenance of the repressed state of the *Mutator* genes and protects plant genomes from harmful mutations induced by transposons. The function of the CW domain in this context is yet unclear [91]. The CW domain of the MORC3 family of ATPase chromatin remodelers recruits the

MORC3 to the chromatin and negatively regulates ATPase activity [92,93]. CW is present in the ZCWPW1 and ZCWPW2 PWWP-domain containing proteins which can recognize H3K4me3 and H4K20 methylation marks, but the function of these proteins is yet to be understood [94-96]. CW containing LSD2/AOF1/KDM1B protein, which belongs to the amine oxidase family, functions as transcriptional co-repressor by demethylation of mono- and dimethyl H3K4 marks. The CW domain in its context appears to be inactive as it was shown to be sterically inaccessible and not able to bind histone tails, but it still contributes to the overall structural stability and regulates the activity of the enzyme and its association with mitotic chromosomes [96-99].

Sequence alignment of CW from different protein families shows a high degree of variability with only a few conserved regions (Figure 4B). Structure of the domain is characterized by the formation of two antiparallel β-sheets, which scaffolds a π-electron based binding pocket formed by two conserved tryptophan side chains. Cysteines coordinate a Zn^{2+} ion in the core of the domain, maintaining its fold. Sequence variability is translated into disordered loops that differ for CW homologs (Figure 4C) [90,95,100].
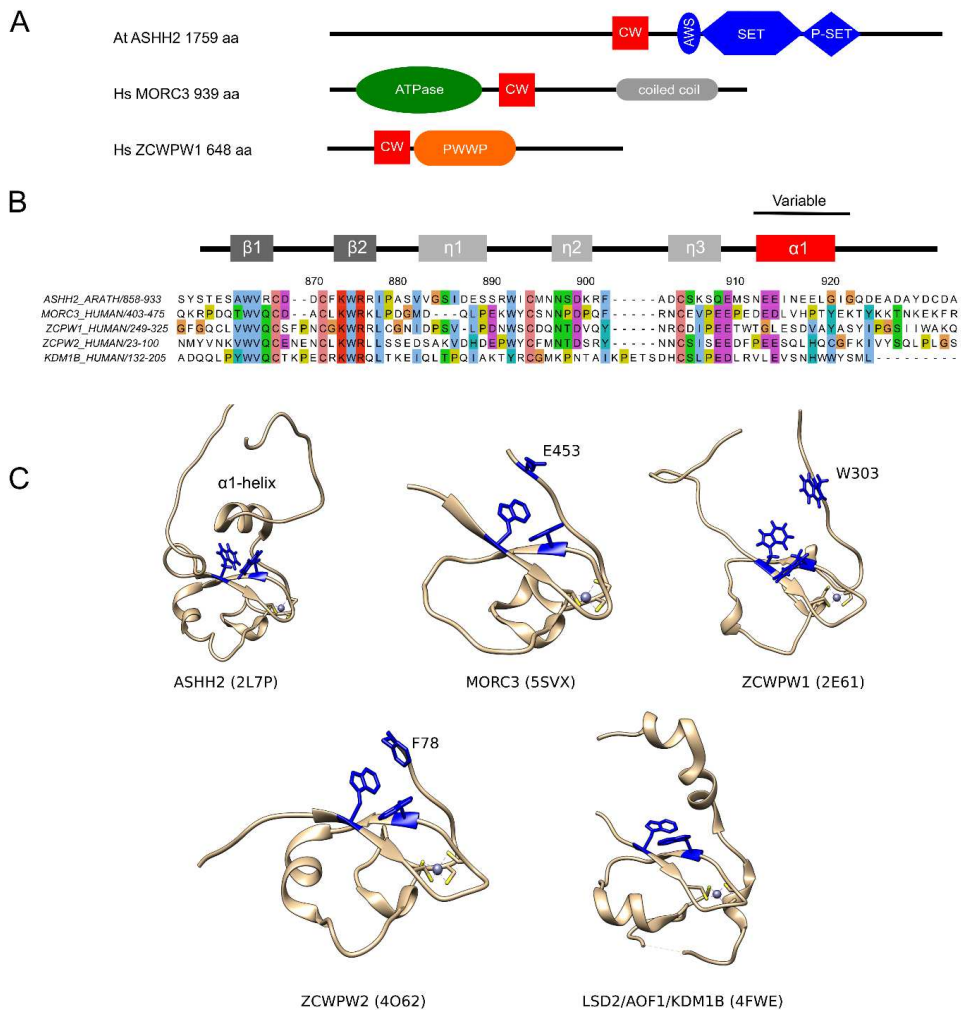
Figure 4. Comparison of CW domain containing proteins and CW structures. A – domain organization of CW containing proteins: Arabidopsis ASHH2, Human MORC3 and ZCWPW1; B – sequence alignment of different CW domains. Structural features are indicated, C-terminal region is variable; C – representative structures of CW-domains from ASHH2, MORC3, ZCWPW1, ZCWPW2 and LSD2 proteins. Sequence alignment was prepared using Jalview software with ClustalO algorithm with default parameters, Clustalx coloring scheme was used. Graphics were prepared using UCSF Chimera software and pdb files of domains in unbound state.

1.5 Selectivity of CW domains

CW domains from different protein families show different preference towards methylated states of the 4th lysine of the H3 histone. Even though CW can interact with mono- di- and trimethylation modifications, CW domain from ASHH2 methyltransferase was shown to be more specific to H3K4me1 [90,101], and CW from MORC3, ZCWPW1, ZCWPW2 proteins bind stronger to H3K4me3 mark [92,93,95,96].

The mechanism of selective differentiation of a histone modification for various CW specificity is not well understood. Comparison of known structures of the domains from different protein families shows structural differences which might shape and regulate specificity of the domains. Variable C-terminal regions upstream of the CW domain motif might be involved in regulation of specificity towards ligand of different methylation states (Figure 4B, C). The C-terminal end of ZCWPW1 posses a tryptophan residue (Trp303). It finalizes the binding pocket when the ligand is bound and conditions affinity [95]. Its homolog ZCWPW2 has a phenylalanine residue at the same position (Phe78), which serves a similar function and possibly contributes to selectivity between methylation states of H3K4 [96]. CW domain from MORC3 proteins has glutamic acid (Glu453) at this region, which also finalizes the binding pocket upon interaction with a ligand and contributes to binding to H3K4 di- and trimethylated ligands [92,93,96]. CW domains of ASHH2 and LSD2 (other names are AOF1 and KDM1B) are examples of helical structures at the C-terminal region. For LSD2, it was shown by Zhang *et al*., 2013 that the CW domain is lacking any binding activity due to various structural hindrances caused by the neighboring SWIRM and C4H2C2 zinc finger domains, and by electrostatic repulsion of the positively charged histone tail [97]. In ASHH2, the α1-helix is located above the tryptophan binding pocket (Figure 4C). Hoppmann *et al*., 2011 have shown that removal of this helix abolishes the binding activity of the domain [90]. Later Liu and Huang, 2018 have reported Ile915, Asn916 and Leu919 residues from this helix to be the key in determining specificity towards the monomethylated state [101].

1.6 ASHH2 methyltransferase of *Arabidopsis thaliana*

A gene encoding methyltransferase ASHH2 (other names are SDG8, EFS and CCR1) was characterized in a small flowering plant *Arabidopsis thaliana*, with non-redundant role associated with regulation of flowering time, branching, hormone response and other, controlling expression of more than 1000 genes [89,90,102-104]. The primary function of the enzyme is considered to be transferring di- and trimethylation modifications to H3K36 residue, a modification which is associated with increased level of gene expression [89]. The enzyme is 1759 amino acids long and consists of the CW domain followed by AWS, SET and Post-SET domains (Figure 4A). Methyltransferase activity is pertained to the SET-domain. ASHH2 recognizes the H3K4me1 modifications by CW. This interaction brings the SET-domain to the vicinity of the histone it acts upon and allows modification of H3K36 residue, which leads to a subsequent increase in expression of associated genes [90]. Loss-of-function mutation in ASHH2 gene results in repressed expression of genes regulating flowering time and plant development, which correlates with a general reduction of H3K36me2/me3 chromatin modifications [102,105,106]. Function of AWS and Post-SET domains is not yet clear, but they can, conceivably, function as autoinhibitors of the methyltransferase activity of SET-domains and play a role in interaction with the H3 histone, regulating the positioning of the SET [107-109].

1.7 Theoretical models of protein binding

A protein's function is linked to its interactions with other molecules. This, in turn, is related to its fold and dynamical properties [110,111]. There are two extreme cases where a protein exists as a very rigid structure with a well-defined fold, and proteins that are missing any folding, the so-called intrinsically disordered proteins [112].

The first and probably the most intuitive model for protein binding is the one proposed by E. Fischer already in 1894, where he proposed that proteins and their ligands fit each other in a lock-and-key manner [113]. This concept implicitly assumes relatively rigid bodies interacting with each other. However, since protein function and its interaction with binding partner emerges from the fold of the protein, which is never

entirely rigid and often is highly fluctuating, it is useful to look at the binding models from a folding funnels perspective, as suggested by Ruth Nussinov research group [112,114].

Folding funnels represent the energy landscape as a function of conformational properties of a protein folding [115]. A population of denatured conformations of a protein follows the surface of a folding funnel down to the bottom to obtain a fold with minimal energy. Depending on conditions, folding can follow different paths, displaying various intermediate conformations. The shape of the bottom of a folding funnel determines the flexibility and dynamics of the protein fold. Thus, for proteins with a rigid structure, folding funnel will have a deep, well-defined minimum (Figure 5A), while for flexible proteins it will be characterized by presence of several minima or a broad continuous well, which corresponds to presence of an ensemble of conformations (Figure 5B), and the more flexible a protein is, the higher is the number of the conformers it can obtain [114,116].

Interaction of a protein with a rigid structure with its also rigid ligand is best described by the "lock-and-key" interaction model (Figure 5C). Existing as one stable conformation, such protein exhibits high specificity to its ligand with minor structural changes after binding [114,116]. Wedemayer *et al*., 1997 explored the interaction mechanism of a matured antibody in comparison with its corresponding germline antibody [117]. Analysis of X-ray crystal structures of antibodies in their free states and in complex with a hapten antigen showed small conformational changes in matured antibodies upon binding, following the "lock-and-key" type of rigid interaction. On the other hand, structural changes in germline antibodies were more significant after complexation with the hapten antigen, indicating that the interaction corresponds more to a regime where binding stabilizes new conformations in the complex.

Binding of flexible proteins, which exist as an ensemble of conformations, is better described by the "conformational selection" and "induced fit" models. In the "conformational selection" a ligand "chooses" the most favorable conformation of a protein to form a stable complex. And, in the "induced fit" model of binding, the ligand

will induce structural changes in the protein to adopt a conformation that will stabilize the complex. Figure 5C depicts scenarios for these interaction models. These two processes usually occur together, and the predominance of "conformational selection" or "induced fit" is kinetically regulated, meaning that it depends on how fast a protein exchanges between its conformations, and how fast those conformations can accept a ligand in the binding centers [118,119]. The flexible nature of such proteins allows them to interact with several ligands or other proteins. An example of these is p53, which has numerous intrinsically disordered regions crucial for its functions. The unstructured N-terminal region contains functional transactivation domains that form α-helical structures upon binding to its ligand. The C-terminal end has regions responsible for oligomerization and tetramerization, and a regulatory region that can also obtain different structures and configurations depending on the interacting partner [120-123].
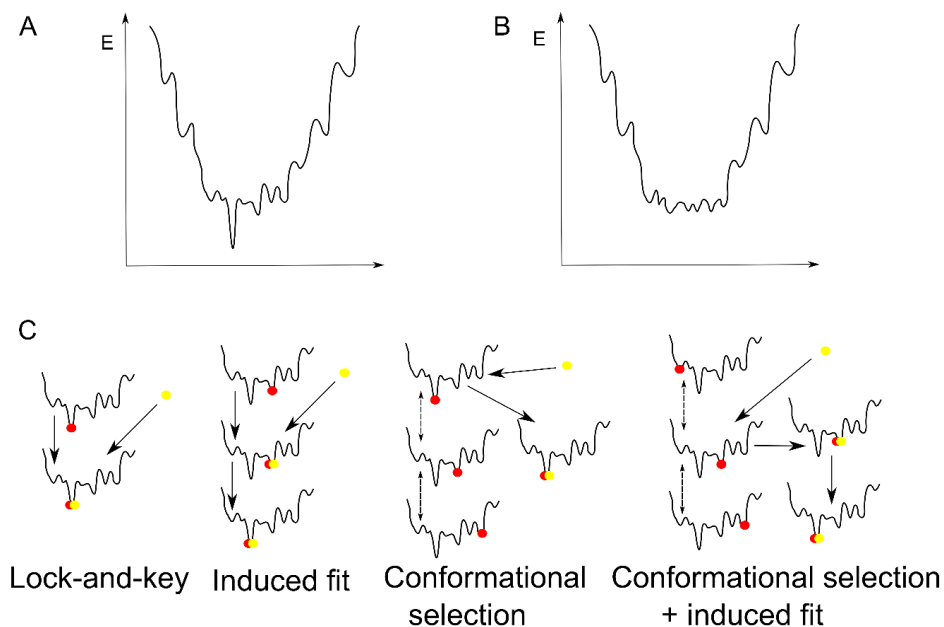
Figure 5. Schematic depiction of landscapes around the bottom of folding funnels and protein binding models. A – funnel bottom with a well-defined energy minimum; corresponds to a protein with a stable conformation and lock-and-key binding model. B – funnel with a rugged bottom and low energy barriers between the minima; corresponds to a flexible protein existing as ensemble of conformations and conformational selection binding model. C – schematic scenarios of the interaction models: Lock-and-key, where a rigid protein (red dot, ●) binds its ligand (yellow dot, ●) as the interacting surfaces match; Induced fit, where a ligand binds present conformation of a protein and induces conformational changes of the partner; Conformational selection, where a ligand binds one of the present fluctuating conformations without inducing further conformational changes; Conformational selection and induced fit, where a ligand binds one of the present fluctuating conformations of a protein and induce conformational changes in the partner. Figure adapted from Csermely *et al*. [112].

1.8 Conformational selection in structural biology. Relevance and methods

All biological processes and their regulation are based on mechanisms of molecular recognition between a ligand and its target, associated with conformational changes of the interacting partners [119,124,125]. Understanding of these mechanisms in terms of structure, energy and kinetics of interaction is crucial for efficient drugs and therapeutics development, engineering of new enzymes, as well as answering fundamental questions in biology. Different models proposed to characterize these processes were reviewed earlier in the text in Section 1.7. For this purpose, purely static models of molecular interaction were not satisfactory, and attention was shifted from Fisher's lock-and-key model towards dynamic models of binding, which try to account for conformational rearrangement of interacting receptor and ligand [126]. Combination of methods like nuclear magnetic resonance (NMR) spectroscopy, molecular dynamic simulations and molecular docking, biophysical methods like isothermal titration calorimetry (ITC) and fluorescence spectroscopy allows deep characterization of the interaction processes taking into account dynamic conformational variations [126-128].

The theoretical framework of conformational selection has contributed to advances in both technology and fundamental biology including gene regulation. For example, the presence of intrinsically disordered regions in transcription factors introduces another level of conformational flexibility that is involved in regulation of selectivity mechanisms in competitive binding [129,130]. Conceptualization of these mechanisms enabled the design of transcription factors which specifically target a given regulatory-DNA sequence of a gene, bringing the effector domain to modulate transcription [131,132]. It also allowed a search for transcription factor inhibitors that potentially could be used as new cancer therapeutics [133]. The proposed involvement of conformational selection in gene regulation and the presence of mobile loops in CW and its flanking regions in the ASHH2 protein suggests that the action of CW may require molecular insights other than those provided by lock-and-key type of mechanisms.

1.8.1 Molecular interaction and dynamics by Nuclear Magnetic Resonance spectroscopy

One of the most useful methods for studying the interaction mechanisms is Nuclear Magnetic Resonance (NMR) spectroscopy. NMR provides a vast set of applications that allow detection of changes in NMR parameters of the ligand or the target molecule, such as chemical shifts, relaxation rates, diffusion properties and intermolecular cross-relaxation [134-137]. A widely used method is the ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectrometry of ^{15}N isotopically labeled proteins in the absence and presence of a ligand. The method enables monitoring of the changes in chemical shifts of the target upon ligand binding. This information can be used for mapping of binding sites and identification of the residues that are directly or indirectly involved in the interaction [138,139].

Under a complexation of a ligand with its target molecule, the overall dynamic properties of flexible molecules tumbling in solution are expected to change. These changes can be detected and quantified by NMR methods which allow measuring relaxation rates, relaxation dispersion or exchange processes, which are directly affected by protein dynamics. The methods are sensitive over a broad time scale of motion (ps to s) Figure 6 [127,140-143].

In later years, the determination of relaxation rates of longitudinal magnetization (T_1 and R_1), transverse magnetization (T_2 and R_2) and relaxation of heteronuclear spin-spin magnetization of Nuclear Overhauser Effects (hNOEs) has been developed into a standardized tool for experimental assessment of different modes of internal protein dynamics at the ps-to-ns time scale. R_1 and R_2 parameters characterize the overall tumbling and the local flexibility of a protein backbone, and hNOE parameters describe the motion of the NH-bond vector [144-147].

For deeper characterization of protein dynamics, the internal protein motions need to be separated from its global tumbling. For this purpose, application of a so-called model-free analysis was proposed by Lipari and Szabo in 1982. In contrast to alternative approaches being used at the time, the model-free approach does not

explicitly model individual atoms in the protein being examined [148]. Rather, it interprets relaxation data, assuming statistical independence of the global motion of a protein and its internal motion [149,150]. By using the T_1, T_2 and hNOE data it allows characterization and analysis of spatial restrictions of the motion of intermolecular vectors. These motions are generalized by the order parameter S^2, which indicates the relative contribution to the relaxation from the overall molecular motion and from additional local motion. The S^2 parameter lies in the 0 to 1 range, where 1 corresponds to no local motion and 0 represents a fully flexible protein. It also allows characterization of the oscillation time for the complex motion of a given residue, described by the effective internal correlation time parameter τ_e, and calculation of conformational exchange contribution, R_{ex}, to the observed transverse relaxation rate R_2, which characterizes the rate of transition between two distinct states of a residue [145,149-153].
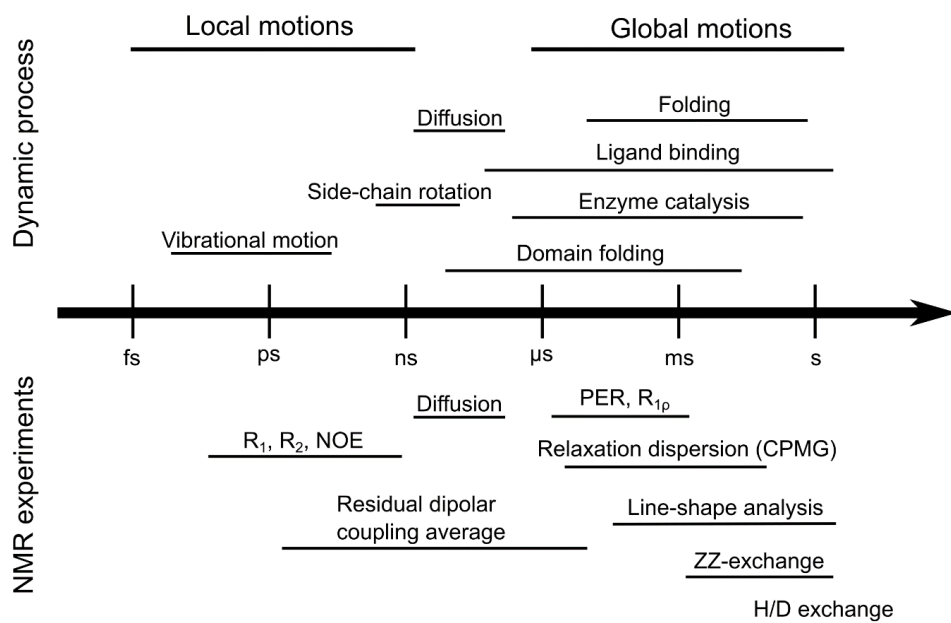
Figure 6. Time scale of protein dynamic processes and NMR methods used to study them. (PRE – paramagnetic relaxation enhancement; H/D exchange – hydrogen/deuterium exchange; ZZ-exchange also called exchange spectroscopy EXSY). Figure prepared based on Li *et al.* and Kleckner *et al.* [143,154]

While relaxation measurements with subsequent model-free analysis are used for assessment of the dynamics at ps-to-ns time scale, the relaxation dispersion (RD) method can provide information about the processes at µs-to-ms time window. This timescale includes protein secondary structure changes, loop motions and side chains reorientation, related to ligand binding and folding [155,156]. The method analyzes the peak broadening caused by conformational exchange. Named after inventors Carr Purcell Meiboom Gill (CPMG), this pulse sequence allows to quench the relaxation due to exchange and then calculate effective relaxation rates R_2^{eff} [157,158]. The dispersion profile of CPMG experiment can be fitted to dynamical models (no exchange, slow exchange and fast exchange), which parameterize the exchange between two conformational states in terms of rate, k_{ex}, populations of states and chemical shift differences between states, $\delta\omega$.

In summary, NMR provides a very versatile set of methods for exploring protein dynamics. Depending on the particular techniques being employed, NMR allows characterization of the dynamics underlying allostery, conformational fluctuations and mechanisms of conformational selection [154,158-161], mechanisms of protein folding [162], oligomerization [163], and enzymatic mechanisms [164-166]. It was also utilized for structural probing of nucleic acids [167-170].

1.8.2 Molecular dynamics by computer simulation

Computation methods such as molecular dynamic (MD) simulations can contribute to and provide information on the understanding of molecular flexibility and interaction mechanism, especially when experimental data is limited [171,172]. Molecular dynamic simulations and NMR data together can shed the light on internal motion of a protein, including the dynamics of the backbone and side chains, and characterize protein folding. For interactions, MD simulations can be used to explore conformational responses upon binding, such as the induction of secondary structure, stabilization or release of loops as well as determining intermolecular contacts between ligand and target protein [173-176]. Another advantage of molecular dynamics simulation is the possibility of modeling the systems at different temperatures to access

how the solvent affects the internal motion of proteins and predict the interaction mechanism [177-180].

At the beginning of an MD simulation, a system is initialized by creating a state with the defined types and positions of the particles. Then interaction potentials, referred to as force fields or potential energy functions, are introduced to the system by using mathematical equations that describe the interaction between the particles. For proteins, these force fields express covalent (bond length and dihedral angles) and non-covalent interactions (electrostatics, van der Waals, polar, etc) [181]. Among the most popular force field programs are AMBER and CHARMM, which are designed as a collection of codes (scripts) which work together [182-184]. The next step includes computation of forces acting on each particle of the system, based on their interaction potentials. By solving classical equations of motion, new position and velocity of the particles based on the force exerted on each particle is then calculated. After that, the system is adjusted to control the thermodynamics parameters such as temperature and pressure. By repeating this algorithm for a required number of cycles, the trajectory of the particles can be computed, and the behavior of the whole system can be tracked and characterized at the atomic level of detail [128,177,181,185].

1.8.3 Isothermal Titration Calorimetry

A system of interacting molecules is associated with changes in global thermodynamic parameters. Isothermal titration calorimetry (ITC) allows characterization of these parameters and quantification of driving forces that lead interaction in a certain direction towards equilibrium. This information can help to understand the thermodynamical basis of binding mechanisms and provide binding constants [186]. By comparison within a set of related interactions it can also answer questions of the specificity of the interacting molecules [186,187]. Direct measurement by ITC of the heat released or consumed by the interacting system allows calculation of change in free energy of the system ΔG, which is connected to equilibrium binding constant K_a; change in enthalpy ΔH, which is the heat associated with formation or breaking of non-covalent bonds during formation of the complex; and change in entropic state ΔS,

which characterize the change in degree of freedom of the system [188-191]. Interdependence of these parameters is expressed in the next equations:

$$\Delta G = -RTlnK_a \quad (1)$$
$$\Delta G = \Delta H - T\Delta S \quad (2)$$

Where R is the universal gas constant (8.314 J/mol·K), T is the temperature in degrees Kelvin, K_a is the binding constant (M^{-1}), ΔG is the Gibbs free energy (J/mol), ΔH is the enthalpy term (J/mol), and ΔS is the entropy term (J/mol·K) [187].

Enthalpy and entropy are the two components of the overall energy of the system and the balance of enthalpy-entropy contribution will define the mode of interaction. Thus, lock-and-key type of interaction is dominated by gain in solvent entropy, induced fit mechanism is enthalpy driven, and conformational selection is characterized by sequential solvent entropy gain and enthalpy decrease of the system, that dives the conformational adjustment [192-195].

1.8.4 Fluorescence spectroscopy

Structural characterization of a protein fold and its complex with a ligand can be supplemented with information that addresses questions of stability. Monitoring the fold of a protein as a function of e.g. ligand binding, temperature or denaturants is a well-established method of gaining such information [196-199]. Fluorescence spectroscopy exploits the properties of aromatic amino acids to emit light after excitation by UV-light. Tryptophan is the most popular probe as it has the highest absorptivity [200,201]. The emission peak (λ_{max}) of tryptophan is sensitive to its immediate environment that depends on solvent accessibility and protein fold. Shorter wavelengths of λ_{max} is associated with thryptophans buried inside the structure, while tryptophans that are exposed to the polar environment of the solvent are characterized by longer λ_{max} [197]. Recording emission spectra of a folded protein under a subsequent gradual increase in temperature (or increase of concentration of a denaturing compound) until it unfolds, the change in λ_{max} can be monitored [202,203]. By analyzing this data using sigmoidal curve fitting, a conclusion of protein fold stability can be made, expressed as the melting temperature T_m. Adding interacting partners to

the system can provide information on the change in stability of the formed complexes [200,204,205]. By the same principle, the changes in emission spectra can be used to monitor protein association with ligands and conformational transitions of a protein in titration series with a gradual increase in concentration of the added interaction partner [198,199,206].

1.8.5 Size measurements of proteins and their complexes in solution

In contrast to detailed characterization of molecular interaction in terms of structure, thermodynamics or different exchange constants, estimation of molecular qualities like the size of a molecule or a complex is a comparatively simple and intuitively interpretable information [207]. The changes of the size and shape of a protein and associated with them hydrodynamic coefficients can be used to address conformational variations, folding and aggregation processes and interaction with ligands [208-211]. There are many ways of determining the sizes of proteins in solution, including size exclusion chromatography (SEC), light scattering techniques and NMR diffusion.

Size exclusion chromatography enables separation of molecules based on their excluded volume and is able to separate monomers, dimers, aggregates or molecular complexes. Traveling through the SEC column, molecules with a bigger size will exit faster than smaller ones, which are retained longer within the maze-like structure of the matrix [212,213]. Techniques like static light scattering (SLS), multi-angle light scattering (MALS) and dynamic light scattering (DLS) utilize properties of molecules in a solution to scatter light. Estimation of size, molecular weight and diffusion properties can be achieved by measuring how much light was scattered, and this directly correlates with the dimensions and motions of a molecule [211,214,215]. Size exclusion chromatography and multi-angle light scattering (MALS) can be coupled together when the flow from SEC is going through a MALS detector that measures time-average intensity and proportion of light scattered by an analyte at multiple angles in relation to the incident laser beam. Analysis of the variations in the scattered light allows determination of the size of the molecules and molecular complexes expressed as radius of gyration, R_g [207,214,216,217].

Another useful method is NMR diffusion. The method allows measurement of diffusion coefficients, which are directly related to the protein size (through the Stokes-Einstein equation, equation 3). NMR diffusion experiments are set up as pulsed-field gradient (PFG) pulse sequence with varying gradient strengths. The first gradient pulse, which defocus the signals, is followed by a delay at which the molecules are let to diffuse. After this time interval, a second gradient pulse with the same magnitude and duration, but with the opposite effective sign is applied to refocus the signals. The resulting intensity of the signal is dependent on the pulse strength and duration, delay time and diffusion coefficient D of the studied molecules. Thus, for the molecules that move quickly, the signal will not be refocused and will result in low intensity. For larger slow moving molecules, the refocusing regains the signal [208,218,219]. The data of a diffusion experiment is a recording of a set of 1D spectra, each with a different gradient strength or diffusion delays. Processing of the data will result in a 2D map relating chemical shifts to the diffusion coefficient. By detecting the changes of the diffusion coefficient, the method enables characterization of folding and aggregation processes, interaction with ligands, and separation and identification of individual compounds in complex mixtures [208,210,220,221].

$$D = \frac{k_B T}{6\pi\eta R} \qquad\qquad (3)$$

In equation 3, D is the diffusion coefficient, k_B is the Boltzmann constant, η is the viscosity of the solution and R is the radius of a hard sphere.

2. AIMS OF THE STUDY

The overall objective of the work was to elucidate and understand the structural and biophysical requirements underlying the specificity of the ASHH2 CW domain interacting with histone tails with variable methylation states. The aims are 1) to determine the structure, dynamics, and thermodynamical binding parameters of the CW-H3K4me1 complex, and 2) to investigate the role of structural features like the α1-helix, unstructured regions and geometry of the binding pocket with respect to the domain's binding preferences, conformational stability and selection mechanism. Initially, there was also an aim to characterize the function of CW in the context of ASHH2 mini-enzyme construct, but this attempt was not achieved as we failed to establish an expression system for the enzyme production.

2.1 ASHH2 CW-H3K4me1 complex structure

Since publishing of the ASHH2 CW domain structure in its free state, the mechanism of CW binding and selectivity to the H3K4me1 peptide remained unclear. Analysis of available CW domain structures in complex with corresponding peptides suggests a common mechanism of interaction, with variable elements specific to particular CW subtypes. Liu and Huang, 2018 reported a crystal structure of the ASHH2 CW domain in complex with H3K4me1 peptide, highlighting important features and aspects of histone tail recognition, such as involvement of residues from α1-helix and the η1-loop region. The binding mechanism was discussed primarily in terms of lock-and-key interaction model. The published structure is not without limitations, as to solve it, a mutant E917A was used in order to achieve crystallization, and the domain was also truncated after residue I921, which is close to the functionally important α1-helix.

The objectives of the first part were to solve the NMR structure of ASHH2 CW domain in complex with H3K4me1 peptide, avoiding the limitations of the previously published structure, supplementing with analysis of the domain's internal dynamics to provide deeper and more detailed understanding of the contribution of the disordered regions in the binding mechanism.

2.2 Biophysical aspects of ligand binding and specificity

In the second part we aimed to analyze the primary determinants of interaction and understand how different structural features shape and affect interaction forces. Functionality of the unique to the ASHH2 subtype α1-helix, geometry of the binding pocket, and the role of disordered regions of the domain are of particular interest. With thermodynamical and structural characterization of the wild type interaction with histone mimicking peptides, supplemented with mutation studies, we aimed to understand what conditions binding specificity and affinity.

3. METHODOLOGY

Objectives formulated in Section 2 were approached by a combination of different NMR methods to characterize mechanism of interaction of ASHH2 CW domain complex structurally and study its dynamics. Molecular dynamics simulation (MD) was employed to make comparison between solved structures of ASHH2 CW, focusing on flexibility. Furthermore, methods of biophysical characterization (ITC, tryptophan fluorescence spectroscopy, NMR diffusion measurements) allowed to asses thermodynamic parameters of CW-ligang interactions, and physical properties like size and stability. An overview of the methods and their principles is given in section 1.8 and in Figure 7, and in each paper of the thesis as indicated at the end of each subsection.
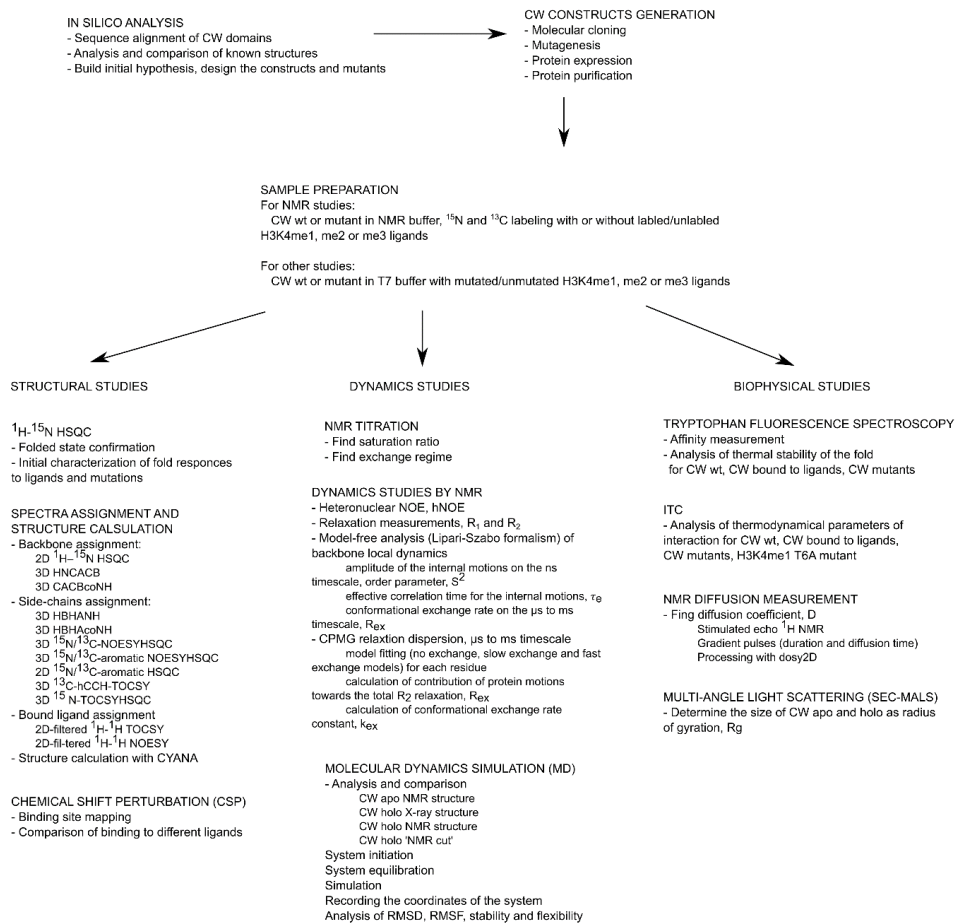
IN SILICO ANALYSIS
- Sequence alignment of CW domains
- Analysis and comparison of known structures
- Build initial hypothesis, design the constructs and mutants

CW CONSTRUCTS GENERATION
- Molecular cloning
- Mutagenesis
- Protein expression
- Protein purification

SAMPLE PREPARATION
For NMR studies:
  CW wt or mutant in NMR buffer, ^{15}N and ^{13}C labeling with or without labled/unlabled H3K4me1, me2 or me3 ligands

For other studies:
  CW wt or mutant in T7 buffer with mutated/unmutated H3K4me1, me2 or me3 ligands

STRUCTURAL STUDIES

^{1}H-^{15}N HSQC
- Folded state confirmation
- Initial characterization of fold responces to ligands and mutations

SPECTRA ASSIGNMENT AND STRUCTURE CALSCULATION
- Backbone assignment:
  2D ^{1}H–^{15}N HSQC
  3D HNCACB
  3D CACBcoNH
- Side-chains assignment:
  3D HBHANH
  3D HBHAcoNH
  3D ^{15}N/^{13}C-NOESYHSQC
  3D ^{15}N/^{13}C-aromatic NOESYHSQC
  2D ^{15}N/^{13}C-aromatic HSQC
  3D ^{13}C-hCCH-TOCSY
  3D ^{15}N N-TOCSYHSQC
- Bound ligand assignment
  2D-filtered ^{1}H-^{1}H TOCSY
  2D-fil-tered ^{1}H-^{1}H NOESY
- Structure calculation with CYANA

CHEMICAL SHIFT PERTURBATION (CSP)
- Binding site mapping
- Comparison of binding to different ligands

DYNAMICS STUDIES

NMR TITRATION
- Find saturation ratio
- Find exchange regime

DYNAMICS STUDIES BY NMR
- Heteronuclear NOE, hNOE
- Relaxation measurements, R_1 and R_2
- Model-free analysis (Lipari-Szabo formalism) of backbone local dynamics
  amplitude of the internal motions on the ns timescale, order parameter, S^2
  effective correlation time for the internal motions, τ_e
  conformational exchange rate on the μs to ms timescale, R_{ex}
- CPMG relaxtion dispersion, μs to ms timescale
  model fitting (no exchange, slow exchange and fast exchange models) for each residue
  calculation of contribution of protein motions towards the total R_2 relaxation, R_{ex}
  calculation of conformational exchange rate constant, k_{ex}

MOLECULAR DYNAMICS SIMULATION (MD)
- Analysis and comparison
  CW apo NMR structure
  CW holo X-ray structure
  CW holo NMR structure
  CW holo 'NMR cut'
System initiation
System equilibration
Simulation
Recording the coordinates of the system
Analysis of RMSD, RMSF, stability and flexibility

BIOPHYSICAL STUDIES

TRYPTOPHAN FLUORESCENCE SPECTROSCOPY
- Affinity measurement
- Analysis of thermal stability of the fold for CW wt, CW bound to ligands, CW mutants

ITC
- Analysis of thermodynamical parameters of interaction for CW wt, CW bound to ligands, CW mutants, H3K4me1 T6A mutant

NMR DIFFUSION MEASUREMENT
- Fing diffusion coefficient, D
  Stimulated echo ^{1}H NMR
  Gradient pulses (duration and diffusion time)
  Processing with dosy2D

MULTI-ANGLE LIGHT SCATTERING (SEC-MALS)
- Determine the size of CW apo and holo as radius of gyration, Rg

Figure 7. Overview of the methodological approach of the thesis.

3.1 Materials

The Histone H3 tail mimicking peptides were synthetized by Lifetein and had 95% purity as assessed by mass spectrometry. For some NMR experiments the peptides also were selectively ^{15}N and ^{13}C labled in positions A1, A7, R2 and R8. A mutated T6A monomethylated peptide was also used as a control. An additional tyrosine was introduced to the peptide sequences for concentration determination by absorption at 280 nm (NanoDrop), and an extinction coefficient of 1490 M^{-1} cm^{-1} was used. The peptide sequences are summarized in Table 1. D_2O, ^{15}N enriched (99%) NH_4Cl and ^{13}C

enriched (99%) glucose were purchased from Cambridge Isotopes, and SVCP-Super-3-103.5 NMR tubes were acquired from Norell. Unless otherwise specified, samples were buffered by the T7 solution (25 mM Tris-HCl pH 7.0, 150 mM NaCl, 1mM TCEP) or NMR buffer (20 mM phosphate, pH 6.4, 50 mM NaCl, 1 mM DTT). For protein purification TZNKβT buffer (50 mM Tris-HCl, pH 8.5, 12 mM NaCl, 100 μM $ZnAc_2$, 150 mM KCl, 2 mM MgCl, 10 mM β-mercaptoethanol) was used. Buffer components were acquired from Sigma-Aldrich.

More details on the materials description can be found in the Material and Methods sections of all three papers, under subheadings "Materials and experiments" in Paper I, "Materials" in Paper II, and "Materials" in Paper III.

Table 1. Histone mimicking peptides.

Name	Sequence
H3	ARTKQTARY
H3K4me1	ARTK(me1)QTARY
H3K4me2	ARTK(me2)QTARY
H3K4me3	ARTK(me3)QTARY
H3K4me1 T6A	ARTK(me1)QAARY
H3A1*K4me1A7*	A(^{15}N,^{13}C)RTKme1QTA(^{15}N,^{13}C) RY
H3R2*K4me1	AR(^{15}N,^{13}C)TKme1QTARY
H3K4me1R8*	ARTKme1QTAR(^{15}N,^{13}C)Y

X* indicates isotope-labeled site, where X is the amino acid.

3.2 *In Silico* analysis

Initial analysis of CW domains included sequence alignment using Jalview software [222] with Clustal O algorithm with default parameters. Analysis and comparison of CW structures was performed with UCSF Chimera [223], PyMOL and POSA web tool [224].

More details on the method description can be found in the Material and Methods section under subheading "Analysis of known structures" in Paper III.

3.3 Cloning of CW constructs, site-directed mutagenesis

After *in silico* analysis, in the attempt to generate a construct suitable for crystallization, several constructs with various length of N- and C-terminal ends were produced by ligation independent cloning into KpnI/SacI restriction sites of pET-49b vector (Novagen). Figure 8A presents used primers maped on CW sequence. The constructs fused to an N-terminal GST-tag were subsequently cloned into pSXG vector (R. Aasland, V. De Marco, V. Christodoulou and S. J. Gamblin, unpublished). Between CW and GST-tag there was a thrombin cleavage site, and this design would allow purification and cleavage of the constructs on the column. The most successful constructs, in terms of ligand binding, were CW33, CW37 and CW42 (analysed by fluorescence spectroscopy). These constructs in complex with H3K4me1 ligand were used for a crystallization attempt. The crystallization efforts failed, and structural analysis proceed using NMR. The CW42-construct was chosen for further studies as it was expressing with high yields and had affinity comparable to previously studied CWs [90].

The CW42 construct was also subjected to site directed mutagenesis using PCR with different pairs of primers carrying a mutation. These were mutations in the α1-helix (I915A and L919A), η1 and η3 loops (D886A, S907P, S907G and Q908E), C-terminal coil (Q923A) and loop connecting the β-sheets (CW-M3loop and CW-Z1loop). Position of mutations mapped on the structure are shown in Figure 8B.

Generated plasmids were then used to transform chemicompetent cells. Cells were plated on agar-plates supplemented with ampicillin for the selection of successful transformants. Bacterial cultures grown from individual colonies were then used to prepare mini-preps of plasmids. The nucleotide sequences of the plasmids were verified by sequencing.

In the course of the thesis work, expression and purification an ASHH2 mini-enzyme construct was attempted. This construct would allow studying possible effects of CW on the neighbouring methyltransferase SET domain. Two sequences were chosen for design: one contained all domains from ASHH2 (CW-AWS-SET-PostSET, residues

3 2

801-1290, denoted as SDG8) and methyltransferase domains only (AWS-SET-PostSET, residues 938-1212, denoted as SDG8s). The sequences were cloned into pGEX-4T-1, pET His6 MBP TEV LIC (1M) and pET His6 SUMO TEV LIC (1S) vectors (Adgene) by ligation independent cloning. The constructs were fused with GST and His-tags for purification, and with MBP and SUMO-tags in an attempt to improve the solubility of the expression product [225,226]. Overview of the constructs is presented in Figure 8C. The resulted vectors were used to transform chemicompetent cells, and the obtained mini-prep plasmids were verified by sequencing.
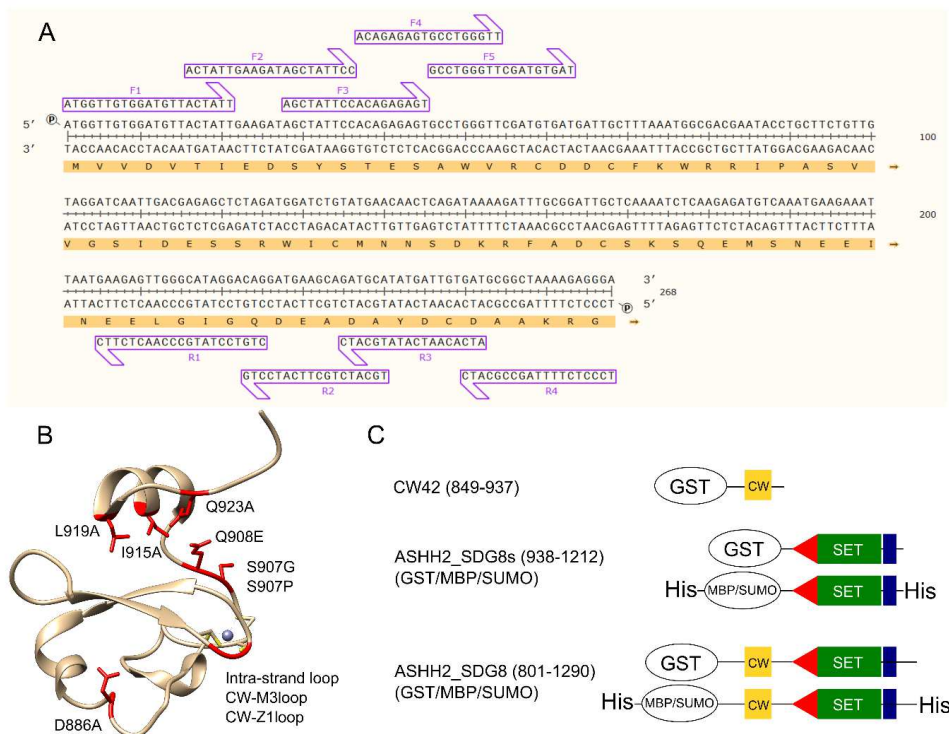

Figure 8. Design of CW constructs. A – primers used to generate CW constructs mapped on the CW sequence. CW42 was made with F4 and R2 primer pair. B – amino acids subjected to mutagenesis are shown on CW structure (highlighted in red). C – design of ASHH2 mini-enzyme constructs.

More details on the method description can be found in the Material and Methods sections of all three papers, under subheadings "Protein expression" in Paper I, "Materials" in Paper II, "Cloning of CW constructs, site-directed mutagenesis, protein expression and purificationand" in Paper II supplementary materials, and "Proteinexpression and purification" in Paper III.

3.4 Protein expression and purification

The BL21 bacterial strain for protein expression was transformed with the CW plasmids described above. Transformed bacteria were cultivated on 2xYTG media supplemented with ampicillin at 37 °C. During cultivation, the optical density (OD(A600)) of the media was monitored, and at the point when it was reaching 0.6 the expression was induced by adding IPTG to the media. After that, the culture was left to grow overnight at 18 °C.

For production of labeled samples for NMR studies, the M9 minimal media was used. The media contained minimal required components for bacterial growth: trace elements, vitamins, nitrogen and carbon sources. Use of ^{15}N ammonia sulphate and ^{13}C glucose as the nitrogen and carbon sources allowed uniform labelling of the proteins. Culturing was done in a few steps: 2xYTG media mini-prep incubation overnight at 37 °C, inoculation of M9 pre-culture (overnight incubation at 37 °C), transfer of pre-culture to 1L M9 media and incubation until OD(A600) was reaching 0.6 with subsequent induction with IPTG. Overnight culturing was performed at 18 °C.

The BL21 bacterial strain was also transformed with prepared mini-enzyme vectors and screened for optimal expression conditions. Cultures of transformed cells, picked from a single colony from agar plate, were grown at 16, 25 and 37 °C, and induced with IPTG after the OD was reaching 0.6. To monitor expression progress, the samples were taken from the media within 5 to 36 hours. This screening showed that none of the construct resulted in the expression of stable soluble constructs.

After incubation the cells were harvested by centrifugation and resuspended in lysis buffer named TZNKβT, supplementing with EDTA-free protease inhibitor (Roche).

Resuspended cells were disrupted by sonication, and centrifuged to clear the lysate of cell debris. For affinity-based protein purification, Glutathione Sepharose 4B beads (GE Healthcare) were used. Before loading, the column was pre-equilibrated with TZNKβT buffer. GST-taggfed CW constructs were bound to the column by flushing the column or incubating it with the lysate supernatant. The loaded column was then washed with TZNKβT to remove all unbound materials. To cut the CW constructs off from the GST-tag, the column was incubated with biotinylated thrombin overnight. The cleaved protein was collected by flushing the column with TZNKβT buffer. Thrombin contained in this eluate was subsequently removed by incubation with Streptavidin Sepharose (GE Healthcare) followed by centrifugation. Purification steps were monitored by SDS-gel electrophoresis. The protein samples were purified further by size exclusion chromatography eluting with T7 or NMR buffer. Figure 9 shows an example of purification steps monitored by SDS-gel and SEC chromatogram of the final purified protein. The results are representative for the most mutants prepared in the project, and confirm in each case the purity and monomeric state of the constructs. The fractions containing purified protein were collected and concentrated using Amicon Centrifugal Filters with 3 kDa cut-off (Millipore).
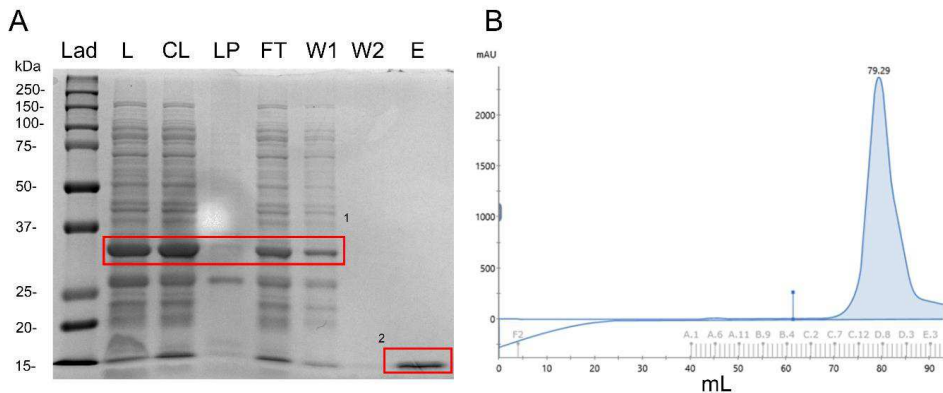

Figure 9. Monitoring of purification of CW constructs. A – SDS-gel (Lad – ladder, L – lysate, CL – cleared lysate after centrifugation, LP – lysed pellet after centrifugation, FT – flowtrough, W1 – first fraction of wash, W2 – last fraction of wash, E – elute). Red rectangles highlight position of GST-CW (1) and cleaved CW (2). B – SEC chromatogram, shows the elution peak of purified CW between 70-90 mL.

More details on the method description can be found in the Material and Methods sections of all three papers, under subheadings "Protein expression" and "Protein purification" in Paper I, "Materials" in Paper II, "Cloning of CW constructs, site-directed mutagenesis, protein expression and purificationand" in Paper II supplementary materials, and "Proteinexpression and purification" in Paper III.

3.5 NMR spectroscopy

Sample preparation: Purified labled CW constructs were dissolved in NMR buffer containing 10% (v/v) of D_2O. When needed, peptides were added at the concentrations corresponding to saturation ratio or higher. The samples were kept in sealed NMR sample tubes stored at 4 °C.

Data collection: Data was acquired with 850 MHz Bruker Avance III HD spectrometer at 25 °C. The instrument was equipped with a $^1H/^{13}C/^{15}N$ TCI CryoProbe and a SampleJet with temperature control set to 4 °C for sample storage. Data was collected at 25 °C.

Backbone assignment, titration NMR and chemical shift perturbation: For sequential backbone assignment of CW in its free state and bound to peptides 2D ^1H–^{15}N HSQC, 3D HNCACB and 3D CACBcoNH spectra were acquired. The assignment was performed using the CARA program v 1.9.1.2 [227]. HSQC type of experiment was also used to indicate conformational exchange regime (fast or slow) and saturation ratios of CW binding to peptides. For this, the titration series were prepared with increasing ratios of CW:peptide and for each point shift of the peaks positions or peaks intensities were monitored. HSQC spectra are sensitive to changes in the backbone conformations, and comparison spectra of CW in its free state and bound to peptides allowed the calculation of chemical shift perturbations (CSP) from the change of the peaks positions. This change indicates structural responses to binding. The sensitivity of HSQC was also used to verify the folded state of the mutants and estimate structural disturbances caused by them.

Analysis of backbone local dynamics with heteronuclear NOE, relaxation measurements and model-free analysis: Dynamics analysis was performed for CW in free state and bound to peptides. In order to perform model-free NMR analysis, ^{1}H-^{15}N NOE values, ^{15}N Longitudinal ($R_1 = 1/T_1$) and transverse ($R_2 = 1/T_2$) relaxation rates were acquired. For determination of heteronuclear NOEs, two ^{1}H-^{15}N HSQC data sets were recorded, using 3 s recycling delay between transients. For T_1 and T_2 measurements sets of different relaxation delays were used (Table 2).

Table 2. Relaxation delays used for T_1 and T_2 measurements.

T_1 relaxation delays, ms	20	60	80	100	200	400	600	800	1000	1200	1400	
T_2 relaxation delays, ms	16	30	60	95	125	160	190	220	250	345	440	500

Analysis of the relaxation data was carried out in Bruker Dynamics Center 2.5.3 (Bruker BioSpin). ^{1}H-^{15}N steady-state NOE, longitudinal and transverse relaxation rates (R_1 and R_2) were determined from cross-peak intensities/integrals in ^{1}H-^{15}N HSQC spectra. The heteronuclear NOE values were estimated as the ratio of the steady-state intensities measured in the presence and absence of saturation of the proton magnetization. R_1 and R_2 parameters were obtained by fitting the peak intensities acquired at different relaxation delays to a two-parameter exponential decay function. The errors of the fitted parameters were calculated from the inverse of the weighted curvature matrix.

More details on the methods description can be found in the Material and Methods sections of all three papers, under subheadings "NMR sample preparation", "NMR data acquisition, processing and assignment" and "Titration experiments for CW42" in Paper I, "NMR spectroscpy" in Paper II, "NMR spectroscopy" and "Dynamics measurement" in Paper III.

3.6 Intrinsic tryptophan fluorescence spectroscopy

Affinity measurements: Affinity measurements by intrinsic tryptophan fluorescence spectroscopy was used at the stage of the CW constructs selection. Most successful candidates with stable expression CW33, CW37 and CW42 in comparison to CWs from [90] were analysed. For this analysis two solutions were prepared, each containing equal amounts of a given CW version, to ensure constant concentration throughout the titration. In the second solution, an excess amount of peptide was added, and this solution was titrated into the first one. Typical titration series consisted of 9-12 concentration points. The titration series were performed in triplicates for each CW in combination with H3K4me1, H3K4me2 and H3K4me3. The unmodified H3 histone-tail peptide was used as a control. Representative blanks containing only histone peptides were titrated into T7 buffer. Intrinsic tryptophan fluorescence measurements were done at 25 °C, with excitation at 290 nm. The emission data was recorded in the 300-450 nm interval. Intensities, F_{OBS}, at the wavelength giving the largest fluorescence change (319-322nm interval) were plotted as a function of histone peptide concentration. Then, the dissociation constant K_d was determined using a non-linear least-squares fit to equation 4 [228,229]:

$$F_{obs} = \frac{F_p P_o + (F_{pl} - F_p)}{\sqrt{(K_d + P_o + L_o) - (K_d + P_o + L_o)^2 - 4 P_o L_o}} \qquad (4)$$

Here, F_p is fluorescence of the protein and F_{pl} is fluorescence of protein-ligand complex, P_o and L_o are the total concentrations of the protein and the ligand.

Thermal denaturation measurements: To assess the thermal stability of the CW domain and its change upon binding to peptides or due to mutations, the samples were subjected to thermal denaturation with monitoring the changes in tryptophan fluorescence in the 310 nm and 450 nm region after excitation at 295 nm. CW in free state, bound to peptides, and CW α1-helix mutants, also in free state and bound to H3K4me1 peptide, were analysed in the 5 – 90 °C temperature range, with 5 °C – 10 °C stepwise increase in temperature. The samples containing only histone peptides were used as blanks. The measurements were done in triplicates. To find the melting

temperature T_m, the ratio of the fluorescence intensity at 335 nm and 355 nm [230] at each temperature point was plotted against the temperature with the subsequent fitting of a 4-parameter sigmoidal curve (equation 5):

$$f(x) = y_0 + \frac{y_f}{1+10^{\left(-\frac{Tm-x}{b}\right)}} \qquad (5)$$

Here, y_0 is the slope of the pre-transition state, y_f is the slope of the post-transition state, T_m is the inflection point in the curve, and b is the slope at the inflection point.

More details on the method description can be found in the Material and Methods sections under subheadings "Intrinsic tryptophan fluorescence" in Paper II, "Intrinsic tryptophan fluorescence affinity measurements" and "Thermal denaturation monitored by intrinsic tryptophan fluorescence" in Paper II supplementary materials, and "Thermal denaturation" in Paper III.

3.7 Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) was used to find thermodynamical parameters of interaction like the change in enthalpy ΔH, entropy ΔS and free energy ΔG, and find dissociation constants K_d. CW wt was titrated with methylated peptides and for control unmethylated H3 peptide and T6A mutant were also included in analysis. ITC was also performed for all CW mutants. Typical titration setup was 22 stepwise injections of a peptide to a sample, with 300 seconds intervals in between, performed in triplicates on Nano ITC from TA Instruments at 25 °C with a stirring rate of 300 rpm. As a blank, a peptide was titrated into T7 buffer only. Corrected heat flow peaks were integrated, plotted and fitted by independent modelling to determine binding parameters using the NanoAnalyze V 2.4.1 software.

More details on the method description can be found in the Material and Methods sections under subheadings "Isothermal calorimetry" in Paper II, and "Isothermal titration calorimetry (ITC)" in Paper III.

4. SUMMARY OF THE RESULTS

The general aim of the research was to understand and characterize the interaction mechanism of selective recognition of H3K4me1 histone modification by the CW domain of ASHH2 methyltransferase. The results of this research are presented in three papers that highlight different approaches and aspects of the study, supplementing each other.

Paper I is a documentation paper describing back-bone, side-chains and peptide assignments of NMR spectra of the CW domain in free and bound states and analysis of protein-peptide ratio at which the complex is saturated. This information was a foundation for the next work in solving the structure and analyzing protein dynamics. (*Published in Biomol NMR Assign 12(1):215-220*).

Paper II aimed to solve the NMR structure of the CW-H3K4me1 complex and assess the dynamics and internal motion of the protein in free and bound states. Complexation with the ligand was inspected with NMR diffusion, multiple angle light scattering (MALS) and size exclusion chromatography (SEC), as well as thermal denaturation analysis. A comprehensive assessment of the interaction mechanism was performed in comparison with previously reported by Liu and Huang, 2018 X-ray structure by the means of MD simulation. (*FEBS J, published online ahead of print, 2020 Feb 21 DOI: 10.1111/febs.15256*).

Paper III was focusing on biophysical aspects of the interaction mechanism, analyzing the leading forces of the interaction. Comparison of known CW domain structures revealed different structural features that were hypothesized to shape these interaction forces, therefor, along with the key interaction features highlighted in Paper II, were subjected to mutagenesis. Analysis of mutants together with wild type CW in free and bound to different peptides states was performed by isothermal titration calorimetry (ITC), thermal denaturation monitored by intrinsic tryptophan fluorescence spectrometry and NMR. (*Manuscript*).

4.1 CW-H3K4me1 complex structure and dynamics

The study described in Paper II aimed to solve the structure of the CW-H3K4me1 complex and determine the key elements involved in the interaction, as well as the overall interaction mechanism. Figure 10A shows the structural ensemble of 20 minimized structural conformers consistent with the 1056 NMR determined restrains. By comparing NMR structures in free (*apo*) and bound (*holo*) states, four residues I915, L919, I921 and Q923 were identified to form the "ceiling" of the binding pocket, closing the monomethylated lysine (Figure 10B and C). Binding of the peptide causes moderate reorganization of the backbone of the protein. The most prominent differences between the backbones of the *apo* and *holo* states are approximately 8 Å shift of the η1-loop, due to interaction with the N-terminal of the ligand, and a minor reposition of the α1-helix (≈ 2 Å), which allows insertion of the methylated lysine into the binding pocket. NMR data and MD simulations also indicate involvement of the C-terminal coil in the mediation of the interaction as it tends to shift towards the N-terminal end of the ligand upon binding. MD simulation, hydrogen bond analysis of the complex and secondary structure analysis verified a stable intermolecular β-sheet augmentation, reinforced by hydrogen bonds between ligand and CW domain, and expansion of the domain's β-sheets. ITC analysis of Q923A mutation indicated the presence of unspecific binding as the titration profile did not follow the sigmoidal trend, stating the importance of the Q923 residue in the interaction with the ligand.

Characterization of internal motions of the *apo* and *holo* states by NMR indicates that the CW domain is quite flexible in both states, but acquiring an overall stabilization upon peptide binding. The most flexible regions of the *apo* state are shown in Figure 10D. Complexation of CW with H3K4me1 ligand was also assessed for stability by intrinsic tryptophan fluorescence spectroscopy and hydrodynamic size estimations with size-exclusion chromatography (SEC) along with multi-angle light scattering (MALS) and diffusion constant measurements using pulsed-field NMR. Results from these methods showed that CW gets stabilized upon complexation with the ligand and gain a more compact form. Evidence for binding through conformational selection were found through applying CPMG relaxation dispersion experiments. The residues

D886 (from η1-loop), S907 and M910 (both from η3-loop) are characterized by slow exchange, i.e. a shifting between conformation on the ms-to-s timescale. In addition to η1-loop, which was shown to be important for binding, the dynamics and flexibility of the η3-loop, preceding the α1-helix, suggests a mechanism by which the α1-helix samples conformations upon binding with subsequent consolidation when the correct ligand is bound.

In order to test the role of flexibility and the residues highlighted by the NMR analysis, mutagenesis with subsequent ligand affinity measurements were performed. Specifically, S907 and Q908 residues from the η3-loop were subjected to mutagenesis and analysis of the interaction with ITC. The flexible serine residue was mutated to glycine (more flexible) and proline (least flexible) residues [231]. S907P mutation resulted in increase of K_d value by approximately 3-fold (3.22 ± 0.29 μM, versus 1.09 ± 0.21 μM for wt), and the effect of S907G mutation on the interaction constant was insignificant when tested with H3K4me1. S907P and S907G failed to bind di- and trimethylated ligands. Q908 residue was replaced with a conserved glutamic acid, and the Q908E mutation resulted in a significant drop in affinity by 14-fold (18.75 ± 1.02 μM).
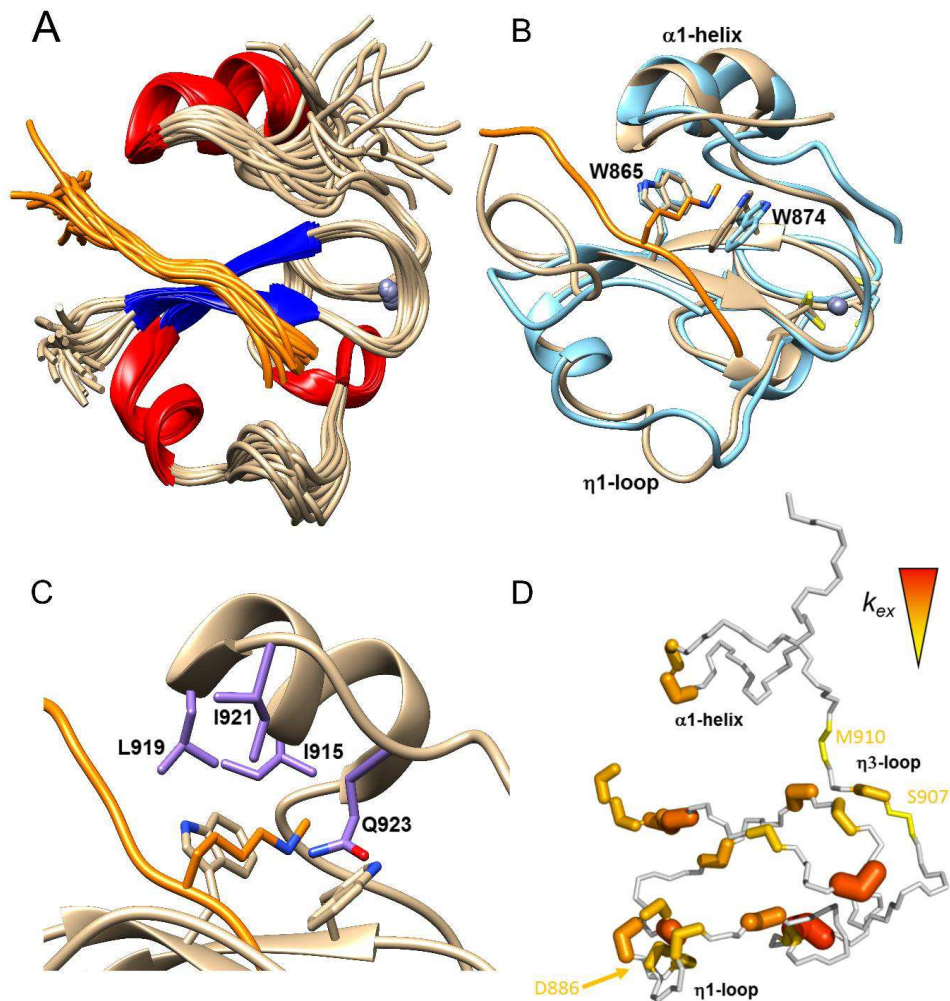
Figure 10. NMR structure of CW bound to H3K4me1 and dynamics analysis. A – The structural ensemble of 20 minimized NMR-derived structures. B – structural superposition of CW in free state (brown, PDB code: 2L7P) and bound state (blue, PDB code: 6QXZ); the H3K4me1 ligand shown in orange. C – organization of binding pocket with I915, L919, I921 and Q932 residues colored in purple. D – conformational exchange values k_{ex} plotted on CW domain structure in free state (intence red color and thick stickes represent higher k_{ex} values and quicker motion; yellow color and thin stickes represent smaller k_{ex} values and slower motion). A, B and C prepared with Chimera, and D prepared with PyMol.

4.2 Aspects of ASHH2's CW-domain selectivity

The study in Paper III aimed to identify the determinants that account for the selectivity of the domain towards H3K4me1. First, CW wild type in complex with three peptides (H3K4me1, H3K4me2 and H3K4me3) was characterized by HSQC experiments and chemical shift perturbations (CSP) indexing. Overall, the binding of all three peptides affected the structure of the domain in a similar way. Complexation with H3K4me3 peptide showed higher CSP values, as a larger methylated residue introduces a steric strain on the whole CW's structure. When interacting with the H3K4me1 peptide, a disordered region I921-Q923 succeeding the α1-helix showed higher CSP values. This suggests that this region might be involved into selectivity towards H3K4me1 peptide, which is in agreement with the mutation analysis of Q923, a residue which is also part of a NOE network that stabilize the domain in complex with H3K4me1. Closer inspection of HSQC signals from side chains of tryptophans forming the binding pocket and asparagines at the α1-helix indicated that complexation of wild type with the peptides affected the structure of the binding pocket and the α1-helix orientation differently, resulting in three different conformational states of the domain.

Interaction of the domain with peptides was also assessed by ITC. Results show that the interaction is driven by enthalpy. From the enthalpy-entropy contribution, it is seen that the interaction with H3K4me1 peptide is the most favorable, resulting in the lowest free energy of the system and the lowest K_d value.

Comparison of known CW structures from different protein families revealed a small loop region between β-sheets that differs between the paralogs. The length and composition of this loop might hypothetically affect the positioning of the β-sheets in relation to each other, and thus affect the orientation of tryptophan's side chains that form the binding pocket, and, therefore, this loop might regulate selectivity of the domain. To analyze this hypothesis, the CW-M3loop and CW-Z1loop mutants were prepared with the loop in ASHH2 matching the MORC3 and ZCWPW1. The mutants were assessed for changes in the interaction constants by ITC. HSQC fingerprinting of these mutations showed significant structural disturbance for the CW-Z1loop sample, which resulted in weakened binding and the interaction could not be characterized.

CW-M3loop construct did not show any big structural reorganization and resulted in an insignificant increase of K_d value (1.37 ± 0.32 µM). None of these mutations resulted in a change of specificity.

Paper II and work previously published by Liu and Huang, 2018, identified I915 and L919 residues as key residues in locking the ligand in the binding pocket. Liu and Huang, 2018, prepared and determined by ITC the affinities of I915A and L919A mutants towards mono-, di- and trimethylated peptides. They reported a change of selectivity for I915A mutant. We repeated the analysis of these mutants and subjected them to thermal denaturation, HSQC fingerprinting and ITC analysis to assess the contribution of these amino acids to selectivity and stability of the domain. Both of these mutations resulted in a prominent structural disturbance as about 60% of all the domain signals at HSQC spectra were significantly shifted (more than 0.2 ppm). The mutations also affected the thermal stability of the domain, reducing it by 10 °C, and increasing the interaction constants K_d by ten-fold (12.17 ± 2.10 µM, versus 1.09 ± 0.21 µM for wt).

5. GENERAL DISCUSSION

The aim of the study was to explore the determinants underlying the specificity and stability of the ASHH2 CW-domain with a particular interest in the role of the C-terminal α1-helix and disordered regions. Structural analysis and comparison with other known structures of CW domains also suggested a hypothesis that the specificity of the domain might be conditioned by the flexible loops. To approach this objective, we analyzed the interaction of the wild type in more detail, followed by mutation studies with the aim of directly test the relevance of residues positioned in loops and other high-mobility elements of the domain.

5.1 Binding mechanism of ASHH2's CW-domain

Liu and Huang, 2018, explored the interaction mechanism by the means of crystallography and explained it in terms of lock-and-key binding model [101]. Trying to avoid the limitations of their work (use of the E917A mutant and a short construct to enable crystallization), the NMR structure was solved with the assessment of the dynamic behavior of the structure in free and bound states. From these results, the structure of the CW domain appears to be flexible even when complexed with peptides. HSQC fingerprinting and inspection of side chains signals of tryptophans, forming part of the binding pocket, and asparagines from α1-helix presented in Paper III, showed that the domain attains different conformational equilibriums when bound to different peptides. MD simulations indicated that complexation with monomethylated peptide is followed by β-sheet augmentation of the ligand. These observations lead to the conclusion that the interaction mechanism of the CW domain is better characterized by conformational selection and induced fit models of interaction, rather than lock-and-key.

The C-terminal α1-helix is flanked by two unstructured regions: the η3 "hinge" region and the I921-Q923 region. Mutation analysis indicates that S907 residue, which shows high mobility, allows movement of the α1-helix to open the binding pocket for the methylated residue of the peptide. The residue and the helix are getting stabilized when the domain is complexed with H3K4me1 ligand. Interaction with the peptide is

finalized by locking it with the I921-Q923 region. The flexibility of the η3 and I921-Q923 regions allows positional adjustment of the α1-helix into a favorable orientation when the CW domain is bound to H3K4me1 peptide.

5.1.1 Assessment of NMR and Crystal structures

Coinciding with the convergence of the NMR structure solved in this work, an X-ray structure of CW in complex with H3K4me1 peptide was published by Liu and Huang, 2018 [101]. They also identified the η1-region, I915 and L919 residues as the key features of the binding mechanism and discussed the interaction in terms of lock-and-key binding. While both structures represent a domain removed from its functional context, the crystal structure has several additional limitations. To solve the crystal structure, the E917A mutant was prepared in order to generate stable crystals. Moreover, a shorter version of CW, terminated at I921, and therefore devoid of C-terminal coil region, was used in their work. The H3K4me1 mimicking peptide was also shorter and included residues from 1st to 7th, while the peptide from 1st to 9th residues of the H3 tail was used for NMR structure. It could also be argued that crystal structures, while usually high in resolution, are inherently biased towards static and compact structures [232,233].

Having access to structural models solved by independent techniques lends a more complete picture when considering the strengths and weaknesses of both cases. While overall, the structures match, there are differences in positioning of C-terminal α1-helix and displacement of C$_\alpha$s by roughly 4 Å, which are likely a result of the limitations of the Liu and Huang's approach. In the NMR structure, the ligand in the binding pocket is "surrounded" by the I921-Q923 coil, which is missing in the crystal structure, leaving the ligand open. Other differences were observed in the orientation of the ligand for residues Q5-A7, and some conformers of the NMR structure suggest the formation of β-sheet by peptide. This, however, was not seen in the crystal structure. To comprehensively asses the structures, MD simulations were run for the NMR structure, X-ray structure, and a shortened NMR structure, made to match the X-ray in length. We observed that the crystal structure is much more restricted in its movements, which suggests that the structure is trapped in a conformation that is difficult to escape from.

Hydrogen bond and secondary structure analysis derived from MD simulations of both structures showed β-augmentation, but with some differences in the ligand's residues involved in the interaction (A2 to Q5 for the X-ray structure and T3 to A6 in the NMR structure). MD simulations run on the full length and truncated NMR structures indicated an increase in flexibility of the α1-helix and β1-sheet in the structure devoid of the C-terminal coil, which, again, suggests a role of the C-terminal coil in the stabilization of the ligand and the overall complex upon its binding.

Identified by both X-ray and NMR structures as the key residues that stabilize the methylated lysine sidechain inside the binding pocket, I915, L919 and Q923 were subjected to mutagenesis. ITC analysis performed by Liu and Huang, 2018 did not include Q923, as the construct used in their work was devoid of this residue. The conclusion was that I915 is a regulatory residue responsible for selectivity towards H3K4me1 peptide. Close inspection of their results indicated that the experimental setup was not properly optimized, as the reported stoichiometric coefficients of interaction derived from their analysis deviate substantially from 1 [187]. The mutation analysis was repeated and the function of I915 and L919 residues was accessed by ITC and fluorescence spectrometry. Results presented in Paper III did not verify the involvement of these residues into regulation of selectivity, and suggested that these residues function by stabilization of the domain's fold, probably by protecting the tryptophan residues from unfavorable solvent interactions in the absence of ligand. HSQC fingerprinting of these mutants showed significant shift of the signals as the result of structural disturbances, which indicates that the CW domain is very sensitive to mutations in this region.

5.2 Characterization of CW domain selectivity

After characterization of the interaction mechanism of the CW domain with H3K4me1 peptide in Paper II, the question of selectivity was addressed further in Paper III.

The CW domain in complex with histone mimicking peptides was analyzed by HSQC type of experiments followed by chemical shift perturbation indexing of the CW's

backbone. Binding to all three peptides causes similar structural disturbances, which also affect the domain outside the immediate vicinity of the binding pocket. Notable difference is that binding to H3K4me3 peptide resulted in slightly higher values of CSP for most of the residues. Strain induced by the H3K4me3 peptide in the binding pocket is also reflected in thermal denaturation analysis, where complexation with the peptide decreased the thermal stability of the domain. Binding to H3K4me1 had the highest CSP values for the I921-Q923 region, suggesting its involvement in selectivity.

Results from analysis of NMR dynamics of the CW complexed with peptides showed that the I921-Q923 region is getting stabilized the most when the CW binds H3K4me1, again suggesting involvement of this region in selectivity.

Calorimetric analysis by ITC showed that the interaction is driven by enthalpy that overcomes an entropic barrier. In the lock-and-key binding model it is the entropy that drives the interaction [192,193] and as our results show that the interaction of CW with the ligand is enthalpy driven, it is an argument in favor of conformational selection and induced fit mechanism [192,194]. A balance between enthalpy and entropy contributions is reflected in differences in dissociation constants for CW in complex with peptides. Binding to H3K4me1 has the lowest K_d value and H3K4me3 has the highest, which is in agreement with Hoppmann et al., 2011 and Liu and Huang, 2018. Enthalpy has the lowest value when CW is bound to H3K4me1 peptide, and, as the enthalpy in the non-strict terms characterizes the sum of the energies of non-covalent bonds, which were formed upon complexation [192,195], the CW-H3K4me1 complex obtains more of the protein-ligand and intradomain connections. These connections are likely the contacts made by the I921-Q923 region with the ligand and the formation of hydrogen bonds by β-sheet augmentation.

Taking together, the results from chemical shift perturbation analysis, analysis of the domain's dynamics and calorimetric analysis of interaction it is possible to conclude that I921-Q923 region is an important determinant that contributes to selectivity of the CW domain towards H3K4me1 peptide, as in cases of interaction with H3K4me2 and

H3K4me3 this region fails to make stable contacts and stabilize the structure of the complex, resulting in reduced K_d values and reduced thermal stability.

5.3 Sub-type specific determinants of CW domain's selectivity

CW domain family is represented by several known structures originated from different protein families. These CW share core structural elements, like tryptophan binding pocket, located at the antiparallel β-sheets, and Zn^{2+} ion that maintains the structure via coordination by cysteines. The rest of the fold is variable for the individual domains. The domains exhibit different preferences towards the methylation state of the ligands. While ASHH2's CW is more specific towards H3K4me1, others were shown to prefer H3K4me3 and me2 modifications [90,92,95,96]. Comparative analysis of known structures revealed variable structural elements which might be involved into selectivity of a domain. In case of ASHH2's CW domain, the C-terminal α1-helix is one of such features. A closer look at the geometry of the binding pocket showed that the angle between the tryptophans differ for the domains, and ASHH2 has the smallest of them. Limited space of the binding pocket might, therefor, also contribute to the domain's preference towards "smallest" monomethylated ligand. Since the tryptophans that form the binding pocket are located at the β-sheets, their orientation and, thus, the angle between them will depend on these β-sheets relative arrangement. The β-sheets are connected by a loop, variable in length and amino acid composition, which hypothetically might affect the geometry of the binding pocket.

Possible involvement of these structural features into selectivity was investigated by analysis of the mutants. Mutant that was lacking the α1-helix was first generated by Hoppmann et al., 2011 and was shown to lose its binding activity [90]. Going deeper into studying this construct revealed that the loss of activity was due to fold disruption, suggesting the conclusion that the α1-helix's function is fold maintenance. Thermal stability analyses of I915A and L919A α1-helix mutants indicated that the fold is stabilized by I915 and L919 amino acids, most likely due to their orientation inside the binding pocket, where these amino acids together with tryptophans create and maintain a hydrophobic environment. HSQC spectra of these mutants also revealed great

structural disturbances introduced by these mutations on the overall fold, including the chemical environment around the tryptophan side chains in the binding pocket. It is, however, very difficult to completely deconvolute the α1-helix effects on fold stability and binding specificity, as they inevitably affect each other.

Analysis of mutants with variable loops, when ASHH2 CW domain was generated to have the loop connecting β-sheets to be the same as for CW from MORC3 and ZCWPW1 proteins, did not indicate any change of selectivity. HSQC spectra showed a disturbance of the overall fold, resulting in loss of binding (for CW-Z1loop) or reduced affinity (for CW-M3loop).

These observations contribute to the conclusion that the interaction mode of the CW domain and its recognition of ligand is better characterized by flexible and dynamic conformational selection model, rather than lock-and-key binding, as none of the analyzed mutations resulted in a shift of preference, but rather introduced structural disturbances that often encompassed the entire fold.

5.4 CW function in context of full length ASHH2 enzyme

Although the objectives of the present study were concerned mainly with the interaction mechanism of CW alone, the presented findings can be discussed in the light of the full length protein. Changes in conformation associated with CW binding to the histone tail modified at H3K4me1 residue might potentially contribute to regulation of SET-domain positioning and orientation. Thus, repositioning of the C-terminal region closer to the backbone of the histone might allow adjustment of the SET-domain positioning towards H3. Formation of β-sheet by the region of the histone bound by CW might be required for the orientation of the target H3K36 residue in the vicinity of the SET domain's active site.

Another possible function of ASHH2 CW-domain can be hypothesized by extrapolation from other CW containing proteins. For example, for human MORC3 ATPase it was shown that CW regulated its enzymatic activity by autoinhibition. The enzyme is activated only after binding to H3 histone tail [234]. CW of ASHH2 might

be involved in the activation of the SET-domain by a similar mechanism. For MORC3, it was also shown that the active state of the enzyme bound to DNA and histone tail drives phase separation and concentration of free DNA within its droplets [235]. Although ASHH2 does not interact directly with DNA, a similar effect can be inherent to its function. The interaction environment of ASHH2 with H3 histone and the binding of the methyl group donor S-adenosyl-methionine (SAM) utilized by the SET-domain might shift the equilibrium of the phase separated heterochromatin towards its active, loose structure. Another example is the CW domain of LSD2 amine oxidase. Despite being inactive and not binding histone tails, it contributes to the overall stability of the enzyme and is required for its activity [96-99]. CW of ASHH2 can also be an important determinant in the overall structural maintenance of the whole enzyme.

6. CONCLUDING REMARKS

In this study we aimed to characterize and understand the interaction mechanism by which CW domain from ASHH2 methyltransferase selectively recognize H3K4me1 histone modification and differentiate between mono-, di- and trimethylated state. A complementary structural and biophysical approach, combining high and low resolution techniques, prompted us to question the fundamental nature of the interaction mechanism as assumed in existing literature. Calorimetric, dynamic and mutational analysis showed that the binding and specificity of ASHH2 CW domain towards H3K4me1 is better described by the conformational selection mechanism and conditioned by energetically favorable enthalpy driven interaction. Furthermore, low affinity towards H3K4me3 peptide is associated with a small size of the binding pocket at equilibrium and decreased stability of the CW-H3K4me3 complex. NMR fingerprinting of the mutants showed how sensitive the CW domain is to mutations, especially in the C-terminal α1-helix region, and analysis of the mutants verified that the α1-helix and the C-terminal tail is an important determinant of the structure stabilization and binding affinity, functioning via intradomain interaction of L919 and I915 residues with the tryptophan side chains in the binding pocket. Finally, upon complexation with H3K4me1 ligand, the repositioning of the α1-helix is mediated by unstructured I921-Q923 and η3 regions.

6.1 Future perspectives

The role of the CW domain in the context of full length protein is yet to be understood. The movement of the unstructured C-terminal coil and the formation of β-sheet by histone tail might be a part of the mechanism which regulates the positioning of the SET-domain along H3 histone ensuring its activity. Zhang *et al*., 2019 presented evidences that CW is involved in the regulation of the enzymatic activity of the ATPase domain of the human MORC3 protein, releasing autoinhibition by CW after binding to H3 histone tail [234]. ASHH2's CW might as well function in a similar way. Solving the structure of the full length enzyme in the free state and in the complex with H3 histone, will help to characterize the function of the enzyme. Use of histone mimicking

peptides or recombinant histones will not allow analysis of possible effects of neighbouring modifications and associated chromatin DNA. However, these limitations can be overcome by the use of purified nucleosomes.

Another possible function can be hypothesized from works done by Dong *et al*., 2008 and Xu *et al*., 2008. Comparison of the results from methyltransferase assays with radiography labeled methyl group donor that used constructs of different architecture (full length construct and a construct devoid of CW domain, consisted of AWS-SET-PostSET domains only) suggests that the CW domain of ASHH2 methyltransferase might function by regulating the SET domain activity restricting it specifically towards H3 histone, as the construct that was lacking CW domain was able to transfer the methyl group *in vitro* not only to H3 histone but also to H4 [105,236]. This setup could be repeated to verify the hypothesis and expanded with structural biology analysis to understand the function of CW in such context.

As ASHH2 methyltransferase itself appears in the wider context at the chromatin level, likely as a part of a cromatin remodeling or mediator complexes, it is worth also switching the attention from *in vitro* characterization to *in vivo* and try to assess its function in the context of a living cell. To shift the focus from the domain's function and answer the questions of the composition and structure of ASHH2 containing complexes and the identification of associated genes that are regulated by them, a combinatorial approach of modern methods of proteomics, sequencing and cryo-EM would have to be utilized. Proteomics could provide qualitative and quantitative characterization of the protein complexes and highlight their structure and organization by cross-linking mass spectrometry. Proteomics coupled with sequencing could identify the genomic regions at which the ASHH2 containing complexes function and their dynamics during cell cycle. Cryo-EM can provide structural information at higher resolution and reveal organization of a protein complex together with the nucleosome it is attached to.

Phase separation effects have been attributed to the activity of CW containing MORC3 ATPase enzyme [235]. ASHH2 potentially might also be involved in similar

mechanisms of chromatin structure regulation. To assess possible involvement of ASHH2 in phase separation events, methods combining experiments with fluorescently labeled enzyme in mixture with purified nucleosomes and methyl group donor S-adenosyl-methionine (SAM) will allow its characterization *in vitro*, and studying genome-edited cells expressing such fluorescently tagged protein will potentially reveal its action in native condition *in vivo*.

7. REFERENCES

[1]   Bednar, J., Horowitz, R.A., Grigoryev, S.A., Carruthers, L.M., Hansen, J.C., Koster, A.J. and Woodcock, C.L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc Natl Acad Sci U S A 95, 14173-8.

[2]   Olins, D.E. and Olins, A.L. (2003). Chromatin history: our view from the bridge. Nat Rev Mol Cell Biol 4, 809-14.

[3]   Jansen, A. and Verstrepen, K.J. (2011). Nucleosome positioning in Saccharomyces cerevisiae. Microbiol Mol Biol Rev 75, 301-20.

[4]   Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389, 251-60.

[5]   Maeshima, K., Hihara, S. and Eltsov, M. (2010). Chromatin structure: does the 30-nm fibre exist in vivo? Curr Opin Cell Biol 22, 291-7.

[6]   Coverley, D. and Laskey, R.A. (1994). Regulation of eukaryotic DNA replication. Annu Rev Biochem 63, 745-76.

[7]   Parker, M.W., Botchan, M.R. and Berger, J.M. (2017). Mechanisms and regulation of DNA replication initiation in eukaryotes. Crit Rev Biochem Mol Biol 52, 107-144.

[8]   Heyer, W.D., Ehmsen, K.T. and Liu, J. (2010). Regulation of homologous recombination in eukaryotes. Annu Rev Genet 44, 113-39.

[9]   Reményi, A., Schöler, H.R. and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat Struct Mol Biol 11, 812-5.

[10]   Bertoli, C., Skotheim, J.M. and de Bruin, R.A. (2013). Control of cell cycle transcription during G1 and S phases. Nat Rev Mol Cell Biol 14, 518-28.

[11]   Maeshima, K., Hihara, S. and Takata, H. (2010). New insight into the mitotic chromosome structure: irregular folding of nucleosome fibers without 30-nm chromatin structure. Cold Spring Harb Symp Quant Biol 75, 439-44.

[12]   Arnosti, D.N. and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94, 890-8.

[13]   Li, H. and Johnson, A.D. (2010). Evolution of transcription networks--lessons from yeasts. Curr Biol 20, R746-53.

[14]   Merika, M. and Thanos, D. (2001). Enhanceosomes. Curr Opin Genet Dev 11, 205-8.

[15]   Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7, 703-13.

[16]   Li, Q., Barkess, G. and Qian, H. (2006). Chromatin looping and the probability of transcription. Trends Genet 22, 197-202.

[17]   Björklund, S. and Gustafsson, C.M. (2005). The yeast Mediator complex and its regulation. Trends Biochem Sci 30, 240-4.

[18]   Kim, Y.J. and Lis, J.T. (2005). Interactions between subunits of Drosophila Mediator and activator proteins. Trends Biochem Sci 30, 245-9.

[19]   Kornberg, R.D. (2005). Mediator and the mechanism of transcriptional activation. Trends Biochem Sci 30, 235-9.

[20]   Malik, S. and Roeder, R.G. (2005). Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends Biochem Sci 30, 256-63.

[21]   Hargreaves, D.C. and Crabtree, G.R. (2011). ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. Cell Res 21, 396-420.

[22]   Henikoff, S. and Shilatifard, A. (2011). Histone modification: cause or cog? Trends Genet 27, 389-96.

[23]   Wang, X., Bai, L., Bryant, G.O. and Ptashne, M. (2011). Nucleosomes and the accessibility problem. Trends Genet 27, 487-92.

[24]   Narlikar, G.J., Fan, H.Y. and Kingston, R.E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. Cell 108, 475-87.

[25]   Smith, R.L. and Johnson, A.D. (2000). Turning genes off by Ssn6-Tup1: a conserved system of transcriptional repression in eukaryotes. Trends Biochem Sci 25, 325-30.

[26]   Simon, J.A. and Kingston, R.E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol 10, 697-708.

[27]   Liu, Z. and Karmarkar, V. (2008). Groucho/Tup1 family co-repressors in plant development. Trends Plant Sci 13, 137-44.

[28]   Klose, R.J. and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci 31, 89-97.

[29]   He, X.J., Chen, T. and Zhu, J.K. (2011). Regulation and function of DNA methylation in plants and animals. Cell Res 21, 442-65.

[30]   Ehrlich, M. and Lacey, M. (2013). DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. Epigenomics 5, 553-68.

[31]   Zhang, H., Lang, Z. and Zhu, J.K. (2018). Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol 19, 489-506.

[32]   Li, E. and Zhang, Y. (2014). DNA methylation in mammals. Cold Spring Harb Perspect Biol 6, a019133.

[33]   Jin, B. and Robertson, K.D. (2013). DNA methyltransferases, DNA damage repair, and cancer. Adv Exp Med Biol 754, 3-29.

[34]   Feng, S. et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107, 8689-94.

[35]   Lyko, F. (2018). The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet 19, 81-92.

[36]   Kouzarides, T. (2007). Chromatin modifications and their function. Cell 128, 693-705.

[37]   Ruthenburg, A.J., Li, H., Patel, D.J. and Allis, C.D. (2007). Multivalent engagement of chromatin modifications by linked binding modules. Nat Rev Mol Cell Biol 8, 983-94.

[38]   Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D. and Patel, D.J. (2007). How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat Struct Mol Biol 14, 1025-1040.

[39]   Turner, B.M. (2007). Defining an epigenetic code. Nat Cell Biol 9, 2-6.

[40]   Tollervey, J.R. and Lunyak, V.V. (2012). Epigenetics: judge, jury and executioner of stem cell fate. Epigenetics 7, 823-40.

[41]   Mellor, J. (2006). It takes a PHD to read the histone code. Cell 126, 22-4.

[42]   Aasland, R., Gibson, T.J. and Stewart, A.F. (1995). The PHD finger: implications for chromatin-mediated transcriptional regulation. Trends Biochem Sci 20, 56-9.

[43]   Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C. and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410, 120-4.

[44]   Langini, C., Caflisch, A. and Vitalis, A. (2017). The ATAD2 bromodomain binds different acetylation marks on the histone H4 in similar fuzzy complexes. J Biol Chem 292, 19121.

[45]   Rack, J.G.M., Lutter, T., Kjæreng Bjerga, G.E., Guder, C., Ehrhardt, C., Värv, S., Ziegler, M. and Aasland, R. (2014). The PHD finger of p300 influences its ability to acetylate histone and non-histone targets. J Mol Biol 426, 3960-3972.

[46]   Dhalluin, C., Carlson, J.E., Zeng, L., He, C., Aggarwal, A.K. and Zhou, M.M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. Nature 399, 491-6.

[47]   Schneider, J. and Shilatifard, A. (2006). Histone demethylation by hydroxylation: chemistry in action. ACS Chem Biol 1, 75-81.

[48]   Nakamura, T. et al. (2000). huASH1 protein, a putative transcription factor encoded by a human homologue of the Drosophila ash1 gene, localizes to both nuclei and cell-cell tight junctions. Proc Natl Acad Sci U S A 97, 7284-9.

[49]   Tran, K. and Green, E.M. (2019). SET domains and stress: uncovering new functions for yeast Set4. Curr Genet 65, 643-648.

[50]   Green, E.M., Mas, G., Young, N.L., Garcia, B.A. and Gozani, O. (2012). Methylation of H4 lysines 5, 8 and 12 by yeast Set5 calibrates chromatin stress responses. Nat Struct Mol Biol 19, 361-3.

[51]   Porras-Yakushi, T.R., Whitelegge, J.P. and Clarke, S. (2006). A novel SET domain methyltransferase in yeast: Rkm2-dependent trimethylation of ribosomal protein L12ab at lysine 10. J Biol Chem 281, 35835-45.

[52]   Dillon, S.C., Zhang, X., Trievel, R.C. and Cheng, X. (2005). The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biol 6, 227.

[53]   Kanyal, A., Rawat, M., Gurung, P., Choubey, D., Anamika, K. and Karmodiya, K. (2018). Genome-wide survey and phylogenetic analysis of histone acetyltransferases and histone deacetylases of Plasmodium falciparum. FEBS J 285, 1767-1782.

[54]   Karmodiya, K., Anamika, K., Muley, V., Pradhan, S.J., Bhide, Y. and Galande, S. (2014). Camello, a novel family of Histone Acetyltransferases that acetylate histone H4 and is essential for zebrafish development. Sci Rep 4, 6076.

[55]   Roth, S.Y., Denu, J.M. and Allis, C.D. (2001). Histone acetyltransferases. Annu Rev Biochem 70, 81-120.

[56]   DesJarlais, R. and Tummino, P.J. (2016). Role of Histone-Modifying Enzymes and Their Complexes in Regulation of Chromatin Biology. Biochemistry 55, 1584-99.

[57]   Patel, D.J. (2016). A Structural Perspective on Readout of Epigenetic Histone and DNA Methylation Marks. Cold Spring Harb Perspect Biol 8, a018754.

[58]   Teske, K.A. and Hadden, M.K. (2017). Methyllysine binding domains: Structural insight and small molecule probe development. Eur J Med Chem 136, 14-35.

[59]   Filippakopoulos, P. et al. (2012). Histone recognition and large-scale structural analysis of the human bromodomain family. Cell 149, 214-31.

[60]   Haynes, S.R., Dollard, C., Winston, F., Beck, S., Trowsdale, J. and Dawid, I.B. (1992). The bromodomain: a conserved sequence found in human, Drosophila and yeast proteins. Nucleic Acids Res 20, 2603.

[61]   Dutto, I., Scalera, C. and Prosperi, E. (2017). CREBBP and p300 lysine acetyl transferases in the DNA damage response. Cell Mol Life Sci

[62]   Chiu, L.Y., Gong, F. and Miller, K.M. (2017). Bromodomain proteins: repairing DNA damage within chromatin. Philos Trans R Soc Lond B Biol Sci 372

[63]   Muller, S., Filippakopoulos, P. and Knapp, S. (2011). Bromodomains as therapeutic targets. Expert Rev Mol Med 13, e29.

[64]   Owen, D.J. et al. (2000). The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p. EMBO J 19, 6141-9.

[65]   Marchand, J.R. and Caflisch, A. (2015). Binding Mode of Acetylated Histones to Bromodomains: Variations on a Common Motif. ChemMedChem 10, 1327-33.

[66]   Eissenberg, J.C. (2001). Molecular biology of the chromo domain: an ancient chromatin module comes of age. Gene 275, 19-29.

[67]   Flanagan, J.F. et al. (2005). Double chromodomains cooperate to recognize the methylated histone H3 tail. Nature 438, 1181-5.

[68]   Eissenberg, J.C. and Elgin, S.C. (2000). The HP1 protein family: getting a grip on chromatin. Curr Opin Genet Dev 10, 204-10.

[69] Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410, 116-20.

[70] Min, J., Zhang, Y. and Xu, R.M. (2003). Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. Genes Dev 17, 1823-8.

[71] Schuettengruber, B., Bourbon, H.M., Di Croce, L. and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell 171, 34-57.

[72] Schwartz, Y.B. and Cavalli, G. (2017). Three-Dimensional Genome Organization and Function in Drosophila. Genetics 205, 5-24.

[73] Vidal, M. and Starowicz, K. (2017). Polycomb complexes PRC1 and their function in hematopoiesis. Exp Hematol 48, 12-31.

[74] Akhtar, A., Zink, D. and Becker, P.B. (2000). Chromodomains are protein-RNA interaction modules. Nature 407, 405-9.

[75] Gelbart, M.E. and Kuroda, M.I. (2009). Drosophila dosage compensation: a complex voyage to the X chromosome. Development 136, 1399-410.

[76] Grimaud, C. and Becker, P.B. (2009). The dosage compensation complex shapes the conformation of the X chromosome in Drosophila. Genes Dev 23, 2490-5.

[77] Eissenberg, J.C. (2012). Structural biology of the chromodomain: form and function. Gene 496, 69-78.

[78] Nishibuchi, G. and Nakayama, J. (2014). Biochemical and structural properties of heterochromatin protein 1: understanding its role in chromatin assembly. J Biochem 156, 11-20.

[79] Zeng, L., Zhang, Q., Gerona-Navarro, G., Moshkina, N. and Zhou, M.M. (2008). Structural basis of site-specific histone recognition by the bromodomains of human coactivators PCAF and CBP/p300. Structure 16, 643-52.

[80] Jacobs, S.A. and Khorasanizadeh, S. (2002). Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. Science 295, 2080-3.

[81] Li, H., Ilin, S., Wang, W., Duncan, E.M., Wysocka, J., Allis, C.D. and Patel, D.J. (2006). Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature 442, 91-5.

[82] Musselman, C.A. and Kutateladze, T.G. (2011). Handpicking epigenetic marks with PHD fingers. Nucleic Acids Res 39, 9061-71.

[83] Champagne, K.S. and Kutateladze, T.G. (2009). Structural insight into histone recognition by the ING PHD fingers. Curr Drug Targets 10, 432-41.

[84] Mansfield, R.E. et al. (2011). Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9. J Biol Chem 286, 11779-91.

[85] Musselman, C.A. et al. (2009). Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications. Biochem J 423, 179-87.

[86] Sanchez, R. and Zhou, M.M. (2011). The PHD finger: a versatile epigenome reader. Trends Biochem Sci 36, 364-72.

[87] Lange, M. et al. (2008). Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. Genes Dev 22, 2370-84.

[88] Zeng, L., Zhang, Q., Li, S., Plotnikov, A.N., Walsh, M.J. and Zhou, M.M. (2010). Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. Nature 466, 258-62.

[89] Grini, P.E., Thorstensen, T., Alm, V., Vizcay-Barrena, G., Windju, S.S., Jørstad, T.S., Wilson, Z.A. and Aalen, R.B. (2009). The ASH1 HOMOLOG 2 (ASHH2) histone H3 methyltransferase is required for ovule and anther development in Arabidopsis. PLoS One 4, e7817.

[90]   Hoppmann, V., Thorstensen, T., Kristiansen, P.E., Veiseth, S.V., Rahman, M.A., Finne, K., Aalen, R.B. and Aasland, R. (2011). The CW domain, a new histone recognition module in chromatin proteins. EMBO J 30, 1939-52.

[91]   Questa, J.I., Rius, S.P., Casadevall, R. and Casati, P. (2016). ZmMBD101 is a DNA-binding protein that maintains Mutator elements chromatin in a repressive state in maize. Plant Cell Environ 39, 174-84.

[92]   Andrews, F.H. et al. (2016). Multivalent Chromatin Engagement and Inter-domain Crosstalk Regulate MORC3 ATPase. Cell Rep 16, 3195-3207.

[93]   Li, S. et al. (2016). Mouse MORC3 is a GHKL ATPase that localizes to H3K4me3 marked chromatin. Proc Natl Acad Sci U S A 113, E5108-16.

[94]   Wang, Y., Reddy, B., Thompson, J., Wang, H., Noma, K., Yates, J.R. and Jia, S. (2009). Regulation of Set9-mediated H4K20 methylation by a PWWP domain protein. Mol Cell 33, 428-37.

[95]   He, F. et al. (2010). Structural insight into the zinc finger CW domain as a histone modification reader. Structure 18, 1127-39.

[96]   Liu, Y., Tempel, W., Zhang, Q., Liang, X., Loppnau, P., Qin, S. and Min, J. (2016). Family-wide Characterization of Histone Binding Abilities of Human CW Domain-containing Proteins. J Biol Chem 291, 9000-13.

[97]   Zhang, Q. et al. (2013). Structure-function analysis reveals a novel mechanism for regulation of histone demethylase LSD2/AOF1/KDM1b. Cell Res 23, 225-41.

[98]   Yang, Z. et al. (2010). AOF1 is a histone H3K4 demethylase possessing demethylase activity-independent repression function. Cell Res 20, 276-87.

[99]   Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J.R., Cole, P.A. and Casero, R.A. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119, 941-53.

[100] Perry, J. and Zhao, Y. (2003). The CW domain, a structural module shared amongst vertebrates, vertebrate-infecting parasites and higher plants. Trends Biochem Sci 28, 576-80.

[101] Liu, Y. and Huang, Y. (2018). Uncovering the mechanistic basis for specific recognition of monomethylated H3K4 by the CW domain of Arabidopsis histone methyltransferase SDG8. J Biol Chem

[102] Li, Y. et al. (2015). The histone methyltransferase SDG8 mediates the epigenetic modification of light and carbon responsive genes in plants. Genome Biol 16, 79.

[103] Ko, J.H., Mitina, I., Tamada, Y., Hyun, Y., Choi, Y., Amasino, R.M., Noh, B. and Noh, Y.S. (2010). Growth habit determination by the balance of histone methylation activities in Arabidopsis. EMBO J 29, 3208-15.

[104] Thorstensen, T., Grini, P.E. and Aalen, R.B. (2011). SET domain proteins in plant development. Biochim Biophys Acta 1809, 407-20.

[105] Xu, L., Zhao, Z., Dong, A., Soubigou-Taconnat, L., Renou, J.P., Steinmetz, A. and Shen, W.H. (2008). Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in Arabidopsis thaliana. Mol Cell Biol 28, 1348-60.

[106] Zhao, Z., Yu, Y., Meyer, D., Wu, C. and Shen, W.H. (2005). Prevention of early flowering by expression of FLOWERING LOCUS C requires methylation of histone H3 K36. Nat Cell Biol 7, 1256-60.

[107] An, S., Yeo, K.J., Jeon, Y.H. and Song, J.J. (2011). Crystal structure of the human histone methyltransferase ASH1L catalytic domain and its implications for the regulatory mechanism. J Biol Chem 286, 8369-74.

[108] Qiao, Q., Li, Y., Chen, Z., Wang, M., Reinberg, D. and Xu, R.M. (2011). The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation. J Biol Chem 286, 8361-8.

[109] Lee, Y., Yoon, E., Cho, S., Schmähling, S., Müller, J. and Song, J.J. (2019). Structural Basis of MRG15-Mediated Activation of the ASH1L Histone Methyltransferase by Releasing an Autoinhibitory Loop. Structure 27, 846-852.e3.

[110] Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 579, 3346-54.

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

[112] Csermely, P., Palotai, R. and Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35, 539-46.

[113] Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der deutschen chemischen Gesellschaft 27, 2985-2993.

[114] Ma, B., Kumar, S., Tsai, C.J. and Nussinov, R. (1999). Folding funnels and binding mechanisms. Protein Eng 12, 713-20.

[115] Dobson, C.M. (2004). Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol 15, 3-16.

[116] Tsai, C.J., Kumar, S., Ma, B. and Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. Protein Sci 8, 1181-90.

[117] Wedemayer, G.J., Patten, P.A., Wang, L.H., Schultz, P.G. and Stevens, R.C. (1997). Structural insights into the evolution of an antibody combining site. Science 276, 1665-9.

[118] Hammes, G.G., Chang, Y.C. and Oas, T.G. (2009). Conformational selection or induced fit: a flux description of reaction mechanism. Proc Natl Acad Sci U S A 106, 13737-41.

[119] Galburt, E.A. and Tomko, E.J. (2017). Conformational selection and induced fit as a useful framework for molecular motor mechanisms. Biophys Chem 223, 11-16.

[120] Uversky, V.N. (2016). p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. Int J Mol Sci 17

[121] Feng, H. et al. (2009). Structural basis for p300 Taz2-p53 TAD1 binding and modulation by phosphorylation. Structure 17, 202-10.

[122] Tong, Q. et al. (2015). Structural plasticity of methyllysine recognition by the tandem tudor domain of 53BP1. Structure 23, 312-21.

[123] Wang, L. et al. (2011). Structure of human SMYD2 protein reveals the basis of p53 tumor suppressor methylation. J Biol Chem 286, 38725-37.

[124] Chakrabarti, K.S., Agafonov, R.V., Pontiggia, F., Otten, R., Higgins, M.K., Schertler, G.F.X., Oprian, D.D. and Kern, D. (2016). Conformational Selection in a Protein-Protein Interaction Revealed by Dynamic Pathway Analysis. Cell Rep 14, 32-42.

[125] Galburt, E.A. and Rammohan, J. (2016). A Kinetic Signature for Parallel Pathways: Conformational Selection and Induced Fit. Links and Disconnects between Observed Relaxation Rates and Fractional Equilibrium Flux under Pseudo-First-Order Conditions. Biochemistry 55, 7014-7022.

[126] Vogt, A.D., Pozzi, N., Chen, Z. and Di Cera, E. (2014). Essential role of conformational selection in ligand binding. Biophys Chem 186, 13-21.

[127] Ghitti, M., Musco, G. and Spitaleri, A. (2014) NMR and Computational Methods in the Structural and Dynamic Characterization of Ligand-Receptor Interactions. In Protein Conformational Dynamics (Han, K.-l., Zhang, X. and Yang, M.-j., ed.^eds), pp. 271-304. Springer International Publishing, Cham.

[128] Śledź, P. and Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. Curr Opin Struct Biol 48, 93-102.

[129] Staby, L., O'Shea, C., Willemoës, M., Theisen, F., Kragelund, B.B. and Skriver, K. (2017). Eukaryotic transcription factors: paradigms of protein intrinsic disorder. Biochem J 474, 2509-2532.

[130] Minezaki, Y., Homma, K., Kinjo, A.R. and Nishikawa, K. (2006). Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. J Mol Biol 359, 1137-49.

[131] Blancafort, P., Segal, D.J. and Barbas, C.F. (2004). Designing transcription factor architectures for drug discovery. Mol Pharmacol 66, 1361-71.

[132] Dreier, B., Fuller, R.P., Segal, D.J., Lund, C.V., Blancafort, P., Huber, A., Koksch, B. and Barbas, C.F. (2005). Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. J Biol Chem 280, 35588-97.

[133] Lambert, M., Jambon, S., Depauw, S. and David-Cordonnier, M.H. (2018). Targeting Transcription Factors for Cancer Treatment. Molecules 23

[134] Cala, O., Guillière, F. and Krimm, I. (2014). NMR-based analysis of protein-ligand interactions. Anal Bioanal Chem 406, 943-56.

[135] Furukawa, A., Konuma, T., Yanaka, S. and Sugase, K. (2016). Quantitative analysis of protein-ligand interactions by NMR. Prog Nucl Magn Reson Spectrosc 96, 47-57.

[136] Aguirre, C., Cala, O. and Krimm, I. (2015). Overview of Probing Protein-Ligand Interactions Using NMR. Curr Protoc Protein Sci 81, 17.18.1-17.18.24.

[137] Akke, M. (2012). Conformational dynamics and thermodynamics of protein-ligand binding studied by NMR relaxation. Biochem Soc Trans 40, 419-23.

[138] Williamson, M.P. (2013). Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Reson Spectrosc 73, 1-16.

[139] Becker, W., Bhattiprolu, K.C., Gubensäk, N. and Zangger, K. (2018). Investigating Protein-Ligand Interactions by Solution Nuclear Magnetic Resonance Spectroscopy. Chemphyschem 19, 895-906.

[140] Goldflam, M., Tarragó, T., Gairí, M. and Giralt, E. (2012) NMR Studies of Protein–Ligand Interactions. In Protein NMR Techniques (Shekhtman, A. and Burz, D.S., ed.^eds), pp. 233-259. Humana Press, Totowa, NJ.

[141] Schneider, R., Blackledge, M. and Jensen, M.R. (2018). Elucidating binding mechanisms and dynamics of intrinsically disordered protein complexes using NMR spectroscopy. Curr Opin Struct Biol 54, 10-18.

[142] Bibow, S. and Hiller, S. (2019). A guide to quantifying membrane protein dynamics in lipids and other native-like environments by solution-state NMR spectroscopy. FEBS J 286, 1610-1623.

[143] Li, C., Tang, C. and Liu, M. (2013). Protein dynamics elucidated by NMR technique. Protein Cell 4, 726-30.

[144] Kay, L.E., Torchia, D.A. and Bax, A. (1989). Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. Biochemistry 28, 8972-9.

[145] Palmer, A.G. (1997). Probing molecular motion by NMR. Curr Opin Struct Biol 7, 732-7.

[146] Srb, P. et al. (2017). Triple resonance ^{15}N NMR relaxation experiments for studies of intrinsically disordered proteins. J Biomol NMR 69, 133-146.

[147] Jarymowycz, V.A. and Stone, M.J. (2006). Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. Chem Rev 106, 1624-71.

[148] d'Auvergne, E.J. and Gooley, P.R. (2006). Model-free model elimination: a new step in the model-free dynamic analysis of NMR relaxation data. J Biomol NMR 35, 117-35.

[149] Lipari, G. and Szabo, A. (1982). Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. Journal of the American Chemical Society 104, 4546-4559.

[150] Sapienza, P.J. and Lee, A.L. (2010). Using NMR to study fast dynamics in proteins: methods and applications. Curr Opin Pharmacol 10, 723-30.

[151] Woźniak-Braszak, A., Jurga, K. and Baranowski, M. (2016). The Lipari-Szabo Model-Free Analysis as a Method for Study of Molecular Motion in Solid State Heteronuclear Systems Using NMR Off-Resonance. Appl Magn Reson 47, 567-574.

[152] Daragan, V.A. and Mayo, K.H. (1999). Using the model free approach to analyze NMR relaxation data in cases of anisotropic molecular diffusion. The Journal of Physical Chemistry B 103, 6829-6834.

[153] Renner, C., Moroder, L. and Holak, T.A. (2001). Analytical solution to the Lipari-Szabo model based on the reduced spectral density approximation offers a novel protocol for extracting motional parameters. J Magn Reson 151, 32-9.

[154] Kleckner, I.R. and Foster, M.P. (2011). An introduction to NMR-based approaches for measuring protein dynamics. Biochim Biophys Acta 1814, 942-68.

[155] Gopalan, A.B., Hansen, D.F. and Vallurupalli, P. (2018). CPMG Experiments for Protein Minor Conformer Structure Determination. Methods Mol Biol 1688, 223-242.

[156] Palmer, A.G., Kroenke, C.D. and Loria, J.P. (2001). Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods Enzymol 339, 204-38.

[157] Baldwin, A.J. (2014). An exact solution for R2,eff in CPMG experiments in the case of two site chemical exchange. J Magn Reson 244, 114-24.

[158] Farber, P.J. and Mittermaier, A. (2015). Relaxation dispersion NMR spectroscopy for the study of protein allostery. Biophys Rev 7, 191-200.

[159] Ishima, R. (2014). CPMG relaxation dispersion. Methods Mol Biol 1084, 29-49.

[160] Hansen, D.F., Vallurupalli, P. and Kay, L.E. (2008). An improved 15N relaxation dispersion experiment for the measurement of millisecond time-scale dynamics in proteins. J Phys Chem B 112, 5898-904.

[161] Bieri, M. and Gooley, P.R. (2011). Automated NMR relaxation dispersion data analysis using NESSY. BMC Bioinformatics 12, 421.

[162] Sekhar, A., Rumfeldt, J.A., Broom, H.R., Doyle, C.M., Bouvignies, G., Meiering, E.M. and Kay, L.E. (2015). Thermal fluctuations of immature SOD1 lead to separate folding and misfolding pathways. Elife 4, e07296.

[163] Rennella, E., Sekhar, A. and Kay, L.E. (2017). Self-Assembly of Human Profilin-1 Detected by Carr-Purcell-Meiboom-Gill Nuclear Magnetic Resonance (CPMG NMR) Spectroscopy. Biochemistry 56, 692-703.

[164] Shi, L. and Kay, L.E. (2014). Tracing an allosteric pathway regulating the activity of the HslV protease. Proc Natl Acad Sci U S A 111, 2140-5.

[165] Whittier, S.K., Hengge, A.C. and Loria, J.P. (2013). Conformational motions regulate phosphoryl transfer in related protein tyrosine phosphatases. Science 341, 899-903.

[166] Ishima, R., Louis, J.M. and Torchia, D.A. (2001). Characterization of two hydrophobic methyl clusters in HIV-1 protease by NMR spin relaxation in solution. J Mol Biol 305, 515-21.

[167] LeBlanc, R.M., Longhini, A.P., Tugarinov, V. and Dayie, T.K. (2018). NMR probing of invisible excited states using selectively labeled RNAs. J Biomol NMR 71, 165-172.

[168] Juen, M.A., Wunderlich, C.H., Nußbaumer, F., Tollinger, M., Kontaxis, G., Konrat, R., Hansen, D.F. and Kreutz, C. (2016). Excited States of Nucleic Acids Probed by Proton Relaxation Dispersion NMR Spectroscopy. Angew Chem Int Ed Engl 55, 12008-12.

[169] Nußbaumer, F., Juen, M.A., Gasser, C., Kremser, J., Müller, T., Tollinger, M. and Kreutz, C. (2017). Synthesis and incorporation of 13C-labeled DNA building blocks to probe structural dynamics of DNA by NMR. Nucleic Acids Res 45, 9178-9192.

[170] Ding, J., Swain, M., Yu, P., Stagno, J.R. and Wang, Y.X. (2019). Conformational flexibility of adenine riboswitch aptamer in apo and bound states using NMR and an X-ray free electron laser. J Biomol NMR 73, 509-518.

[171] Sugita, Y. and Okamoto, Y. (2005). Molecular mechanism for stabilizing a short helical peptide studied by generalized-ensemble simulations with explicit solvent. Biophys J 88, 3180-90.

[172] Okamoto, Y. (2019). Protein structure predictions by enhanced conformational sampling methods. Biophys Physicobiol 16, 344-366.

[173] Kamiya, N., Higo, J. and Nakamura, H. (2002). Conformational transition states of a beta-hairpin peptide between the ordered and disordered conformations in explicit water. Protein Sci 11, 2297-307.

[174] Zhou, R., Berne, B.J. and Germain, R. (2001). The free energy landscape for beta hairpin folding in explicit water. Proc Natl Acad Sci U S A 98, 14931-6.

[175] Shaw, D.E. et al. (2010). Atomic-level characterization of the structural dynamics of proteins. Science 330, 341-6.

[176] Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H. and Shaw, D.E. (2012). Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 41, 429-52.

[177] Geng, H., Chen, F., Ye, J. and Jiang, F. (2019). Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins. Comput Struct Biotechnol J 17, 1162-1170.

[178] Pande, V.S. and Rokhsar, D.S. (1999). Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. Proc Natl Acad Sci U S A 96, 9062-7.

[179] Mohseni-Shahri, F.S., Moeinpour, F. and Nosrati, M. (2018). Spectroscopy and molecular dynamics simulation study on the interaction of sunset yellow food additive with pepsin. Int J Biol Macromol 115, 273-280.

[180] Das, S., Hazarika, Z., Sarmah, S., Baruah, K., Rohman, M.A., Paul, D., Jha, A.N. and Singha Roy, A. (2020). Exploring the interaction of bioactive kaempferol with serum albumin, lysozyme and hemoglobin: A biophysical investigation using multi-spectroscopic, docking and molecular dynamics simulation studies. J Photochem Photobiol B 205, 111825.

[181] Hollingsworth, S.A. and Dror, R.O. (2018). Molecular Dynamics Simulation for All. Neuron 99, 1129-1143.

[182] Brooks, B.R. et al. (2009). CHARMM: the biomolecular simulation program. J Comput Chem 30, 1545-614.

[183] Case, D.A. et al. (2005). The Amber biomolecular simulation programs. J Comput Chem 26, 1668-88.

[184] Ponder, J.W. and Case, D.A. (2003). Force fields for protein simulations. Adv Protein Chem 66, 27-85.

[185] Karplus, M. and McCammon, J.A. (2002). Molecular dynamics simulations of biomolecules. Nat Struct Biol 9, 646-52.

[186] Pierce, M.M., Raman, C.S. and Nall, B.T. (1999). Isothermal titration calorimetry of protein-protein interactions. Methods 19, 213-21.

[187] Perozzo, R., Folkers, G. and Scapozza, L. (2004). Thermodynamics of protein-ligand interactions: history, presence, and future aspects. J Recept Signal Transduct Res 24, 1-52.

[188] Leavitt, S. and Freire, E. (2001). Direct measurement of protein binding energetics by isothermal titration calorimetry. Curr Opin Struct Biol 11, 560-6.

[189] Ladbury, J.E. (2010). Calorimetry as a tool for understanding biomolecular interactions and an aid to drug design. Biochem Soc Trans 38, 888-93.

[190] Ladbury, J.E. and Chowdhry, B.Z. (1996). Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. Chem Biol 3, 791-801.

[191] Wiseman, T., Williston, S., Brandts, J.F. and Lin, L.N. (1989). Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem 179, 131-7.

[192] Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L. and Liu, S.Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci 17

[193] Li, H., Xie, Y., Liu, C. and Liu, S. (2014). Physicochemical bases for protein folding, dynamics, and protein-ligand binding. Sci China Life Sci 57, 287-302.

[194] Corbett, P.T., Tong, L.H., Sanders, J.K. and Otto, S. (2005). Diastereoselective amplification of an induced-fit receptor from a dynamic combinatorial library. J Am Chem Soc 127, 8902-3.

[195] Hariharan, P. and Guan, L. (2014). Insights into the inhibitory mechanisms of the regulatory protein IIA(Glc) on melibiose permease activity. J Biol Chem 289, 33012-9.

[196] Chen, Y. and Barkley, M.D. (1998). Toward understanding tryptophan fluorescence in proteins. Biochemistry 37, 9976-82.

[197] Duy, C. and Fitter, J. (2006). How aggregation and conformational scrambling of unfolded states govern fluorescence emission spectra. Biophys J 90, 3704-11.

[198] Elalaoui, A., Divita, G., Maury, G., Imbach, J.L. and Goody, R.S. (1994). Intrinsic tryptophan fluorescence of bovine liver adenosine kinase, characterization of ligand binding sites and conformational changes. Eur J Biochem 221, 839-46.

[199] Epps, D.E., Raub, T.J., Caiolfa, V., Chiari, A. and Zamai, M. (1999). Determination of the affinity of drugs toward serum albumin by measurement of the quenching of the intrinsic tryptophan fluorescence of the protein. J Pharm Pharmacol 51, 41-8.

[200] Ghisaidoobe, A.B. and Chung, S.J. (2014). Intrinsic tryptophan fluorescence in the detection and analysis of proteins: a focus on Förster resonance energy transfer techniques. Int J Mol Sci 15, 22518-38.

[201] TEALE, F.W. and WEBER, G. (1957). Ultraviolet fluorescence of the aromatic amino acids. Biochem J 65, 476-82.

[202] Hellmann, N. and Schneider, D. (2019). Hands On: Using Tryptophan Fluorescence Spectroscopy to Study Protein Structure. Methods Mol Biol 1958, 379-401.

[203] Tarafdar, P.K., Vedantam, L.V., Podile, A.R. and Swamy, M.J. (2013). Thermally stable harpin, HrpZPss is sensitive to chemical denaturants: probing tryptophan environment, chemical and thermal unfolding by fluorescence spectroscopy. Biochimie 95, 2437-44.

[204] Royer, C.A. (2006). Probing protein folding and conformational transitions with fluorescence. Chem Rev 106, 1769-84.

[205] Gorinstein, S. et al. (2000). Intrinsic tryptophan fluorescence of human serum proteins and related conformational changes. J Protein Chem 19, 637-42.

[206] Akbar, S.M., Sreeramulu, K. and Sharma, H.C. (2016). Tryptophan fluorescence quenching as a binding assay to monitor protein conformation changes in the membrane of intact mitochondria. J Bioenerg Biomembr 48, 241-7.

[207] Some, D., Amartely, H., Tsadok, A. and Lebendiker, M. (2019). Characterization of Proteins by Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS). J Vis Exp

[208] Pagès, G., Gilard, V., Martino, R. and Malet-Martino, M. (2017). Pulsed-field gradient nuclear magnetic resonance measurements (PFG NMR) for diffusion ordered spectroscopy (DOSY) mapping. Analyst 142, 3771-3796.

[209] Dehner, A. and Kessler, H. (2005). Diffusion NMR spectroscopy: folding and aggregation of domains in p53. Chembiochem 6, 1550-65.

[210] Li, D., Keresztes, I., Hopson, R. and Williard, P.G. (2009). Characterization of reactive intermediates by multinuclear diffusion-ordered NMR spectroscopy (DOSY). Acc Chem Res 42, 270-80.

[211] Stetefeld, J., McKenna, S.A. and Patel, T.R. (2016). Dynamic light scattering: a practical guide and applications in biomedical sciences. Biophys Rev 8, 409-427.

[212] Burgess, R.R. (2018). A brief practical review of size exclusion chromatography: Rules of thumb, limitations, and troubleshooting. Protein Expr Purif 150, 81-85.

[213] Ó'Fágáin, C., Cummins, P.M. and O'Connor, B.F. (2017). Gel-Filtration Chromatography. Methods Mol Biol 1485, 15-25.

[214] Sahin, E. and Roberts, C.J. (2012). Size-exclusion chromatography with multi-angle light scattering for elucidating protein aggregation mechanisms. Methods Mol Biol 899, 403-23.

[215] Tarazona, M.P. and Saiz, E. (2003). Combination of SEC/MALS experimental procedures and theoretical analysis for studying the solution properties of macromolecules. J Biochem Biophys Methods 56, 95-116.

[216] Gillis, R.B., Rowe, A.J., Adams, G.G. and Harding, S.E. (2014). A review of modern approaches to the hydrodynamic characterisation of polydisperse macromolecular systems in biotechnology. Biotechnol Genet Eng Rev 30, 142-57.

[217] Li, Y., Weiss, W.F. and Roberts, C.J. (2009). Characterization of high-molecular-weight nonnative aggregates and aggregation kinetics by size exclusion chromatography with inline multi-angle laser light scattering. J Pharm Sci 98, 3997-4016.

[218] Price, W.S. (1997). Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part 1. Basic theory. Concepts in Magnetic Resonance 9, 299-336.

[219] Price, W.S. (1998). Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part II. Experimental aspects. Concepts in Magnetic Resonance 10, 197-237.

[220] Macchioni, A., Ciancaleoni, G., Zuccaccia, C. and Zuccaccia, D. (2008). Determining accurate molecular sizes in solution through NMR diffusion spectroscopy. Chem Soc Rev 37, 479-89.

[221] Kandiyal, P.S., Kim, J.Y., Fortunati, D.L. and Mok, K.H. (2019). Size Determination of Protein Oligomers/Aggregates Using Diffusion NMR Spectroscopy. Methods Mol Biol 2039, 173-183.

[222] Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-91.

[223] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25, 1605-12.

[224] Li, Z., Natarajan, P., Ye, Y., Hrabe, T. and Godzik, A. (2014). POSA: a user-driven, interactive multiple protein structure alignment server. Nucleic Acids Res 42, W240-5.

[225] Marblestone, J.G., Edavettal, S.C., Lim, Y., Lim, P., Zuo, X. and Butt, T.R. (2006). Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. Protein Sci 15, 182-9.

[226] Kapust, R.B. and Waugh, D.S. (1999). Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein Sci 8, 1668-74.

[227] Keller, R.L.J. (2005) Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignmented.^eds)

[228] Martin, S.R., Schilstra, M.J. and Siligardi, G. (2011) Chapter 7 Circular Dichroism. In Biophysical Approaches Determining Ligand Binding to Biomolecular Targets: Detection, Measurement and Modelling ed.^eds), pp. 226-246. The Royal Society of Chemistry

[229] Martin, S.R. and Schilstra, M.J. (2008). Circular dichroism and its application to the study of biomolecules. Methods Cell Biol 84, 263-93.

[230] Mårtensson, L.G., Jonasson, P., Freskgård, P.O., Svensson, M., Carlsson, U. and Jonsson, B.H. (1995). Contribution of individual tryptophan residues to the fluorescence spectrum of native and denatured forms of human carbonic anhydrase II. Biochemistry 34, 1011-21.

[231] Huang, F. and Nau, W.M. (2003). A conformational flexibility scale for amino acids in peptides. Angew Chem Int Ed Engl 42, 2269-72.

[232] Schneider, M., Fu, X. and Keating, A.E. (2009). X-ray vs. NMR structures as templates for computational protein design. Proteins 77, 97-110.

[233] Koehler Leman, J., D'Avino, A.R., Bhatnagar, Y. and Gray, J.J. (2018). Comparison of NMR and crystal structures of membrane proteins and computational refinement to improve model quality. Proteins 86, 57-74.

[234] Zhang, Y. et al. (2019). Mechanism for autoinhibition and activation of the MORC3 ATPase. Proc Natl Acad Sci U S A

[235] Zhang, Y., Bertulat, B., Tencer, A.H., Ren, X., Wright, G.M., Black, J., Cardoso, M.C. and Kutateladze, T.G. (2019). MORC3 Forms Nuclear Condensates through Phase Separation. iScience 17, 182-189.

[236] Dong, G., Ma, D.P. and Li, J. (2008). The histone methyltransferase SDG8 regulates shoot branching in Arabidopsis. Biochem Biophys Res Commun 373, 659-64.

II

The *Arabidopsis* (ASHH2) CW domain binds monomethylated K4 of the histone H3 tail through conformational selection

Olena Dobrovolska[1] , Maxim Brilkov[1] , Noelly Madeleine[1,2] , Øyvind Ødegård-Fougner[3] ,
Øyvind Strømland[2] , Stephen R. Martin[4] , Valeria De Marco[5] , Evangelos Christodoulou[4] ,
Knut Teigen[2] , Johan Isaksson[6] , Jarl Underhaug[7] , Nathalie Reuter[7] ,
Reidunn B. Aalen[8] , Rein Aasland[8] and Øyvind Halskau[1]

1 Department of Biological Sciences, University of Bergen, Norway
2 Department of Biomedicine, University of Bergen, Norway
3 Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany
4 Structural Biology Science Technology Platform, Francis Crick Institute, London, UK
5 King's College London, UK
6 Department of Chemistry, The Arctic University of Tromsø, Norway
7 Department of Chemistry, University of Bergen, Norway
8 Department of Biosciences, University of Oslo, Norway

Chromatin post-translational modifications are thought to be important for epigenetic effects on gene expression. Methylation of histone N-terminal tail lysine residues constitutes one of many such modifications, executed by families of histone lysine methyltransferase (HKMTase). One such protein is ASHH2 from the flowering plant *Arabidopsis thaliana*, equipped with the interaction domain, CW, and the HKMTase domain, SET. The CW domain of ASHH2 is a selective binder of monomethylation at lysine 4 on histone H3 (H3K4me1) and likely helps the enzyme dock correctly onto chromatin sites. The study of CW and related interaction domains has so far been emphasizing lock–key models, missing important aspects of histone-tail CW interactions. We here present an analysis of the ASHH2 CW-H3K4me1 complex using NMR and molecular dynamics, as well as mutation and affinity studies of flexible coils. β-augmentation and rearrangement of coils coincide with changes in the flexibility of the complex, in particular the η1, η3 and C-terminal coils, but also in the β1 and β2 strands and the C-terminal part of the ligand. Furthermore, we show that mutating residues with outlier dynamic behaviour affect the complex binding affinity despite these not being in direct contact with the ligand. Overall, the binding process is consistent with conformational selection. We propose that this binding mechanism presents an advantage when searching for the correct post-translational modification state among the highly modified and flexible histone tails, and also that the binding shifts the catalytic SET domain towards the nucleosome.

Introduction

Chromatin structure, and thereby gene expression, is dynamically regulated by post-translational modifications (PTMs) on the N-terminal histone tails protruding from nucleosomes. These PTMs include methylation, acetylation, phosphorylation and many other modifications. They are thought to constitute a histone code, where unique combinations of PTMs are associated with specific effects on gene expression [1]. The PTMs are established and altered by 'writer' and 'eraser' enzymes that add and remove modifications, respectively, and the ensuing pattern of PTMs on the histone tails is interpreted by 'reader' protein domains [2]. Methylation of histone N-terminal tail lysine residues is carried out by methyltransferases that harbour a catalytic SET domain, and target lysine residues can either be mono-, di- or trimethylated on the ε-nitrogen [3]. The modified lysine residue can be demethylated by one of two classes of lysine demethylases, either a flavin adenine dinucleotide-dependent oxidase or a Fe (II) and α-ketoglutarate-dependent hydroxylase [4]. Methylated lysine residues can be recognized by members of the 'royal family' of protein domains, which are the chromo, MBT, chromo barrel, Tudor and PWWP domains [5]. It is also known that some WD40 domains and PHD fingers can recognize unmodified or methylated lysine residues [6]. The CW domain family has also been identified as another family of proteins that can recognize methylated lysine residues both in animals [7] and in plants [8].

The CW domain family is named after and identified by conserved cysteine and tryptophan residues found in its primary structure. Proteins containing the domain have been found in higher-order plants, vertebrates and vertebrate-infecting parasites [9,10]. The CW domain is found in proteins in combination with other domains such as PWWP and SET, and it has also been identified in chromatin remodellers and demethylases [7–13]. The role of the CW domain in most proteins is to recognize and bind to methylated histone H3 (H3) N-terminal tails at the K4 position (H3K4meX, where X is the number of methyl groups). Depending on the protein, the CW domain displays a different specificity for the degree of methylation [7,8,14]. The other mammalian CW domain-containing proteins ZCWPW1, ZCWPW2, MORC3 and MORC4 display specificity for H3K4me2/me3 [7,11]. CW containing multidomain proteins found in animals and plants are not orthologues, and their overall domain organizations are different [9].

The small, flowering plant *Arabidopsis thaliana* codes for an enzyme named ASHH2 which methylates position K36 on H3. This 1759-amino acid-long enzyme contains a CW domain that binds specifically to monomethylated H3K4 followed by an AWS domain and then a SET domain where the methyltransferase activity resides [8]. ASHH2 is a major regulator of growth and development in *Arabidopsis*, as mutations in ASHH2 result in dwarf plants with alterations in flowering time, fertility, branching, organ identity, programmed cell death and pathogen defence [15,16]. ASHH2 di- and trimethylates H3K36 from their monomethylated state, and in loss-of-function mutant plants, a global reduction in H3K36me3/me2 and a corresponding increase in H3K36me1 are linked to an early flowering phenotype [17,18]. Further pleiotropic effects include reduced fertility as well as homeotic changes in floral organs in plants where the ASHH2 gene is mutated [19,20]. *Arabidopsis* also contains another H3K36 methyltransferase ASHH1, but in contrast to ASHH2, this protein lacks a CW domain and based on the severe pleiotropic effects of the ASHH2 mutant the two proteins are not redundant [18].

In recent years, several structures of CW domains in their *apo* and *holo*, that is their unbound and bound states, have been solved [7,8,14,21]. A shared feature of interaction is the conserved tryptophans scaffolded by a β-sheet which provides part of the pocket that accepts the methylated lysine. Another feature of CW domains not highlighted by earlier investigations is the fact that their tertiary structure comprises just a few short secondary structure elements, while flexible coils dominate the rest of the fold. Disorder and flexibility are prevalent both in histone tails and in proteins involved in chromatin remodelling, and recent bioinformatic studies have highlighted the need for investigations focusing on functional flexibility [22,23]. As far as we can determine, the structural biology of CW domains has not been investigated systematically with

functional flexibility in mind. We are therefore interested in whether CW binding is coupled to changes in structure, stability and mobility at the level of individual amino acids, and secondary and tertiary levels of organization. To explore this question, a comprehensive structural and dynamic analysis of the ASHH2 CW-H3K4me1 complex using NMR, molecular dynamics (MD) and lower-resolution techniques were performed, followed by mutagenesis of residues implicated in functional flexibility to assess their effect on affinity. From our analysis, ASHH2 CW emerges as a dynamic domain that undergoes a global reorganization to become more compact but still remains relatively flexible. We found that the mechanism of binding relies on protein flexibility and is best described by a conformational search for the correct histone modification. CW domains have not yet been reported to act through such mechanisms, and it is possible that this mechanism confers an advantage in the highly complex and dynamic chromatin environment.

Results

The CW domain reorganizes to a more compact form upon binding to H3K4me1

The functional domains of ASHH2 lie within long stretches of amino acids that are predicted to be disordered or to contain orphan secondary structure not associated with any known fold. Among the folded domains, CW is N-terminally situated and is also flanked by disordered segments (Fig. 1A). The NMR structure of the unbound ASHH2 CW domain was determined by Hoppmann *et al.* [8] using a construct denoted CWs (PDB code: 2L7P, Fig. 1A). For structure determination of the complex, a screen of 20 additional constructs was subsequently performed, initially aimed at finding expressible and high-affinity binders amenable for co-crystallization with the bound ligand. Of these, the constructs denoted CW33, CW37 and CW42, all covering the evolutionary conserved residues of the CW domain (Fig. 1A), where high-affinity binders of the histone tail mimick H3K4me1 (ARTKme1QTARY, with one methyl substitution at the ζ-position of K4), as determined by an intrinsic fluorescence-based binding assay (K_ds in the range of 0.21–0.85 μM Fig. 1B-D). Their affinities for H3K4me2 and H3K4me3 were also determined (sequence as for H3K4me1, but with two and three methyl substitutions at K4, respectively), and the K_d values ranged from 0.7 to 7.3 μM (Fig. 1D, examples of binding curves in Fig. S1). Several crystallization attempts were unsuccessful, and consequently, the decision was

made to characterize the complex using NMR. For subsequent work, CW42 construct was selected as it expressed well and had an affinity indistinguishable from that of the longer CWs [24].

One noticeable property of the fluorescence binding studies was that the ligand caused a λ_{max} shift towards shorter wavelengths of the spectrum (Fig. 1B). Such behaviour is characteristic of tryptophans entering a more solvent-protected environment within a fold [25]. This makes sense if the ligand covers the two tryptophans of the binding site upon binding, if there is a consolidation of the overall fold of the domain upon binding or both. We, therefore, compared the temperature stability of the *apo-* and *holo-*forms using intrinsic tryptophan fluorescence, as well as estimating their hydrodynamic sizes. We found that the T_m of the CW42-H3K4me1 complex was about 6 °C higher than that for uncomplexed CW42 (T_m of 58.0 ± 1.4 °C vs 64.4 ± 1.0 °C, Fig. 2A,B). For size estimations, size-exclusion chromatography with multi-angle light scattering (SEC-MALS) was used as well as diffusion constant measurements using pulsed-field NMR. MALS data showed lower effective hydrodynamic radius, that is the elution time on an SEC column increases, even as the molecular mass of the complex goes up (Fig. 2C). The *holo-*state also shows a shoulder towards the unbound state. Curiously, increasing the ligand concentration beyond further beyond twofold excess did not remove this feature. The NMR diffusion rate measurements collected for the protein and the complex support this observation. The observed diffusion rates correspond to roughly $2.1 \cdot 10^{-10}$ m^2·s^{-1} and $2.6 \cdot 10^{-10}$ m^2·s^{-1} for the *apo-* and *holo-*forms (Fig. 2D). Using the Stokes–Einstein relationship [26], these diffusion rates correspond to approximate hydrodynamic diameters of 1.7 and 1.4 nm, respectively. The T_m, MALS and diffusion data support the view that the domain undergoes compaction and stabilization of its structure upon ligand binding. In the following, we elucidated how this was reflected in the structure and dynamics of CW42 at a more detailed level.

Apo- vs *holo-*structural comparison shows C-terminal α1-helix differences and posthelical coil involvement in binding

The most suitable approach for exploring the detailed in-solution molecular changes associated with the binding is comparing the NMR structures of the *apo-* and *holo-*forms of CW. Previously, we published the structure of the free ASHH2 CW domain [8], and now, we present the solution structure of the CW42-H3K4me1 complex. The structure was submitted to

the Protein Databank (PDB code: 6QXZ), and a summary of NMR structural statistics and an ensemble representation of the 20 energy-minimized conformers

can be viewed in Fig. 3A,B. In Fig. 4A, the *apo*-structure (PDB code: 2L7P, Hoppmann *et al.* [8]) and the *holo*-structure are superimposed. The chemical shifts
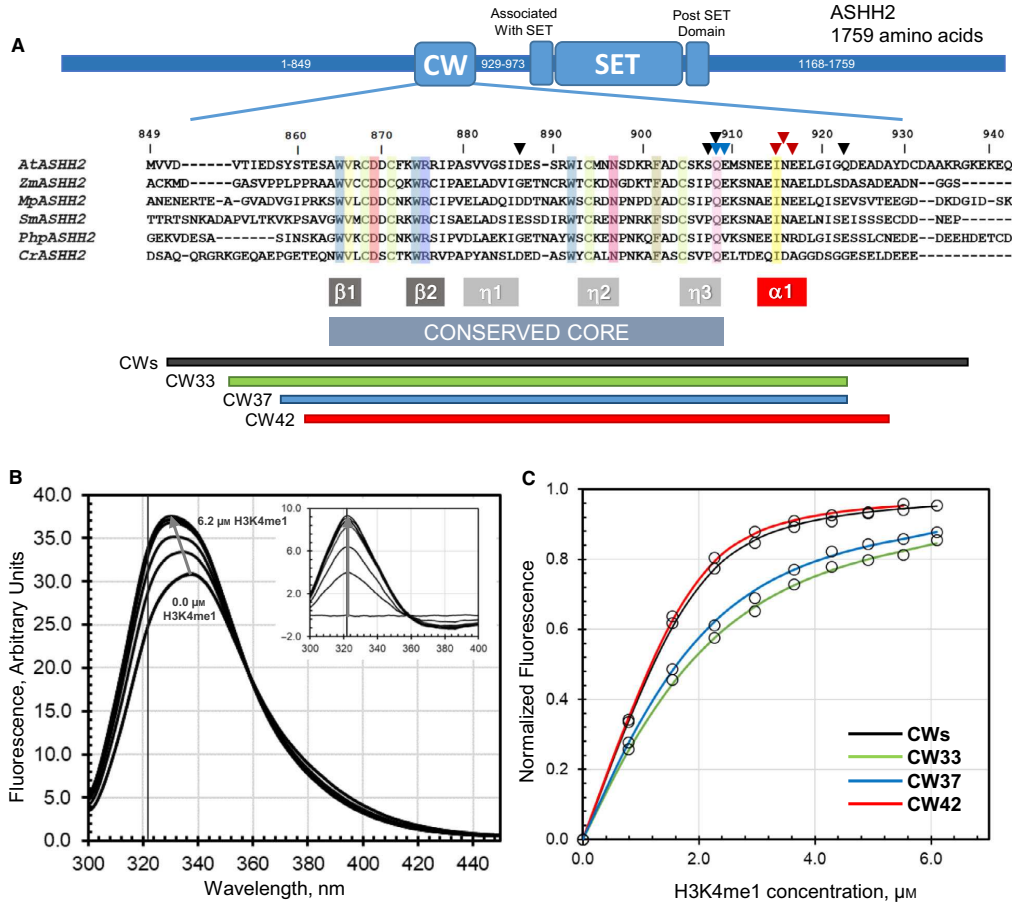
Fig. 1. The CW domain of ASHH2 binds H3K4me1 with high affinity. (A) ASHH2 domain organization, with multiple sequence alignment of the evolutionarily conserved CW domains from dicotyledonous flowering plants (*Arabidopsis thaliana,* Q2LAE1), monocotyledonous (maize, Zea mays, A0A1D6HAE7), liverworts (*Marchantia polymorpha,* A0A2R6W143), spikemosses (*Selaginella moellendorffii,* D8SGM1), mosses (*Physcomitrella patens,* A0A2K1L195) and green algae (*Chlamydomonas reinhardtii,* A0A2K3DEA3). The codes in parenthesis identify the UniProt entries used in the alignment, which was generated using ClustalW. Mutations performed in this study (▼), the Hoppmann *et al* study (▼) and the Liu *et al.* study (▼) are indicated on the sequence. The CW core that is conserved across species, as well as secondary structure elements, is shown below the multiple sequence alignment, as is a subselection of 3 of the 20 constructs initially prepared as possible crystallization candidates and how they relate to the main sequence and the CW construct used in the Hoppmann *et al.* paper. (B) Representative intrinsic tryptophan fluorescence spectra (excitation wavelength 290 nm) used for binding assays of CW42 to H3K4me1. The vertical line indicates the wavelength at which emission intensities were used, as determined in the inset. Inset: ΔFluorescence intensity where spectra of CW42 in the absence of any ligand are subtracted from spectra of increasing amounts of H3K4me1. Units are otherwise the same as in the main panel. The wavelength at which the ΔFluorescence intensity was maximal was 322 nm for CW42, and in the range of 319–322 nm for the other constructs. (C) K_d determinations of CWs (—), CW42 (——), CW37 (——) and CW33 (——). Normalized ΔFluorescence intensity values at the wavelength as determined in the inset of Panel B were plotted against ligand concentrations (0.0–7.2 μM H3K4me1). For CWs and CW42, the wavelengths used were 321 and 322 nm, respectively. Protein concentrations were constant throughout anyone titration, but could vary somewhat from construct to construct (always within 2.0–2.4 μM). The data were fitted using nonlinear least-square methods to Eq. 1, yielding three K_d values in each instance. (D) Tabulated affinities (in μM) of CW33, CW37, CW42 and CWs binding to H3K4me0/1/2/3. N.a., not applicable, n.d., not determined. Values are given as means of each individual K_d determined within sets of matching parallels. Error bars are one standard deviation, $n = 3$.

and the position of the side chain of K4me1 are suggestive of cation–π interactions with the indole group of W874 [27] (Fig. 4A). Four residues comprise the top of the hydrophobic pocket hosting the monomethylated lysine: I915, L919, I921 and Q923 (Fig. 4B,D). All of these residues display NOE connectivities with the ligand, indicating stable contacts in solution (see examples in Fig. 3C). A comparison between the backbone of the *apo-* and the *holo*-NMR structures does not indicate a large reorganization of the protein domain upon binding. The most prominent difference is in the η1-loop that changes position to interact with the N-terminal of the ligand (backbone displaced by up to 8 Å). There is also a minor reposition of the C-terminal α1-helix to accommodate the ligand (backbone displaced by ~ 2 Å), as predicted in the Hoppmann *et al.* paper [8].

A crystal structure of the ASHH2 CW domain in complex with H3K4me1 has also recently been published by Liu *et al.* (PDB code: 5YVX) [14]. The η1 region, as well as the I915 and L919, was identified as crucial for ligand binding, and the interactions were discussed in terms of lock–key arguments for the N-terminal part of the ligand. To exploit all existing structural data, we also include this crystal structure in our analysis. Comparing the two known *holo*-structures shows that their backbones match closely except at the α1-helix (Fig. 4C, RMSD between residues S863-Q908 of the *holo*-forms is 0.913 Å), a part of the domain that is crucial for correct binding [8,14]. In the Liu *et al.* structure, this helix is longer than the *holo*-NMR structure, and the structure terminates immediately after the helix. Moreover, the C_αs is displaced by roughly 4 Å towards the C-terminal end of the helix

(Fig. 4C) relative to our NMR structure. In order to make the protein domain crystallize, Liu *et al.* introduced an E917A mutation into the α1-helix at a site that is partially conserved (Fig. 1A). While this mutation still allowed the ligand to bind with a somewhat reduced affinity [1.3 ± 0.2 mM (WT) vs 2.79 ± 0.36 mM (E917A)], it may together with the lack of the C-terminal coil have caused the α1-helix to become displaced relative to the NMR structure of the wild-type version of the domain. Liu *et al.* [14] report that N916, located within the α1-helix and positioned next to the E917A mutation, is crucial for both binding of the methylated ligand and the methylation-dependent binding profile of the CW domain. In our structure, we find no evidence for stable contacts between it and the ligand, and at the same time, the ligand is in our case surrounded by the key residues (Fig. 4B,D). We also note that between L919, I921 and Q923, there are two glycines (Fig. 4D). These make no contacts with the ligand, but allow the key residues space and flexibility they need to pack tightly around the ligand. This is markedly different than the configuration found in the crystal structure (Fig. 4E).

The CW42-H3K4me1 complex is stabilized by intermolecular β-augmentation, the α1-helix and the C-terminal coil

Several related structures, including that of MORC and zinc finger CW, report β-augmentation as being part of the binding mechanism, that is that an intermolecular β-sheet is formed upon binding [7,21,28]. To examine whether this is a stable feature of the CW42-H3K4me1 complex, three replicates of 50-ns
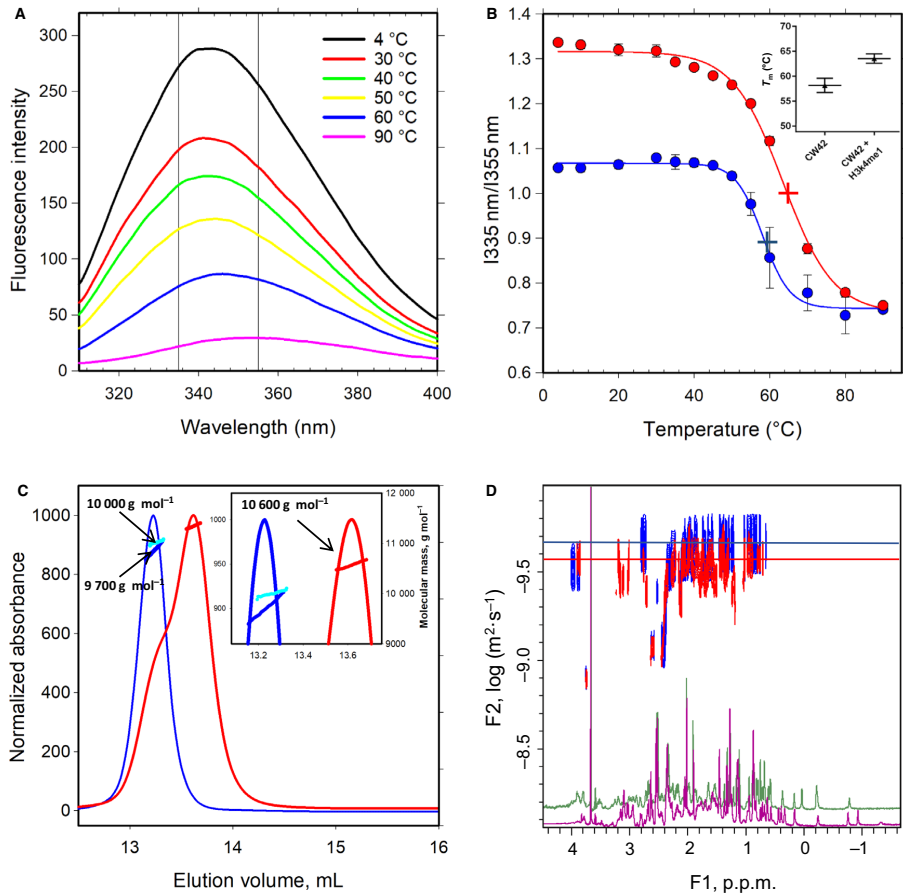
Fig. 2. CW42 becomes more stable upon binding to H3K4me1. (A) Representative intrinsic tryptophan fluorescence spectra of CW42 in the absence of H3K4m1 at 4–90 °C. Tryptophans were excited at 295 nm, and the emission scanned from 310 to 450 nm. Vertical lines at 335 and 355 nm indicate the wavelengths at which intensity values were the intrinsic tryptophan signal dominated by folded and unfolded protein states, respectively. (B) Thermal denaturation profile of bound and unbound CW42. The I335 nm/I355 nm ratios derived from fluorescence data in the presence (●) and absence (●) of H3K4me1 ligand were plotted vs temperature. Each data point represents the mean of three parallels, and error bars are shown as one standard deviation where these exceed the size of the symbols. The data series for the bound (——) and unbound (——) situation were then fitted (nonlinear least squares) to a 4-parameter sigmoidal expression, yielding the midpoint of the denaturation curve, T_m, as an output in the presence (+) and absence (+) of H3K4me1. Inset: summary of T_m for CW42 with and without ligand bound. Error bars show 95% confidence interval of the fits in the main panel. (C) SEC-MALS elution profiles of CW42 in the presence (——) and absence (——) of H3K4me1, where each profile is shown as molecular mass (kDa) vs elution time (min). The molecular mass (g·mol^{-1}) for each elution as determined by static light scattering is shown as red and blue dots for the ligand present and absent situation, respectively. The average molecular masses for each peak are indicated (→). (D) Diffusion measurements of CW42 in the presence (red contours) and absence (blue contours) of H3K4me1. Horizontal axes represent the projection of ^{1}H experiments using bipolar gradient sets separated by diffusion delays and 3-9-19 water suppression. The vertical axis is the logarithm of the diffusion coefficient (D, m^2·s^{-1}). Cross-peaks represent fits of peaks extracted by fitting the 64 ^{1}H experiments to the decay function given by Eq. 3 in the Supplementary Information. Only selected peaks from the nonexchanging, upfield region were used to estimate the mean D, as either buffer components or the ligand do not influence this spectral region. The log D value for CW42 in the presence and absence of ligand is indicated by horizontal lines, (——) and (——), respectively. Representative 1D ^{1}H spectra acquired in the presence (magenta) and absence (green) of the ligand are shown at the bottom of the panel.

A

Restraints used in structure calculation	Number
Total number of NOE distance restraints	1056
Intra-residual NOEs	285
Short-range, $|i\text{-}j| = 1$	330
Medium-range, $1 < |i\text{-}j| < 5$	204
Long-range, $|i\text{-}j| \geq 5$	147
Intermolecular NOEs	74
Number of upper distance limits for Zn^{2+}	8
Number of lower distance limits for Zn^{2+}	8
TALOS N ϕ/ψ dihedral angle restraints	94
Structure statistics, 20 conformers	
CYANA target function value (Å^2)	2.78 ± 0.22
Maximal distance constraint violation (Å^2)	0.36 ± 0.04
Maximal torsion angle constraint violation (Å^2)	0.54 ± 0.49
AMBER energies in implicit solvent (kcal/mol)	-3867.0074
OneDep – Ramachandran statistics	
Residues in favorable regions (%)	89
Residues in allowed regions (%)	9
Residues in outlier regions (%)	2
Root mean square deviation to average coordinates (Å)	
N, C^α, C' (860–910)	0.36 ± 0.09
Heavy atoms (860–910)	0.97 ± 0.14

Fig. 3. NMR structure of the CW42-H3K4me1 complex. (A) NMR restraints and structural statistics for CW42-H3K4me1 complex. (B) The structural ensemble of the 20 minimized NMR-derived structures, backbone Cα atoms aligned to the medoid structure, conformer 15. (C) Strip plots for residues Q923 (HE22/NE22), I921 (HD1/CD1) and W874 (HZ/CZ) derived from the filtered-edited 3D NOESY experiments showing intra- and intermolecular NOE connectivities to both the CW domain and the bound ligand. Graphical representations of structures were prepared in PYMOL 1.5 (Schrödinger, New York, NY, USA).

MD simulations were performed using a representative conformation from our NMR structure, and Liu *et al.*'s crystal structure. The replicates were identical except for different initial velocities, and we find that both structures of the complex are stable and able to hold the ligand within its binding pocket as evidenced by their RMSD values throughout the simulation (Fig. 5A, Table S1). The high variations observed from 12 to 20 ns in the RMSD values of the NMR structure simulations are due to a displacement of the C-terminal coil in one replicate (the C-terminal coil is receding from the ligand). The C-terminal coil comes back to interact with the ligand from 20 ns until the end of the simulation. Hydrogen bond analysis of the complex combined with secondary structure analysis of the ligand along the MD simulation trajectory of
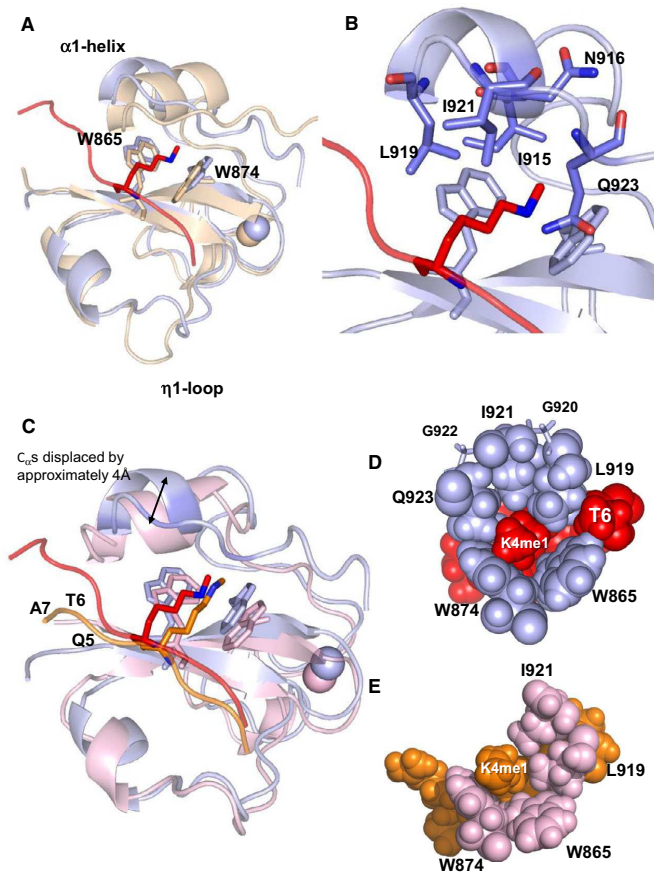
Fig. 4. In-solution structure of CW42-H3K4me1 complex. (A) Structural comparison of the CW42 in the free (PDB code: 2L7P, in beige) and bound state (in light blue, PDB code: 6QXZ). The bound state is represented as the medoid structure, conformer 15. RMSD between the *apo*- and *holo*-structures is 1.6 Å. The ligand is presented in red, and the side chain of the H3K4me1 resides between W865 and W874 of CW domain. Cartoons are rendered with 0.35 transparency setting in PYMOL to increase visibility of key elements. (B) Section of panel A, highlighting the key C-terminal residues forming the hydrophobic pocket – I915, L919, I921 and Q923, clustering around the methylated lysine. N916 is also indicated, but this residue does not contact the ligand. (C) Structural superposition of the CW domain in the bound state – NMR structure (in light blue) and X-ray (in magenta, PDB code: 5YVX). The RMSD value for the core residues 863–908 is 0.913 Å. The H3 residues are positioned differently in the NMR (red) and crystal structure (orange). Ligand residues Q5, T6 and A7 are indicated. The overall backbone of the structures is similar except for the α1-helix, where C_α is displaced by about 4 Å. (D) Space-fill representation of the ligand-binding site, showing key interactions between the protein in light blue and the ligand in red. (E) As in D, but for the crystal structure, protein in pink and ligand in orange. Graphical representations of structures were prepared in PYMOL 1.5 (Schrödinger, New York, NY, USA).

both structures indicates a stable intermolecular β-sheet (Fig. 5B). The secondary structure analysis of the ligand in both structures shows some difference in the residues involved in β-sheet augmentation through the simulation (residues A2 to Q5 and T3 to A6 of the X-ray and NMR structures, respectively, Fig. 5C,D). The same trend has been observed in the

MD simulation replicates. These intermolecular β-sheet interactions are reinforced by hydrogen bonds between the ligand and the CW domain side chains.

In the crystal structure, the ligand is oriented differently for residues Q5-A7, probably due to the misorientation of the α1-helix (Fig. 4C). This is, in turn, a likely consequence of the E917A crystallization mutant
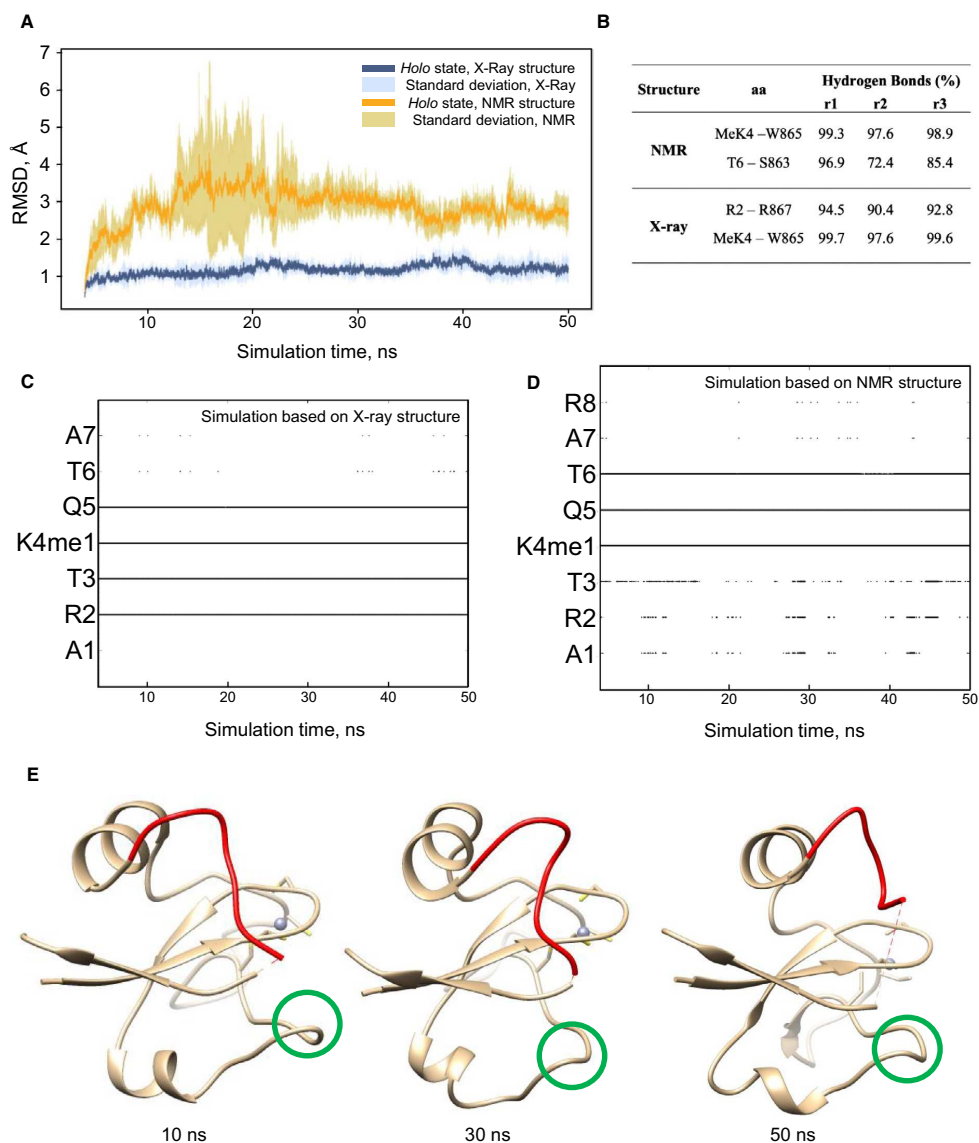
Fig. 5. Induction of β-sheet and rearrangement of C-terminal coil upon ligand binding. (A) Average RMSD evolution (dark colour) and standard deviation (light colour) of the NMR WT CW42-H3K4me1 structure (orange) and crystal structure of the E917A mutant (blue) during the last 46 ns of simulations. (B) Hydrogen bond occupancy between the main chain of amino acids (aa) involved in the intermolecular β-sheet through the last 46 ns of MD simulation performed on the NMR representative structure and the X-ray structure for each replicate (r1, r2 and r3). These hydrogen bonds are present in the initial structures and maintained through MD simulations. (C) and (D) Secondary structure analysis of the ligand along MD simulations performed on the X-ray structure (C) and the NMR structure (D). The results are shown for one replicate of each structure. Points/lines indicate β-strand conformation at a given time during the last 46 ns of the MD simulation. (E) Rearrangement of the C-terminal coil along the simulation. Snapshots from 10, 30 and 50 ns show the C-terminal coil (in red) in interaction with the ligand. The A879-S889 coil that also shifts up to interact with the ligand is encircled in green. Graphical representations of structures were prepared in Chimera.

and the shortened C-terminal part. Although the sequence of the crystal structure ends at I921, just after the α1-helix, both our affinity data for shortened domains (CW33/37) and NMR data suggest that this part of the domain is relevant for binding. Tellingly, there are numerous NOE cross-peaks indicative of stable links from this coil to both the cores of the CW42 domain and the bound ligand (Fig. 3C). For instance, as many as three ligand contacts are mediated by I921, and six are mediated by Q923. In the NMR structure, Q923 resides within a coil, absent in the crystal structure, that appears as an ensemble of fluctuating conformations. Our MD simulations indicate that there is a tendency for this coil to move towards the N-terminal part of the ligand, and together with the η1 region interacts with the ligand but from the opposite side (Fig. 5E). The C-terminal coil's rearrangement with respect to the ligand is observed from around 10 ns and is maintained until the end of the simulation. This observation has been confirmed by the replicates.

Complexation modulates the flexibility of key binding elements

Molecular motions are important for protein function in general and ligand binding in particular [29,30]. We have observed in this study that CW42 responds to binding both at a global level and at a more detailed level. To characterize the motional changes triggered by binding, we compared the local internal motions in the *apo*- and the *holo*-states of the protein using NMR. Steady-state heteronuclear ^{1}H-^{15}N NOE values, and R_1 and R_2 relaxation rates are sensitive to high-frequency motions (10^8–10^{12} s^{-1}) occurring at ps-ns timescale, with R_2 also having contributions from much slower processes occurring at μs-ms timescale [31]. The analysis of these parameters in the free and bound state provides information about the protein local backbone mobility change upon ligand binding (Fig. 6A-C). Overall, all the residues show NOE values near 0.9, indicating backbone motions at the ns scale. Outliers exist in the β2 sheet, and η1 and η3 loop regions. The η1 region and the post-α1-helix flexible loop, including Q923, undergo changes restricting motions upon binding (Figs 4A, 6A and 5E). The R_1 parameter is generally lower for the *holo*-state, indicating an overall stabilization, while the R_2 parameter shows outliers in the D886-R890 interval, as well as M910 and L919.

To further exploit these data, the three relaxation parameters were combined with the structure of the complex using the Lipari–Szabo model-free formalism [31]. Output parameters of this analysis are the order parameter, S^2, reflecting the amplitude of the internal motions on the ns timescale, the effective correlation time for the internal motions, τ_e, and the conformational exchange rate on the μs to ms timescale, R_{ex} (Fig. 6D-F). Overall, the order parameter values, S^2, indicate a quite flexible protein, especially for the *apo*-state. Even its most stable parts have an S^2 value between 0.9 and 0.8, somewhat lower than what is usual for folded proteins and closer to proteins with fluctuating structures [32,33]. Differences between the *apo*- and *holo*-states are found in the loop regions of the protein, post-η1 in particular, but also in the β2-sheet and the α1-helix and its posthelical coil. Overall, the S^2 values suggest a consolidation of the fold upon binding. Residues V882-S889 of the η1 region are restricted upon binding (Fig. 6D). The values of the local correlation time, τ_e, are rather low throughout the protein (within 0.8 ns), indicating overall protein flexibility (Fig. 6E). The R_{ex} parameter, where available, suggests that *apo*-CW undergo conformational exchange on the μs-ms timescale, often associated with conformational shifts related to function [34]. The majority of observed rates are below 2 s^{-1} (Fig. 6F). There are notable outliers, again to be found in the β2-sheet, and near η1 and η3. These residues, with larger values than 2 s^{-1}, are R875, I877, G883, D886, E887, D898, M910, E917, L919 and A926. For these residues and in the α1-helix, we generally observe higher R_{ex} values for the *apo*-form, indicating a slowing down of conformational fluctuations also at the ms-μs timescales. Of these residues, only E887 and L919 make direct contact with the ligand in at least one of the available *holo*-structures, suggesting that lock–key type formalism is not sufficient to understand this binding process.

In contrast, binding through conformational selection may explain why we observe these outliers. Such binding mechanisms postulate that the *apo*-state is flexible and fluctuating and that a small population of the bound conformation exists in equilibrium, also when the ligand is not present [35]. When the ligand is present, binding occurs by stabilizing the pre-organized conformation corresponding to the bound state [36]. Although the preceding Lipari–Szabo model-free analysis implicates dynamic elements in the binding event, the timescales associated with conformational selection are better assessed using relaxation–dispersion NMR experiments. In brief, this approach isolates the contribution of ms-s conformational exchange towards R_2 relaxation [37,38]. We performed these experiments at 600 and 850 MHz, and performed global data fits
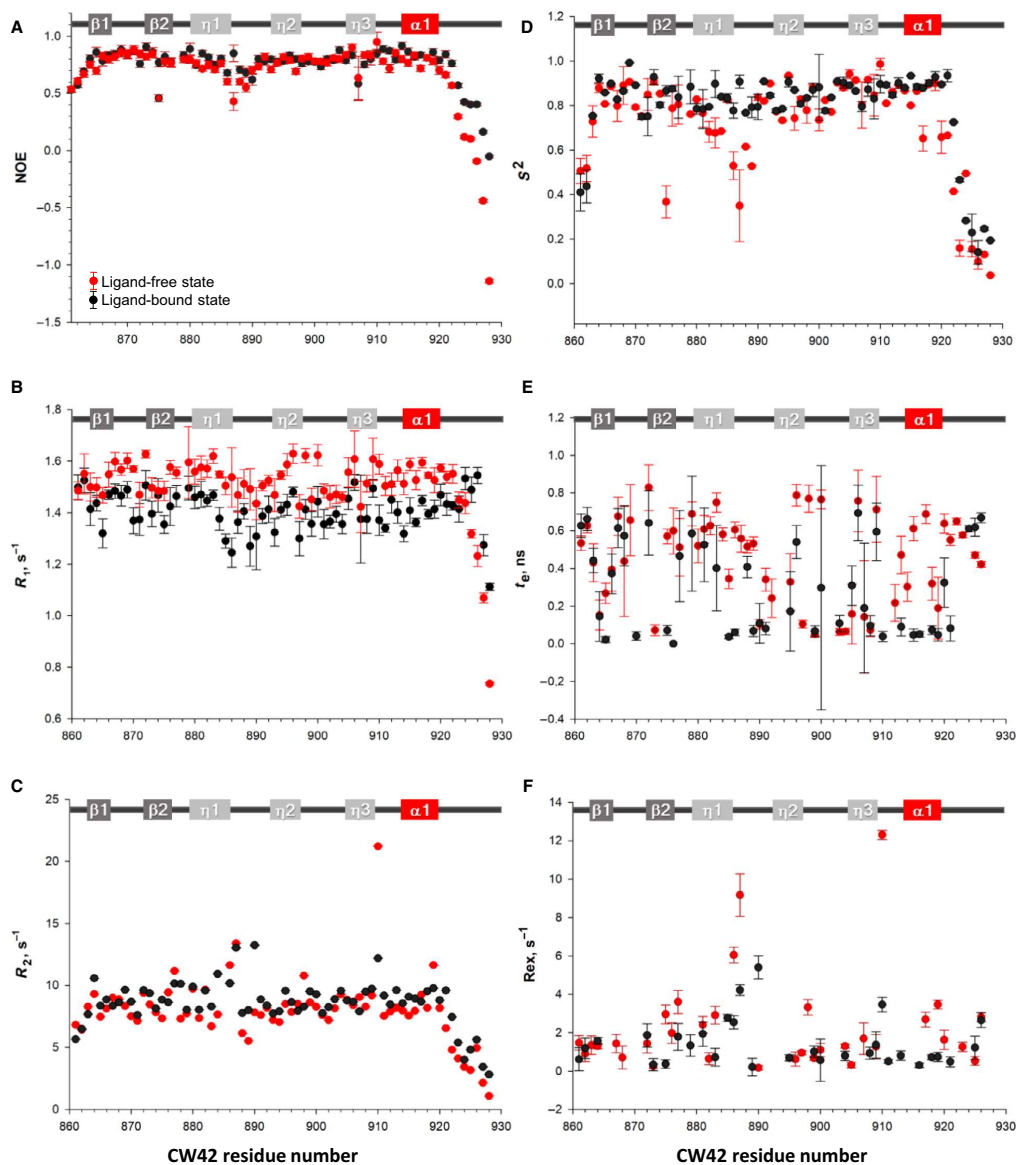
Fig. 6. Relaxation NMR data and Lipari–Szabo model-free dynamic analysis for CW42 in its free (●) and bound to H3K4me1 (•) states. (A) Steady-state ^{1}H-^{15}N NOEs. (B) R_1 relaxation rates. (C) R_2 relaxation rates. Model-free parameters derived for the free and bound states: (D) order parameter S^2, (E) local correlation time τ_e and (F) conformational exchange rate Rex. The CW42 secondary structure is indicated at the top of each panel. Errors are estimated by Monte Carlo simulations (CI, 95%).

using the NESSY software made by Bieri *et al* for this purpose [37]. In NESSY, relaxation dispersion profiles are fitted to models identifying protein motions related to no exchange (i.e. no movement at this timescale), slow-exchange or the fast-exchange limit, for each residue with a backbone amide. The program picks

C

Residue in CW42	K_{ex}, s^{-1}	Error	Model	AICc	Chi^2
R875	n.a.	n.a.	M1	35.68	33.68
D886	728	412.0	M3	92.27	81.27
S889	682	141.0	M7	56.71	45.71
M894	n.a.	n.a.	M1	56.71	27.17
N896	531	99.2	M2	22.67	13.67
S907	16	4.3	M3	397.74	386.74
M910	18	71.9	M3	151.55	140.55
L919	1712	341.8	M2	24.57	15.57

Fig. 7. The ms dynamics of CW. (A) Examples of relaxation dispersion data accumulated at 600 MHz and 850 MHz field strengths (at 300 and 310K), and then globally fitted in NESSY by Monte Carlo simulation. The fitting procedure attempts to fit three models and is selected using χ^2 and AICc tests measuring the goodness of fits. The models M1-M3 allowed are no exchange (M1), two states in the fast-exchange limit exchange (M2) and two states in slow exchange (M3). (B) Box plots of the *apo-* and *holo*-distribution of all R_{ex} output values found by the NESSY fits. Circles designate outliers, whiskers are the highest and lowest nonoutlier values in the data sets, and upper and lower box border is the third and first data quartile, respectively. The data median is indicated by the black bar. Mean values when outliers are removed are 3.1 ± 0.64 rad·s^{-1} and 2.0 ± 1.55 rad·s^{-1} for the *apo-* and *holo*-situation, respectively. The difference between the two data sets is significant ($P < 0.05$, Student's *t*-test, one-tailed, heteroscedastic). (C) Examples of extracted K_{ex} values from fits performed by NESSY, along with the model selected and their χ^2 and AICc scores. Errors are estimated by Monte Carlo simulations (CI, 95%). (D) Plot of determined K_{ex} values for the *apo-* and *holo*-forms of CW42. Values with errors exceeding 500 s^{-1} are not included in this plot. (E) K_{ex} values associated with the *apo*-state plotted onto the unbound NMR structure of CW (2L7P). Residues with quicker motions are drawn using thicker stick representation and more intense red colour. Residues exhibiting slow-exchange behavio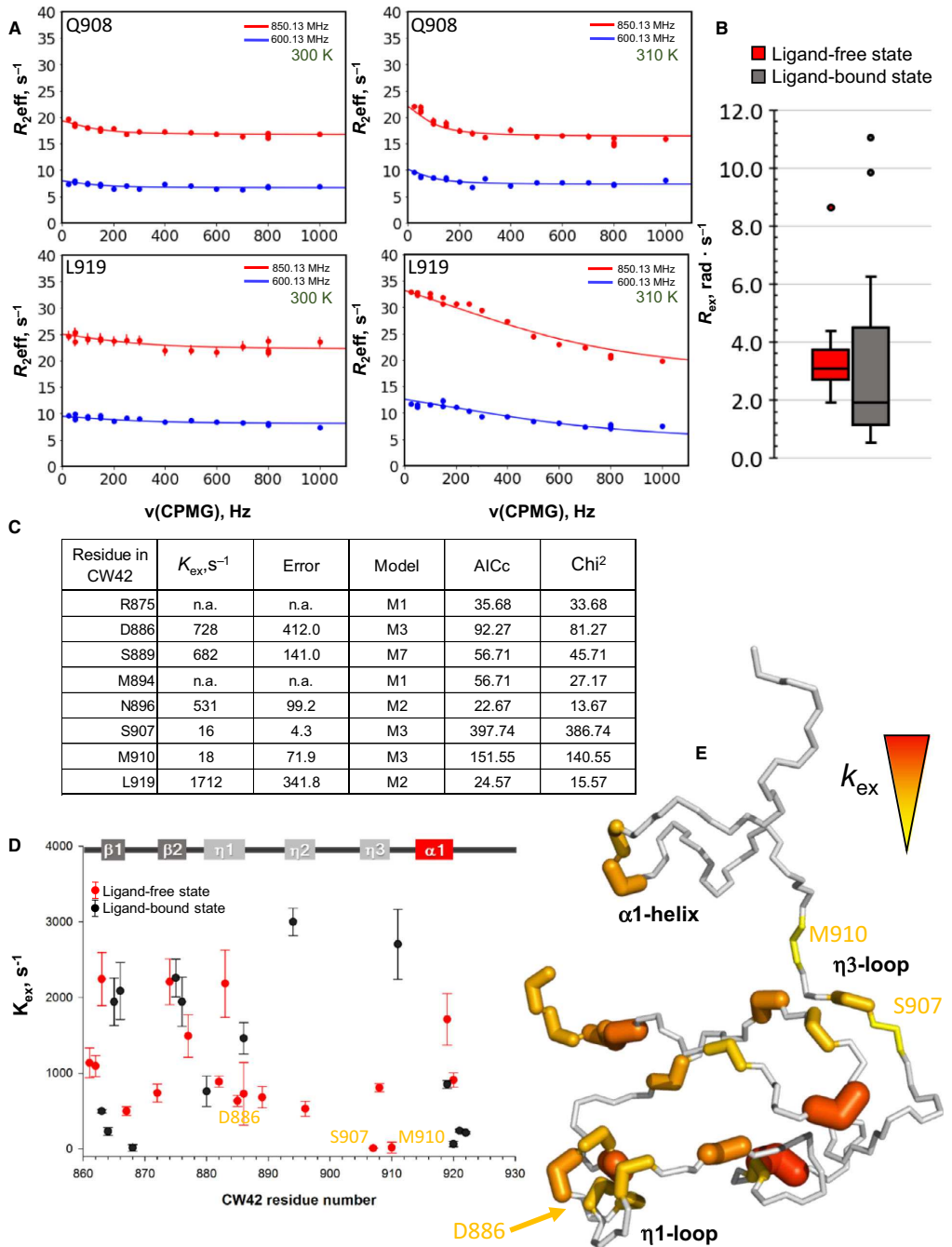ur are indicated in yellow and thinner stick representation. For full, tabulated summaries of NESSY output, see Table S4 and S5. The graphical representation of the structure was prepared in PYMOL 1.5 (Schrödinger, New York, NY, USA).

models using an approach avoiding overfitting based on χ^2 and AICc goodness-of-fit scoring functions [39]. Examples of dispersion curves for the *apo*-state at 300K and 310K are shown in Fig. 7A. In the NESSY models, R_{ex} is an output parameter that can be interpreted as the contributions of relatively slow protein motions towards the total R_2 relaxation behaviour. As the ligand binds, there is a significant (*t*-test, one-sided, heteroscedastic, $P < 0.05$) lowering of R_{ex} values (Fig. 7B). We interpret this as an overall quenching of this type of motions upon binding, a behaviour that is expected for binding through conformational selection [40]. For fast-exchange limit (M2) and slow-exchange (M3) residues, tabulated examples of residues displaying relaxation dispersion behaviours consistent with different models of exchange behaviour are presented in Fig. 7C, along with extracted K_{ex} values representing rates of conformational exchange. A plot of K_{ex} values for the *apo-* and *holo*-states, where residues with large errors (more than 500 s^{-1}) removed, is presented in Fig. 7C. All NESSY-selected models, along with their output values and associated χ^2 scores, can be viewed in Table S4 and S5.

In the following analysis, we focus on the residues that exhibit slow exchange, as this type of behaviour is an indication of minor populations that may be relevant for binding [40]. Three residues, D886, S907 and M910, are in slow exchange (M3), according to the NESSY selection. D886 belongs to the η1 loop that shifts towards the ligand upon binding (Fig. 4A), and its actual K_{ex} value is more similar to fast-exchanging residues fitting the M2 model (Fig. 7C). S907 and M910 are, interestingly, located in the η3 loop which is known to affect binding, is fairly conserved, but also shows ASHH2-specific variations (Fig. 1A). Because of this, we still include S907 in our analysis even though its associated AICc and χ^2 values were notably high for S907. We also show, *vide infra*, that S907 has

significant ($P < 0.05$) effect on the binding of H3K4me1, and abolishes binding of ligands with K4me2 and K4me3 altogether. The rate of conformational exchange, K_{ex}, is very slow and similar for these two residues, around 16–18 s^{-1} (Fig. 7C). The behaviour and location of the slow-exchanging residues in the η3 loop which leads up to the α1-helix (Fig. 7E) is suggestive of a mechanism where the flexibility of the loop allows the α1-helix to sample the binding conformation, which is then consolidated if the correct ligand is present.

The coils flanking the α1-helix are mediators of binding and flexibility

The findings presented above are consistent with a role for protein conformational sampling in binding. We, therefore, returned to our MD simulations and compared the root mean square fluctuation (RMSF) and the radii of gyration (R_g) of the bound and unbound states. In the simulations, the R_g values of the *apo*-state and *holo*-state overlap at a time, suggesting that the *apo*-state can sample the bound conformation (Fig. 8A). RMSF calculations remove the time dimensions in the simulations and allow this measurement for local flexibility to be plotted onto the domain backbone. Overall, the domain fluctuates from 50% to 30% less in the bound form than in the free form (Figs 8B and S3A). The η1, η2 and η3 regions where molecular rearrangement takes place upon binding appear as outliers with increased and unaffected flexibility. The MD and NMR dynamics data generally match; both approaches indicated a restriction of the *holo*-structure as well as showing outliers in the same regions. Both results corroborate the initial low-resolution characterization of ligand binding (Figs 1 and 2).

Our simulations also indicate a difference between the two *holo*-structures as determined by NMR and
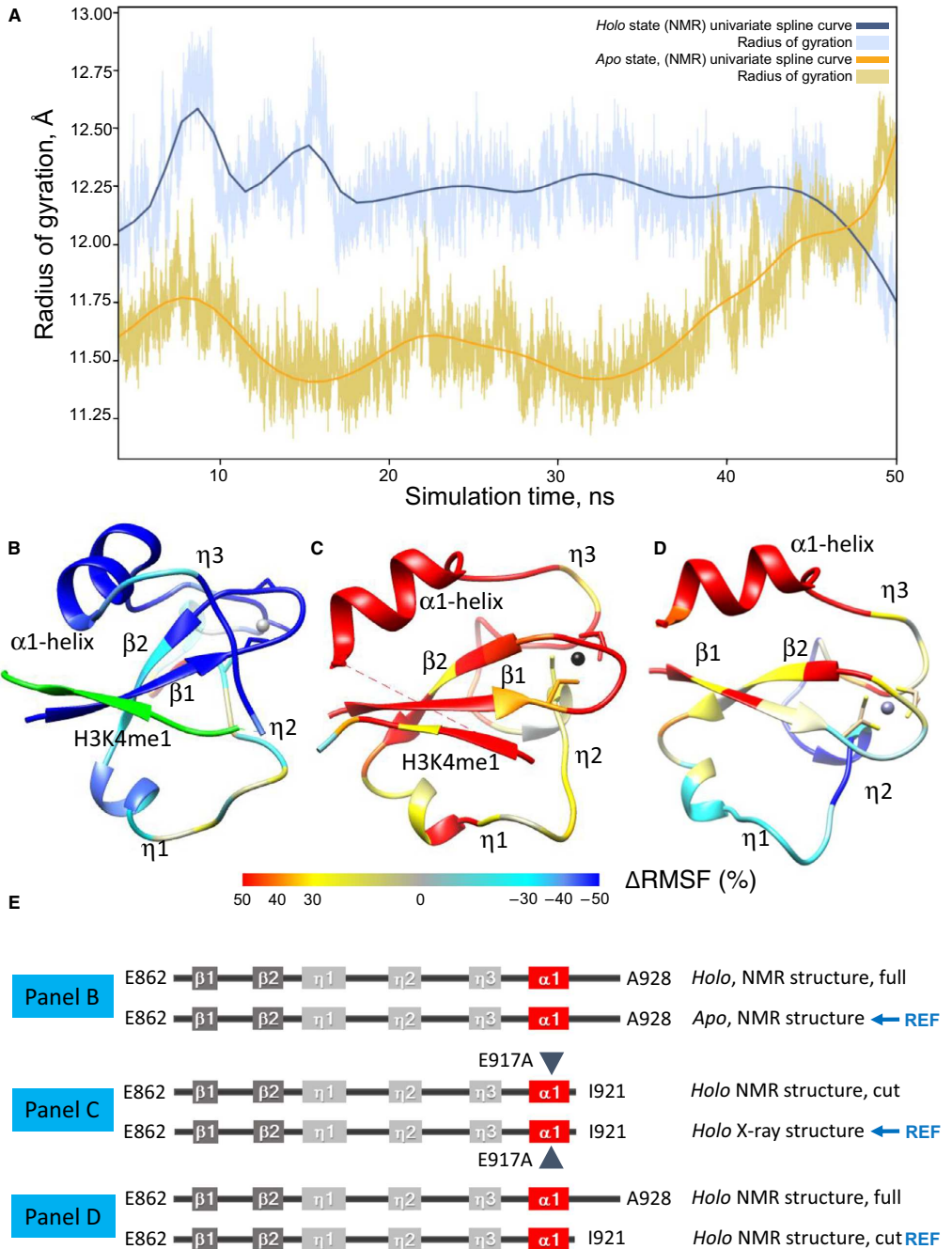
Fig. 8. The flexibility of the domain is influenced by ligand binding and the posthelical coil. (A) The calculated radius of gyration (Å) as a function of simulation time for the *apo* (orange)- and *holo* (blue)-forms of ASHH2 CW for one of the MD simulation replicates ($n = 3$). The lighter colours show the radii of gyrations, and the darker line (univariate spline curve) highlights their overall trend. (B, C, D) Comparative flexibilities of CW structures. Relative changes in flexibilities are calculated by comparing the RMSF values of a given state to a reference state (per cent change). For all panels, cyan to blue colours represent parts of the complex that are less flexible than the reference state. Yellow to red colours represent the parts of the complex that is more flexible than the reference state. For the extreme red and blue colours, the |ΔRMSF| ≥ 50%. The results are shown for one replicate of each state. (B) RMSF differences between the *apo*-state as the reference vs the *holo*-state of CW42, using the NMR structures available. Structures used were 2L7P and 6QXZ, modified to match in length, and see panel E. The ligand is represented in light green. (C) RMSF differences between the *holo*-form crystal structure (mutant) as the reference state vs the NMR *holo*-form NMR structure. Structures used were 5YVX and 6QXZ, where the latter NMR structure also has been modified with the E917A mutation and a shortened the C-terminal coil to match of the 5YVX structure, and see panel E. (D) RMSF differences between the full-length NMR *apo*-complex as the reference state vs the same structure without the C-terminal coil. Structures used were 6QXZ, and a version where the latter has the C-terminal coil removed, and see panel E. In this last case, the ligand is not stably locked in the binding site and is therefore not shown. (E) Schematic overview of the structures used in panels B, C and D. The blue triangle indicates the crystallization mutant E917A. The reference model referred above is indicated by REF. Graphical representations of structures were prepared in Chimera.

crystallography. However, it was unclear whether this is caused by the inherent differences in the crystal vs the NMR model, or whether the sequence difference is the cause. We, therefore, modified the NMR-structure sequence in such a way that it matched the crystal structure sequence (Fig. 8E), and simulated these two states for 50 ns. The method used to resolve the structure does seem to impact the complex flexibility since the NMR structure fluctuates much more than the crystal structure during MD simulations (30–50% more, Figs 8C and S3B). This suggests that Liu *et al.* structure is restricted to a very limited conformational space.

The truncation of constructs scored for H3Kme1 binding affinity indicated that removing the C-terminal adversely affects binding affinity (Fig. 1, constructs CW33 and CW37). The NMR data and MD simulations also supported a role for the C-terminal post-α1-helix coil in binding (Figs 6A,D and 5E). To evaluate the impact of this part of the sequence on the complex flexibility, we compared the RMSF of the complex with and without the C-terminal coil (for a schematic overview of how structures are compared, see Fig. 8E). The results indicate a high (ΔRMSF ≥ 50%) and a moderate (30% < ΔRMSF < 50%) increase in flexibility in the NMR structure lacking the C-terminal coil, residing in the α1-helix and β-sheet, respectively (Figs 8D, S3C). In contrast, the η1 and η2 regions experience a stabilization upon removal of the C-terminal coil. Moreover, the ligand is not stable within the binding site of the truncated structure simulations, underlying the importance of the C-terminal coil for the ligand stability. Nevertheless, the ligand is stable in the X-ray structure simulations and in the comparative simulation with its NMR counterpart (see Fig. 8E schematics), suggesting that the E917A mutation plays an important, albeit artificial, role in the complex stability.

Our structural and dynamics results, as well as sequence alignment and earlier work, strongly implicate the η1 and η3 loops. A graphical summary of this is presented in Figs 1A and 9A. η3 notably contains conserved residues with variation relatable to the loop flexibility, such as Pro to Ser variations. The *holo*- or *apo*-structures do not show much difference in these places (Fig. 4A), yet the conservation pattern and the MD and NMR results related to mobility suggest that these residues might be involved in regulating the equilibrium position of the α1-helix in the free and bound situation. An effect on H3K4me1 binding, or ability to differentiate between methylation states, would be particularly interesting since these residues are in a coil without directly contacting the ligand, and the backbone trace of the *apo*- and *holo*-forms is essentially the same (Fig. 4A).

We note that Hoppmann *et al.* mutated two residues in this region, Q908A and E909A, and both mutations effectively abolished binding in pull-down assays ([8] and Fig. 9A). In reference to this work and our current dynamics data (derived from NMR, as well as the RMSF analysis), we further probe the involvement of S907 and Q908 in modulating the binding affinities of H3K4me1-3 using Isothermal calorimetry (ITC). For unmutated CW42 interacting with H3K4me1, the binding constant and stoichiometry of interaction were determined to be $K_d = 1.09 \pm 0.21$ μM and $n = 0.85 \pm 0.05$, respectively. The reaction is enthalpy-driven (ΔH of -91.63 ± 8.14 kJ·mol^{-1}), while the ΔS term is negative (ΔS = -192.98 ± 25.97 J·mol^{-1}·K) (Fig. 9B-D), in support of the net ordering of the complex reported above. Since the Q908A polar to aliphatic mutation has already been performed by Hoppmann *et al.*, we did a structurally conservative Q908E mutation that converts this polar residue into a charged one. This results in a 17-fold drop in affinity (Fig. 9B, C,

$P \ll 0.01$). For S907, we note that Ser is an amino acid associated with a high amount of flexibility, only surpassed in this regard by Gly according to amino acid flexibility rankings [41]. We, therefore, mutated this residue to confer both higher (S907G) and lower (S907P) flexibility to the η3 coil preceding the α1-helix

D

H3K4 me1	CW	D886A	S907P	S907G	Q908E	Q923A	CW to H3K4me1-T6A
K_d, µM	1.09 ± 0.21	1.54 ± 0.15	3.22 ± 0.59	1.80 ± 0.51	18.75 ± 1.02	No reproducible isotherms found ($n = 4$)	No reproducible isotherms found ($n = 12$)
n	0.85 ± 0.05	1.09 ± 0.01	1.07 ± 0.07	1.02 ± 0.06	0.89 ± 0.04		
ΔH, kJ/mol	−91.63 ± 8.14	−72.85±1.41	−72.00 ± 4.34	−78.92 ± 5.17	−73.88 ± 1.20		
ΔS, J/mol·K	−192.98 ± 25.97	−133.07±4.64	−135.92 ± 15.26	−154.50 ±19.51	−157.30 ± 3.83		

E

H3K4 me2	CW	D886A	S907P	S907G
K_d, µM	For wild type CW affinities for H3K4m2 and H3K4me3, see Figure 1D.	5.15 ± 0.39	No reproducible isotherms found ($n = 3$)	No reproducible isotherms found ($n = 3$)
n		0.96 ± 0.03		
ΔH, kJ/mol		−84.05 ± 0.57		
ΔS, J/mol·K		−180.63±2.57		

F

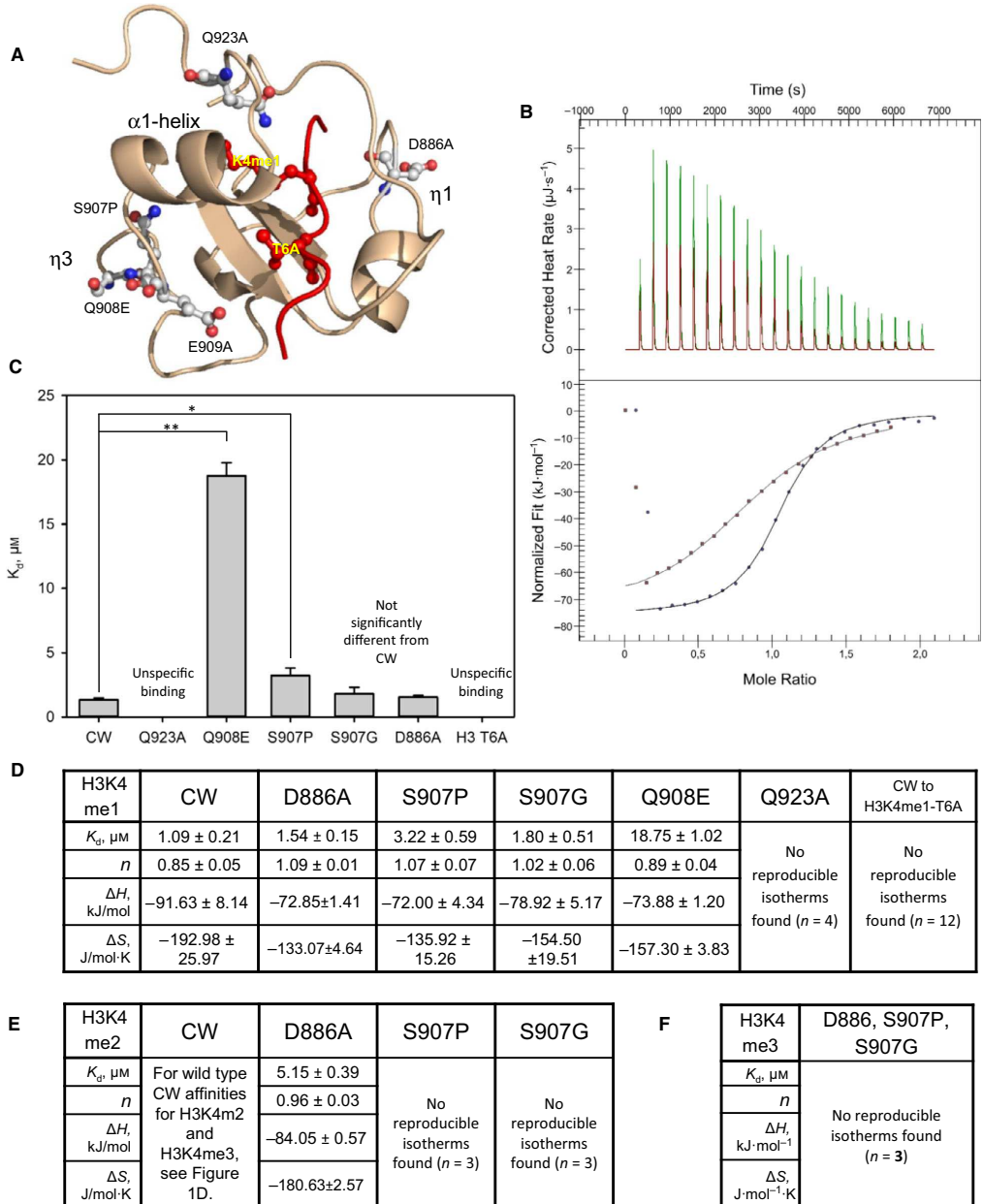| H3K4 me3	D886, S907P, S907G
K_d, µM	No reproducible isotherms found ($n = 3$)
n	
ΔH, kJ·mol⁻¹	
ΔS, J·mol⁻¹·K	

Fig. 9. Mutants in coils and their effects on the binding properties of the CW42-H3K4me1 complex. (A) Ribbon representation of the CW42-H3K4me1 complex. The ligand is shown as sticks in red; mutated residues in this study and Hoppmann *et al.* study are shown in element-specific colouring. There are no direct contacts between the mutated residues and the ligand, except for Q923. The graphical representation of the structure was prepared in PYMOL 1.5 (Schrödinger, New York, NY, USA). (B) Representative ITC data interacting with H3K4me1. Top panel displays representative corrected heat rates plotted against time for CW (——) and the CW Q908E mutant (——) titrated against the ligand. The bottom panel represents the normalized peak areas (CW, •; Q908E, ■) plotted vs the ligand/protein mole ratio. (C) Bar plot representation of mean K_ds where error bars represent one standard deviation ($n = 3$). Statistical significance of pairwise differences as indicated (*t*-test, *, $P < 0.05$, **, $P < 0.01$). (D) Derived thermodynamic parameters for CW42 and mutants binding to H3K4me1. All values are averages based on three determinations, with errors given as standard deviations. N is the binding stoichiometry. (E) Derived thermodynamic parameters for CW42 and mutants binding to H3K4me2. All values are averages based on three determinations, with errors given as standard deviations. N is the binding stoichiometry. (F) Attempts to determine binding and thermodynamic parameters for CW42 mutants binding to H3K4me3 failed after the indicated number of parallels.

and observed that affinity was lowered somewhat. For the restricting S907P mutation, K_d increased almost threefold from 1.09 ± 0.21 μM to 3.22 ± 0.29 μM (*t*-test, $P < 0.05$), while S907G resulted in no significant increase (1.80 ± 0.51 μM). However, when we also investigate the ability of H3K4me2 and H3K4me3 to bind these mutations, we were unable to produce binding isotherms, suggesting that these flexibility-modulating mutations are involved in allowing CW to differentiate between methylation states (Fig. 9E,F). In contrast, the D886A mutation, positioned in the η1 loop that moved towards the ligand upon binding in the structure (Fig. 4A), does not affect H3K4me1 binding much (Fig. 9C) and is still able to bind H3K4me2 at reduced (~ 5-fold) affinity (Fig. 9E).

To examine the effect of the post-C-terminal coil on binding, we also designed a Q923A mutant, which formed NOE contacts with the ligand (Fig. 3C), and performed ITC. These affinity measurements showed that the mutant destroyed specific binding, making it impossible to produce reliable isotherms (Fig. S2), confirming the importance of this residue for binding. We also investigated the effect of altering the H3 peptide. Beyond the central K4me1, the importance of residues 1–3 has been determined by Liu *et al.* [14]. However, the structure presented here suggests that T6 is involved in specific contacts with the ligand site, including L919 and W865 (Fig. 4D). When performing a T6A amino acid substitution in the H3K4me1 peptide, ITC measurements failed to produce reliable binding isotherms (Fig. S2). It may also be relevant that H3 can be phosphorylated at T6 [42]. Such a PTM modification would destroy complementarity (Fig. 4D), and likely abolish binding.

Discussion

Several low-resolution techniques used to assess CW42 to H3K4me1 interaction suggested that a reorganization, compaction and overall slowing down of

dynamics takes place upon binding. However, only a limited amount of reorganization was apparent when comparing the *apo*- and *holo*-structures. Investigation of the dynamic behaviour of the bound and unbound states using NMR and MD provided a more comprehensive picture. The *apo*-state is relatively flexible on the ns timescale (Figs 6A,D and 8B), with several hotspots (η1 and η3) also showing tendencies for dynamics on the μs-ms timescale (Fig. 6F). Using relaxation dispersion experiments, we were able to extend our view of the domain's dynamics to the ms-s timescale, where in particular the η3 loop displayed indications of concerted slow exchange at a rate of about $16–18$ s^{-1} (Fig. 7C). Significantly, η3 mutations at the S907, Q908 and E909 positions adversely affect binding, especially for H3K4me2 and H3K4me3 (Fig. 9E,F). The MD simulations exhibit a variation in the *apo*-form's RMSD across the simulation that is consistent with a dynamic loosening and compaction of the structure at equilibrium. Although each replicate spans 50 ns only, the repeats all show the same tendencies, and the NMR dynamics spanning ns-ms timescales corroborate this. In sum, this suggests that the equilibrium *apo*-state is less compact than determined structures indicate, and may sample a compact, less flexible *holo*-state.

Compaction behaviour like this has been linked to disorder–order transitions [43]. Our ITC data show that there is a significant entropic cost (-192 J·mol^{-1}·K) associated with binding. The entropic cost must be at least partially related to ordering of flexible elements, as ligand binding should lead to entropically favourable desolvation of the hydrophobic residues of the ligand and the binding pocket. While the *apo*-state of CW is certainly folded, there are enough mobile elements for a disorder-to-order transition to occur. Such binding-induced ordering events are relatable to both induced fit and conformational selection mechanisms of binding, where two fairly flexible entities mutually explore conformational space conductive to binding [35,40]. The availability of these states at ambient conditions is

related to both the folding behaviour and flexibility of the protein [44], and both NMR and MD data all point towards CW42 having sufficient flexibility to populate a spectrum of conformations at timescales ranging from ns to ms (Figs 3B and 6-8). For the η3 region, we also find limited evidence that the *apo*-CW exchanging behaviour samples the same state. Throughout these analyses, the η1 and η3 coils consistently display interesting behaviour, and residues therein also show effects on CW binding behaviour when mutated. Taken together, this is consistent with a CW binding behaviour that is suggestive of conformational selection.

There also appears to be an element of induced fit [45], where equilibrium in the bound states of both the ligand and CW42 is shifted towards new structural elements, most notably the β-sheet augmentation observed in the MD simulations (Fig. 5). In addition to the contacts made by the side chain of K4me1, β-sheet augmentation may help dictate the ligand sequence specificity towards the ASHH2 CW domain, as is the case for a number of other instances [46]. The side chains of R2, K4me1 and T6 all orient towards CW42 in the β-sheet and contribute to the final complementarity of the bound state by intercalating as shown in Fig. 4. Conversely, T3, Q5 and A7 are oriented away. This recognition, based on the alternating orientation of side chains, is similar to that reported for the MORC3 CW domain, where β-augmentation is also part of the binding mechanism [21]. Yet, these zipper-like fits are not enough for recognition and binding, as the unmethylated ligand will not bind. Conformational selection may offer an explanation. A lack of or incorrect methylation on K4 will not trigger β-sheet augmentation, and therefore, the zipper-like complementarity will not arise from the ensemble of conformations. The mediator of this could be changes in the α1-helix equilibrium position near the K4me1 side chain, mediated by the η3 loop, followed by rearrangement of coils (the η1 and the C-terminal). In particular, the η1 loop, although fitting models for slow exchange in the relaxation dispersion analysis (Fig. 7B), displays quite fast dynamics, and the D886A mutation is not as crucial for binding of the methylated ligands (Fig. 9). One interpretation of this would be that the first step of binding is mediated by the slowest category of exchange, related to η3 and α1-helix dynamics, and is followed by a more rapid consolidation step mediated by the η1 loop.

Our mutation studies on the coils flanking the α1-helix do show that it is possible to affect binding without directly contacting the ligand or the binding site. Since coils are flexible entities, this suggests that they play the role of tensile regulators of the complementary fit between the ligand, and I915, L919 and Q923. As far as

we know, the involvement and functional importance of flexibility in recognizing and binding histones has not been suggested before for CW domains. There are, however, relevant precedents in the literature. Functional flexibility has been reported for acetyltransferases acting on histones [47], as well as bromo-domains binding such acetylation sites [48]. Flexibility and fluctuating conformations may shed light on how ASHH2 CW differentiates between methylation states while at the same time effectively searching the histones. Conformational selection and induced fit mechanisms have been reported to be important for search and dock tasks, such as effectively scanning DNA for sequence-specific DNA methylations [49]. Being able to sample a range of similar PTM states along histones tails before settling down and activating the full enzyme, rather than locking down at the first favourable interaction, would be an attractive property for any protein acting within the complex environment of chromatin.

Although this study concerns the properties of the CW domain, it is relevant to also discuss results in the context of the function of the full-length multidomain ASHH2 protein. Conformational selection is also implicated in regulation mechanisms [35]. Although a speculation, the CW domain could act not only as a passive reader that docks the full enzyme correctly but also play a role in activation and regulation. Extrapolating from the CW structure presented here suggests that the domains of C-terminal from CW are oriented away from H3; however, the movement of the C-terminal of CW also suggests that a rearrangement takes place that could position the SET domain optimally. The β-augmentation could stiffen both CW and the H3 tail as part of this positioning. Our study, although restricted to the ASHH2 CW domain, suggests that protein flexibility, as well as conformational selection, plays an active role in the function of the ASHH2 CW domain. Future work on the structural biology of both nucleosome and chromatin remodellers would benefit from employing a theoretical framework and methodology that allow for the detection and assessment of functional disorder and flexibility.

Material and methods

Materials

The H3 tail mimicking peptides were synthetized by LifeTein (H3K4me1, ARTKme1QTARY). For NMR samples, the peptides were also synthesized with specific stable isotope labelling sites as follows: A(^{15}N,^{13}C)RTKme1QTA(^{15}N,^{13}C)RY; AR(^{15}N,^{13}C)TKme1QTARY; and ARTKme1QTAR(^{15}N,^{13}C)Y. All peptides had 95% purity as assessed by mass

spectrometry. D$_2$O, ^{15}N-enriched (99%) NH$_4$Cl and ^{13}C-enriched (99%) glucose were purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA, USA), and SVCP-Super-3-103.5 NMR tubes were acquired from Norell Inc. (Morganton, NC, USA). Unless otherwise specified, samples were buffered by the T7 solution (25 mM Tris/HCl pH 7.0, 150 mM NaCl, 1 mM TCEP). Buffer components were acquired from Sigma-Aldrich. CW constructs were prepared using ligation-independent cloning in the KpnI/SacI restriction sites of pET-49b vector (Novagen/Merck, Darmstadt, Germany), and the protein and all mutants were expressed and purified as described [24]. For more details, see Supporting Information: Cloning of CW constructs, site-directed mutagenesis, protein expression and purification. Protein and peptide concentrations for all types of samples were determined using UV-Vis spectroscopy (NanoDrop: absorption at 280 nm, extinction coefficient 19730 M^{-1}·cm^{-1} for CW constructs, 1490 M^{-1}·cm^{-1} for H3K4meX peptides).

Intrinsic tryptophan fluorescence

Affinity measurements

The approach for determining K_ds was adapted from Ref. [50,51] and is further described in Supporting Information: Intrinsic tryptophan fluorescence affinity measurements. Briefly, for each combination of CW construct and H3K4meX, a titration with constant protein concentration and variable ligand (typically 0.0–8.2 μM, up to 15.9 for low-affinity binders) concentration was performed. Stocks were prepared so that the cuvette concentrations for parallel runs always were within 2.0–2.4 μM. The intrinsic tryptophan fluorescence was monitored at wavelengths where the intensity response was greatest in each case (typically 319–322 nm). Δ-intensity values at the wavelengths where the protein-only contribution is subtracted were used as the observable, F_{PL}, when the dissociation constant K_d was determined using a nonlinear least-squares fit to this equation.

$$F_{OBS} = F_P P_o + (F_{PL} - F_P)$$
$$\times \left\{ ((K_d + P_o + L_o) - (K_d + P_o + L_o)^2 - 4P_o L_o)^{1/2} \right\}$$
(1)

F_P and F_{PL} are fluorescence of the protein and protein–ligand complex, respectively, while P_o and L_o are the total concentrations of the protein and the ligand. The shape of the curves, rather than the absolute measurement values, affects the K_ds, which was determined at least three times for each CW construct–ligand combination. Final affinity values and their errors are means of these determinations, and their standards deviations, respectively.

Thermal denaturation measurements

The ratio of the fluorescence intensities recorded at 335 and 355 nm as a function of temperature was monitored. This

ratio is a useful proxy for measuring the unfoldedness of a protein [25] and can be used to fit a 4-parameter sigmoidal curve with T_m as an output parameter. For each sample and temperature point, data were acquired from 310 to 400 nm, and temperature ranged from 4 to 90 °C (5–10 °C stepwise increases). Between each measurement, a 5-min wait was introduced for thermal equilibration. The 100-μL quartz cuvette was equipped with a lid to prevent sample evaporation. For more details, see Supporting Information: Thermal denaturation monitored by intrinsic tryptophan fluorescence.

NMR spectroscopy

Data collection

Data were collected at 25 °C on an 850 MHz Bruker Avance III HD Spectrometer fitted with a ^{1}H/^{13}C/^{15}N TCI CryoProbe and a SampleJet with temperature control for storing samples in between runs (set to 4 °C). Samples were prepared in NMR buffer consisting of 20 mM potassium phosphate, 50 mM NaCl and 1 mM DTT adjusted to pH 6.4, and all NMR experiments related to the backbone and side-chain assignment performed for this study (summarized in Table S2) were collected, processed and analysed as described previously [24]. Protein diffusion measurements were performed on protons at 25 °C using stimulated echo, bipolar gradients and 3-9-19 pulse train for solvent suppression. ^{1}H-^{15}N NOE values, and ^{15}N longitudinal ($R_1 = 1/T_1$) and transverse ($R_2 = 1/T_2$) relaxation rates were acquired using sequences in Table S2. Local backbone dynamics was determined using model-free Lipari–Szabo formalism. CPMG relaxation dispersion experiments were acquired at 600-and 850-MHz fields at 25 and 35 °C using pulse sequences, series of spin-echo pulse elements and relaxation delays described in Table S2. Analysis of the relaxation data was carried out in Bruker Dynamics Center 2.5.3 (Bruker BioSpin, Billerica, MA, USA) and NESSY [37]. For more details, see Supporting Information: Heteronuclear NOE, relaxation measurements and Model free analysis of backbone local dynamics.

Structural calculation, refinement and validation of the CW42-H3K4me1 complex

Assignment of ARTKme1QTARY in complex with CW42 was carried out using 2D-filtered ^{1}H-^{1}H TOCSY and 2D-filtered ^{1}H-^{1}H NOESY spectra. The peptide assignment was used to establish the intermolecular NOE connectivities with CW42 (BMRB ID: 27251, [24]). Intra- and intermolecular NOE cross-peaks were assigned manually using the CARA program v 1.9.1.2 [52]. The structure of the CW42-H3K4me1 complex was then calculated using CYANA v. 3.97 (L.A.Systems, Inc.,Tokyo, Japan) [53] , based on distance constraints converted from the NOESY peak lists and torsion angles obtained from the secondary chemical shifts in TALOS N [54]. Two hundred conformers were calculated using

CYANA with 15 000 simulated annealing steps. To ensure that the Zn^{2+} ion tetrahedral geometry was correctly represented, the position of the Zn^{2+} ion was restricted towards the final stage of calculation by setting the lower and upper distance limits for S^{γ}-Zn^{2+} to 2.2 and 2.4 Å and to 2.9 and 3.4 Å for C^{β}-Zn^{2+} for the residues C868, C871, C893 and C904. The twenty structure conformers of the complex with the lowest target functions were subsequently energy-minimized in implicit solvent (Generalized Born [55]). This was done using the Amber ff14SB force field [56], with modifications for the monomethylated lysine [57] and ZAFF parameters for Zn^{2+} and the Zn-coordinated amino acids [58]. The procedure consisted of 50 000 steps of steepest descent followed by 10 000 steps of conjugate gradient minimization with a 100 nm cut-off for nonbonded interactions. NOE constraints were applied as a square well potential with a force constant of 50 kcal·mol^{-1}·Å^{-2}. The refined conformations of the CW42-H3K4me1 complex were validated in RCSB validation server OneDep [59] and deposited as PDB entry 6QXZ. For more details, see Table S2.

Molecular dynamics simulations

Molecular dynamics simulations were performed on the following four structures: a representative conformation of the CW42-H3K4me1 complex of the NMR structure reported in this work (PDB ID: 6QXZ, 'NMR full'); the crystal structure complex (PDB ID: 5YVX); the uncomplexed NMR structure (PDB ID: 2L7P); and the CW42-H3K4me1 complex again, but with the E917A mutation specific to the crystal structure and with the C-terminal removed (residues G922–A928, 'NMR Cut'). The latter simulation was performed to evaluate the relevance of these differences and to be able to compare simulations with the same number of atoms. For the NMR structures, residues related to the plasmid or peptide design, and thus not native to the sequence, were removed in the N-terminal prior to simulation. Briefly, the structures were subject to 50-ns MD simulations using NAMD and the CHARMM36 force field [60–62]. In order to obtain a better sampling, a total of three simulations have been run for each system, using different initial velocities. Each system was solvated with TIP3 water molecules and neutralized with chloride and potassium ions using CHARMM, and then energy-minimized prior to three equilibration steps: water simulation with constrained structure (100 ps), simulation with protein backbone and key ligand-binding residues constrained (500 ps), and unconstrained simulation prior to main simulation (500 ps). For details of the equilibration steps, see Supporting Information: Molecular Dynamics Simulations.

Isothermal calorimetry

Isothermal calorimetry was performed for the wild-type (WT) CW42 and the mutants at a stirring rate of 300 r.p.m. at 25 °C. The protein concentrations of CW42 and the mutants were typically in the range of 50–180 μM, and the enthalpy of binding was determined by stepwise titration with 400–1800 μM histone peptide (H3K4meX, ARTKmeXQTARKY, where X denotes the number of methyl groups on the lysine, 1–3). Both the protein and the peptide were dissolved in T7 buffer, and the heat of peptide dilution into T7 buffer was subtracted from the measurement using the average of 22 successive titrations. Corrected heat flow peaks were integrated, plotted and fitted by independent modelling to determine binding parameters using the NanoAnalyze V 2.4.1 software. Experiments were performed in triplicate or more on a Nano ITC from TA Instruments.

Conflict of interest

All authors declare no conflicts of interest.

Author contributions

RA, VDM and EC designed, planned and executed screen for CW constructs. RA, VDM and ØS performed MALS experiments. MB, RA, ØS and ØØF prepared the samples. OD, ØH, JI and JU acquired NMR data. OD performed NMR structural calculations. OD, ØH and JI performed NMR dynamics experiments and analysed the data. MB and ØS performed mutagenesis. MB, ØS, ØØF and SRM designed ITC and other affinity studies. OD, NM and KT refined NMR structure. OD, NM and NR performed molecular dynamic simulations. OD, MB, NM, ØØF, SM, ØS, RA and ØH prepared figures.

OD, MB, NM, ØØF, ØS, RBA, RA and ØH wrote the manuscript. ØH, OD, RA, JU and NR supervised the work. RA and ØH designed the research and wrote grants. All authors interpreted data and read and commented on the manuscript.

References

1 Strahl BD & Allis CD (2000) The language of covalent histone modifications. *Nature* **403**, 41–45.

2 Spotswood HT & Turner BM (2002) An increasingly complex code. *J Clin Invest* **110**, 577–582.

3 Zhang Y & Reinberg D (2001) Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev* **15**, 2343–2360.

4 Schneider J & Shilatifard A (2006) Histone demethylation by hydroxylation: chemistry in action. *ACS Chem Biol* **1**, 75–81.

5 Maurer-Stroh S, Dickens NJ, Hughes-Davies L, Kouzarides T, Eisenhaber F & Ponting CP (2003) The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends Biochem Sci* **28**, 69–74.

6 Sanchez R & Zhou MM (2011) The PHD finger: a versatile epigenome reader. *Trends Biochem Sci* **36**, 364–372.

7 He FH, Umehara T, Saito K, Harada T, Watanabe S, Yabuki T, Kigawa T, Takahashi M, Kuwasako K, Tsuda K *et al.* (2010) Structural insight into the zinc finger CW domain as a histone modification reader. *Structure* **18**, 1127–1139.

8 Hoppmann V, Thorstensen T, Kristiansen PE, Veiseth SV, Rahman MA, Finne K, Aalen RB & Aasland R (2011) The CW domain, a new histone recognition module in chromatin proteins. *EMBO J* **30**, 1939–1952.

9 Perry J & Zhao Y (2003) The CW domain, a structural module shared amongst vertebrates, vertebrate-infecting parasites and higher plants. *Trends Biochem Sci* **28**, 576–580.

10 Berg A, Meza TJ, Mahic M, Thorstensen T, Kristiansen K & Aalen RB (2003) Ten members of the Arabidopsis gene family encoding methyl-CpG-binding domain proteins are transcriptionally active and at least one, AtMBD11, is crucial for normal development. *Nucleic Acids Res* **31**, 5291–5304.

11 Liu Y, Tempel W, Zhang Q, Liang X, Loppnau P, Qin S & Min J (2016) Family-wide characterization of histone binding abilities of human CW domain-containing proteins. *J Biol Chem* **291**, 9000–9013.

12 Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, Stroud H, Husmann D, Zhan Y, Lajoie BR, McCord RP *et al.* (2012) MORC family ATPases required for

heterochromatin condensation and gene silencing. *Science* **336**, 1448–1451.

13 Fang R, Barbera AJ, Xu Y, Rutenberg M, Leonor T, Bi Q, Lan F, Mei P, Yuan GC, Lian C *et al.* (2010) Human LSD2/KDM1b/AOF1 regulates gene transcription by modulating intragenic H3K4me2 methylation. *Mol Cell* **39**, 222–233.

14 Liu Y & Huang Y (2018) Uncovering the mechanistic basis for specific recognition of monomethylated H3K4 by the CW domain of Arabidopsis histone methyltransferase SDG8. *J Biol Chem* **293**, 6470-6481.

15 Ko JH, Mitina I, Tamada Y, Hyun Y, Choi Y, Amasino RM, Noh B & Noh YS (2010) Growth habit determination by the balance of histone methylation activities in Arabidopsis. *EMBO J* **29**, 3208–3215.

16 Thorstensen T, Grini PE & Aalen RB (2011) SET domain proteins in plant development. *Biochim Biophys Acta* **1809**, 407–420.

17 Zhao Z, Yu Y, Meyer D, Wu C & Shen WH (2005) Prevention of early flowering by expression of FLOWERING LOCUS C requires methylation of histone H3 K36. *Nat Cell Biol* **7**, 1256–1260.

18 Xu L, Zhao Z, Dong A, Soubigou-Taconnat L, Renou JP, Steinmetz A & Shen WH (2008) Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in *Arabidopsis thaliana*. *Mol Cell Biol* **28**, 1348–1360.

19 Grini PE, Thorstensen T, Alm V, Vizcay-Barrena G, Windju SS, Jørstad TS, Wilson ZA & Aalen RB (2009) The ASH1 HOMOLOG 2 (ASHH2) histone H3 methyltransferase is required for ovule and anther development in *Arabidopsis*. *PLoS One* **4**, e7817.

20 Li Y, Mukherjee I, Thum KE, Tanurdzic M, Katari MS, Obertello M, Edwards MB, McCombie WR, Martienssen RA & Coruzzi GM (2015) The histone methyltransferase SDG8 mediates the epigenetic modification of light and carbon responsive genes in plants. *Genome Biol* **16**, 79.

21 Andrews FH, Tong Q, Sullivan KD, Cornett EM, Zhang Y, Ali M, Ahn J, Pandey A, Guo AH, Strahl BD *et al.* (2016) Multivalent chromatin engagement and inter-domain crosstalk regulate MORC3 ATPase. *Cell Rep* **16**, 3195–3207.

22 Sandhu KS (2009) Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. *J Mol Recognit* **22**, 1–8.

23 Lazar T, Schad E, Szabo B, Horvath T, Meszaros A, Tompa P & Tantos A (2016) Intrinsic protein disorder in histone lysine methylation. *Biol Direct* **11**, 30.

24 Dobrovolska O, Bril'kov M, Odegard-Fougner O, Aasland R & Halskau O (2018) (1)H, (13)C, and (15)N resonance assignments of CW domain of the N-methyltransferase ASHH2 free and bound to the

mono-, di- and tri-methylated histone H3 tail peptides. *Biomol NMR Assign* **12**, 215–220.

25 Martensson LG, Jonasson P, Freskgard PO, Svensson M, Carlsson U & Jonsson BH (1995) Contribution of individual tryptophan residues to the fluorescence spectrum of native and denatured forms of human carbonic anhydrase II. *Biochemistry* **34**, 1011–1021.

26 Macchioni A, Ciancaleoni G, Zuccaccia C & Zuccaccia D (2008) Determining accurate molecular sizes in solution through NMR diffusion spectroscopy. *Chem Soc Rev* **37**, 479–489.

27 Zaric SD (2003) Metal ligand aromatic cation-pi interactions. *Eur J Inorg Chem* **5**, 20–31.

28 Zhang Y, Klein BJ, Cox KL, Bertulat B, Tencer AH, Holden MR, Wright GM, Black J, Cardoso MC, Poirier MG *et al.* (2019) Mechanism for autoinhibition and activation of the MORC3 ATPase. *Proc Natl Acad Sci USA* **116**, 6111–6119.

29 Amaral M, Kokh DB, Bomke J, Wegener A, Buchstaller HP, Eggenweiler HM, Matias P, Sirrenberg C, Wade RC & Frech M (2017) Protein conformational flexibility modulates kinetics and thermodynamics of drug binding, *Nat Commun* **8**, 2276–2289.

30 Yang LQ, Sang P, Tao Y, Fu YX, Zhang KQ, Xie YH & Liu SQ (2014) Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *J Biomol Struct Dyn* **32**, 372–393.

31 Kay LE, Torchia DA & Bax A (1989) Backbone dynamics of proteins as studied by N-15 inverse detected heteronuclear nmr-spectroscopy - application to *Staphylococcal Nuclease. Biochemistry* **28**, 8972–8979.

32 Alderson TR & Markley JL (2013) Biophysical characterization of alpha-synuclein and its controversial structure. *Intrinsically Disord Proteins* **1**, 18–39.

33 Mazzei L, Dobrovolska O, Musiani F, Zambelli B & Ciurli S (2015) On the interaction of Helicobacter pylori NikR, a Ni(II)-responsive transcription factor, with the urease operator: in solution and *in silico* studies. *J Biol Inorg Chem* **20**, 1021–1037.

34 Boehr DD, McElheny D, Dyson HJ & Wright PE (2010) Millisecond timescale fluctuations in dihydrofolate reductase are exquisitely sensitive to the bound ligands. *Proc Natl Acad Sci USA* **107**, 1373–1378.

35 Csermely P, Palotai R & Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci* **35**, 539–546.

36 Korzhnev DM, Salvatella X, Vendruscolo M, Di Nardo AA, Davidson AR, Dobson CM & Kay LE (2004) Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature* **430**, 586–590.

37 Bieri M & Gooley PR (2011) Automated NMR relaxation dispersion data analysis using NESSY. *BMC Bioinformatics* **12**, 421.

38 Farber PJ & Mittermaier A (2015) Relaxation dispersion NMR spectroscopy for the study of protein allostery. *Biophys Rev* **7**, 191–200.

39 d'Auvergne EJ & Gooley PR (2003) The use of model selection in the model-free analysis of protein dynamics. *J Biomol NMR* **25**, 25–39.

40 Chakrabarti KS, Agafonov RV, Pontiggia F, Otten R, Higgins MK, Schertler GFX, Oprian DD & Kern D (2016) Conformational selection in a protein-protein interaction revealed by dynamic pathway analysis. *Cell Rep* **14**, 32–42.

41 Huang F & Nau WM (2003) A conformational flexibility scale for amino acids in peptides. *Angew Chem Int Ed Engl* **42**, 2269–2272.

42 Karimi-Ashtiyani R & Houben A (2013) *In vitro* phosphorylation of histone H3 at threonine 3 by Arabidopsis Haspin is strongly influenced by posttranslational modifications of adjacent amino acids. *Mol Plant* **6**, 574–576.

43 Devarakonda S, Gupta K, Chalmers MJ, Hunt JF, Griffin PR, Van Duyne GD & Spiegelman BM (2011) Disorder-to-order transition underlies the structural basis for the assembly of a transcriptionally active PGC-1alpha/ERRgamma complex. *Proc Natl Acad Sci USA* **108**, 18678–18683.

44 Halskau O Jr, Perez-Jimenez R, Ibarra-Molero B, Underhaug J, Munoz V, Martinez A & Sanchez-Ruiz JM (2008) Large-scale modulation of thermodynamic protein folding barriers linked to electrostatics. *Proc Natl Acad Sci USA* **105**, 8625–8630.

45 Okazaki KI & Takada S (2008) Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc Natl Acad Sci USA* **105**, 11182–11187.

46 Remaut H & Waksman G (2006) Protein-protein interaction through beta-strand addition. *Trends Biochem Sci* **31**, 436–444.

47 Langini C, Caflisch A & Vitalis A (2017) The ATAD2 bromodomain binds different acetylation marks on the histone H4 in similar fuzzy complexes. *J Biol Chem* **292**, 19121.

48 Setiaputra D, Ross JD, Lu S, Cheng DT, Dong MQ & Yip CK (2015) Conformational flexibility and subunit arrangement of the modular yeast Spt-Ada-Gcn5 acetyltransferase complex. *J Biol Chem* **290**, 10057–10070.

49 Estabrook RA & Reich N (2006) Observing an induced-fit mechanism during sequence-specific DNA methylation. *J Biol Chem* **281**, 37205–37214.

50 Martin SR & Schilstra MJ (2008) Circular dichroism and its application to the study of biomolecules. *Method Cell Biol* **84**, 263–293.

51 Martin SR, Schilstra MJ & Siligardi G (2011) Biophysical Approaches Determining Ligand Binding to Biomolecular Targets: Detection, Measurement and Modelling. Royal Society of Chemistry, Cambridge.

52 Keller RLJ (2005) Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment.

53 Guntert P (2004) Automated NMR structure calculation with CYANA. *Methods Mol Biol* **278**, 353–378.

54 Shen Y & Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR* **56**, 227–241.

55 Hawkins GD, Cramer CJ & Truhlar DG (1996) Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J Phys Chem* **100**, 19824–19839.

56 Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE & Simmerling C (2015) ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713.

57 Papamokos GV, Tziatzos G, Papageorgiou DG, Georgatos SD, Politou AS & Kaxiras E (2012) Structural role of RKS motifs in chromatin interactions: a molecular dynamics study of HP1 bound to a variably modified histone tail. *Biophys J* **102**, 1926–1933.

58 Peters MB, Yang Y, Wang B, Fusti-Molnar L, Weaver MN & Merz KM (2010) Structural survey of zinc-containing proteins and development of the Zinc AMBER Force Field (ZAFF). *J Chem Theory Comput* **6**, 2935–2947.

59 Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, Sen S, Gutmanas A, Armstrong DR, Berrisford JM *et al.* (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* **25**, 536–545.

60 Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M & MacKerell AD (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J Chem Theory Comput* **8**, 3257–3273.

61 Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K & Schulten K (1999) NAMD2: greater scalability for parallel molecular dynamics. *J Comput Phys* **151**, 283–312.

62 Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S *et al.* (2009) CHARMM: the biomolecular simulation program. *J Comput Chem* **30**, 1545–1614.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Selected distances between the centres of mass of amino acids (aa) of the binding site and K4me1.

Table S2. List of NMR experiments used in this study.

Table S3. Primers used for Ligation Independent Cloning and site-directed mutagenesis of CW42.

Table S4. Model choice and output summary from NESSY for the unbound state of CW42.

Table S5. Model choice and output summary from NESSY for the bound state of CW42.

Fig. S1. K_d determinations of CWs, CW42, CW37 and CW33 for H3K4me2 and H3K4me3.

Fig. S2. Loss of specific binding caused by mutations in the H3-mimicking peptide and the postα1 loop.

Fig. S3. Average RMSF and standard deviation of the CW constructs assessed by MD.

The *Arabidopsis* (ASHH2) CW domain binds monomethylated K4 of the histone H3 tail through conformational selection

Olena Dobrovolska, Maxim Brilkov, Noelly Madeleine, Øyvind Ødegård-Fougner, Øyvind Strømland, Stephen R. Martin, Valeria De Marco, Evangelos Christodoulou, Knut Teigen, Johan Isaksson, Jarl Underhaug, Nathalie Reuter, Reidunn B. Aalen, Rein Aasland and Øyvind Halskau

Supplementary Information for:

The Arabidopsis (ASHH2) CW domain binds monomethylated K4 of the histone H3 tail through conformational selection

Please note: the supplementary table numbering follows their order of appearance in the main text. Hence, appearance in this document does not start with Table S1.

Cloning of CW constructs, site-directed mutagenesis, protein expression and purification

The CW42-construct fused to N-terminal GST-tag (**MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYID GDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDF LSKLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKK RIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSD**GSTSGGGGSNNNPPTPTPSSGSGH HHHHHSAA*LEVLFQG*PG where the sequence in bold is the GST-tag and in italics is the HRV 3C cleavage site) and subsequently cloned into pSXG vector was subjected to site directed mutagenesis with PCR. For forward and reverse primers used in this process, see Table S3. After the PCR reaction and DpnI enzyme treatment, linearized DNA was ligated in a reaction mixture containing the PCR product, NEB4 buffer, T4 ligase, PNK T4 Kinase, 1 mM ATP and 10 mM DTT (room temperature, 20 minutes). Chemicompetent One Shot TOP10 cells were transformed with ligated plasmids. The mini-prep plasmids obtained were verified by sequencing. The mutant CW42-constructs were then expressed as GST fusion proteins in BL21-CodonPlus (DE3)-RIL chemicompetent cells.

Intrinsic tryptophan fluorescence affinity measurements

Each experiment was performed as follows: 100 µL of CW33, CW37, CW42 or CWs dissolved in T7 buffer was placed in a cuvette (protein concentration 2.2 µM) and then 2 µL aliquots of histone peptide in T7 buffer (~40µM histone peptide, and also 2.2 µM CW) were added. This ensured that the CW concentration was constant at 2.2 µM throughout all titrations. Typical titration series consisted of 9-12 concentration points in the range of 0.0 - 6.2 µM for the H3K4me1 peptide titrated into all CW constructs. Similar concentration values were used for H3K4meX peptides. This usually gave final protein:peptide concentrations of up to 3.3, enough to reach saturation for the high-affinity instances. For the weaker binding proteins, larger aliquots were added. A total of 13 unique titration series were performed, one for each CW in combination with H3K4m1, H3K4me2 and H3K4me3, plus one where CWs was titrated with H3K4me0. The CWs to H3K4me0 titration indicated that no binding took place, and so H3K4me0 binding was not investigated further for the other constructs. Each series was prepared from stocks and repeated up to three times, to ensure repeatability. Intrinsic tryptophan fluorescence data was acquired for each unique series and its repeats on a Jasco FP-8500 Series Fluorimeter at 25 °C. Tryptophans were excited at 290 nm, and emission data was recorded in the interval 300-450 nm. Representative blanks consisting of histone peptide only titrated into T7 buffer were also acquired and subtracted prior to further analysis. Intensities, F_{OBS}, at the wavelength giving the largest fluorescence change (typically in the interval 319-322nm) were recorded as a function of histone peptide concentration. Then, the dissociation constant K_d was determined using a non-linear least-squares fit to this equation.

$$F_{OBS} = F_P P_o + (F_{PL} - F_P) \times \left\{((K_d + P_o + L_o) - (K_d + P_o + L_o)^2 - 4P_o L_o)^{1/2}\right\} \text{ Eq. 1}$$

Here, F_P and F_{PL} are fluorescence of the protein and protein-ligand complex, respectively, while P_o and L_o are the total concentrations of the protein and the ligand. The fitting procedure also includes the option for fixing individual parameters at known values and also a factor that can be used to assess the effects of errors in the ligand concentration. Data was fit for each individual parallel, rather than averages of parallels; K_ds provided in the Fig. 1D in the main document are averages of at least three such fits, and the errors given are one standard deviation. Examples of individual data fits are shown in Fig. 1C in the main document, and in <u>Fig. S1</u> in this document.

Thermal denaturation monitored by intrinsic tryptophan fluorescence

The three tryptophan residues in the CW domain were excited with a 295 ± 2.5 nm light beam in an LS 50B Fluorescence Spectrometer, and using a 100 µL quartz cuvette to hold the sample. Fluorescence was observed between 310 nm and 450 nm with a scanning rate of 100 nm/min, and each measurement was an average of 5 scans. Experiments were performed in triplicates with and without 5 times molar excess of H3K4me1 peptide. Before each experiment, the samples were spun down at 17000 x g for 30 minutes to ensure that any precipitates were removed, and protein concentration (typically 5-15 µM) was measured after this step. Each sample was subjected to a temperature gradient ranging from 4 °C to 90 °C, using a 5 °C – 10 °C stepwise increase in temperature. Between each temperature increase, 5 minute waiting intervals were introduced for the sample and holder to reach thermal equilibrium. The cuvette was equipped with a lid, to prevent evaporation from the sample. Fluorescence data measured from the blank (T7 buffer at 4 °C) was subtracted, and the ratio of the fluorescence intensity at 335 nm (dominated by tryptophan fluorescence of a folded domain) and 355 nm (representative of a tryptophan exposed to the solvent) was plotted against temperature. This parameter, I335nm/I355nm, is a useful proxy for the overall fold of the protein (1). A global curve fit using a 4-parameter sigmoidal curve taking all parallels into account was then used to find the T_m:

$$f(x) = y_0 + \frac{y_f}{1+10^{\left(-\frac{Tm-x}{b}\right)}} \qquad \text{Eq. 2}$$

Here, y_0 is the slope of the pre-transition state, y_f is the slope of the post-transition state, T_m is the inflection point in the curve, and b is the slope at the inflection point.

Determination of molecular weight by multi-angle static light scattering

Analytical size exclusion chromatography was performed by injecting 100 µl of either CW42 at 450 µM or CW42 at 450 µM with a 2-fold molar excess of H3K4me1 peptide onto a Superdex 75 10/300 Increase column (GE Healthcare) connected to an Äkta Purifier (GE Healthcare). The running buffer was T7 and the flow rate was 0.4 ml/ml. The Äkta system was connected to a mini-DAWN TREOS multi-angle static light scattering (MALS) detector (Wyatt) and an Optilab rEX differential refractometer (Wyatt) for absolute molecular weight determination. The molecular weights were determined using the ASTRA software (Wyatt).

Protein diffusion measurements:

Stimulated echo ^{1}H NMR diffusion measurements using bipolar gradients and 3-9-19 water suppression were performed at 25 °C. Briefly, 64 gradients arranged in a linear gradient ranging from 5% to 95% of the maximum strength were recorded as pseudo-2D NMR data. The number of scans were set to 8, and the gradient pulse duration and diffusion time were set to 1000 µs and 300 ms, respectively. Each 1D slice was processed using 2k data points, a squared cosine window function for FID apodization, and qfil

baseline modification. The resulting traces were then phased and subjected to a 5-term polynomial baseline adjustment prior to processing using the Bruker AU-program dosy2d. The program finds the diffusion constant, D, by fitting the decaying signals of each peak (picked using automated peak picking) to:

$$I = I_0 e^{-(\gamma \delta G)^2 \left(\Delta_T - \frac{\delta}{3} - \frac{\tau}{2}\right) D} \qquad \text{Eq. 3}$$

Here, I and I_0 are the volumes in the presence and absence of gradients, γ is the gyromagnetic ratio of ^{1}H, G is the gradient strength, δ is the duration of the bipolar gradients and τ is the gradient recovery time. Δ_T is a constant described as the sum of the diffusion time Δ, δ and 2τ. Only selected peaks from the non-exchanging, upfield region were used to estimate the mean diffusion constant, as either buffer components or the ligand do not influence this spectral region. The Stokes-Einstein equation, Eq. 4, was then used to calculate the hydrodynamic radius of *holo* and *apo* CW.

$$D = \frac{k_B T}{6 \pi \eta R} \qquad \text{Eq. 4}$$

In Eq. 4, D is the diffusion constant of a rigid sphere in solution, k_B the Boltzmann constant, η the viscosity of the solution and R is the radius of the assumed sphere.

Assignment and structure calculation of the CW42-H3K4me1 complex:

The ARTKme1QTARY peptide in complex with CW42 was assigned using 2D filtered ^{1}H-^{1}H TOCSY (mixing times of 40 ms, 80 ms, 120 ms) and 2D filtered ^{1}H-^{1}H NOESY spectra (50 ms, 80 ms, and 120 ms mixing time) (Table S2). This assignment was then used to establish the intermolecular NOE connectivities with CW42 (BMRB ID: 27251, [1]). Intra- and intermolecular NOE cross-peaks, based on data collected on labelled CW42 samples in complex with unlabelled peptide, were assigned manually using CARA program v 1.9.1.2 [2] using the experiments detailed in Table S2. Additionally, a set of 2D ^{1}H-^{15}N/^{13}C - HSQC's and 3D ^{15}N/^{13}C - edited NOESY-HSQC spectra were collected for the samples containing unlabelled CW42 bound to H3K4me1 with selectively ^{13}C- and ^{15}N-labelled residues (A1, A7, R2, and R8, see Table S2). Intensities of the NOE cross-peaks were integrated using the model-based linear equation system method [2] and exported as NOESY peak lists into the torsion angle dynamics program CYANA v. 3.97 [3]. Torsion angles were obtained from the secondary chemical shifts using TALOS N [4].The structure of the CW42-H3K4me1 complex was then calculated based on the distance constraints converted from the NOESY peak lists and the TALOS N derived torsion angles. The monomethylated lysine residue was defined *in silico* using the UCSF Chimera package, based on a modification of the standard lysine residue from CYANA library [5]. 200 structures were calculated using CYANA with 15 000 simulated annealing steps using the distance and dihedral angle restraints.

Heteronuclear NOE, relaxation measurements and Model free analysis of backbone local dynamics

In order to perform model-free NMR analysis, ^{1}H-^{15}N NOE values, ^{15}N Longitudinal ($R_1 = 1/T_1$) and transverse ($R_2 = 1/T_2$) relaxation rates were acquired using sequences in Table S2. For determination of heteronuclear NOEs, two ^{1}H-^{15}N HSQC data sets were recorded. A recycling delay of 10 s was used between transients. T_1 measurements were based on inversion-recovery type experiments recorded using 11 different delays: 20, 60, 80, 100, 200, 400, 600, 800 (performed twice), 1000, 1200, and 1400 ms. T_2 measurements were carried out using a Carr-Purcell-Meiboom-Gill (CPMG, [6]) spin-echo pulse sequences acquired with 12 different delays: 16, 30, 60, 95, 125, 160, 190, 220, 250, 345, 440, and 500 ms. Recycle delays of 10 s were used in both experiments.

^{1}H-^{15}N steady-state NOE, longitudinal relaxation rate R_1 ($1/T_1$), and transverse relaxation rate R_2 ($1/T_2$) were determined from cross-peak intensities/integrals in ^{1}H-^{15}N HSQC spectra. The heteronuclear NOE values were estimated as the ratio of the steady-state intensities measured in the presence and absence of saturation of the proton magnetization. R_1 and R_2 parameters were obtained by fitting the peak intensities acquired at different relaxation delays to a two-parameter (I_0, and T_1 or T_2, respectively) exponential decay function, using a nonlinear Levenberg-Marquard algorithm. The errors of the fitted parameters were estimated from the inverse of the weighted curvature matrix, using a confidence level of 95%. In all cases, the uncertainty on the intensities was estimated from the standard deviation of the noise in each spectrum. Analysis of the relaxation data to determine local backbone dynamics was performed using Model-free Lipari-Szabo formalism [7], carried out in Bruker Dynamics Center 2.5.3 (Bruker BioSpin).

CPMG relaxation dispersion spectra were acquired at 600 MHz and 850 MHz, as a series of pseudo-3D experiments using constant CPMG period (T_{CPMG}) of 0.04 s and 0.08 s, respectively. The following CPMG field strengths, υ_{cp}, were used: 25, 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 800, 1000 Hz, with repeat experiments performed at 50, 100, 150, and 800 Hz.

Molecular Dynamics Simulations

MD simulations of the CW42-H3K4me1 complex: A representative, NMR-derived conformation of the CW42-H3K4me1 complex and the crystal structure complex containing the E917A mutation (PDB ID: 5YVX) were subjected to MD simulations using NAMD [8] and the CHARMM36 force field [9]. Residues G849 to T861 in CW42 were removed from the NMR structures since they derive from the plasmid and not from the ASHH2 sequence [1]. Likewise, the N-terminal Tyr-residue, present in the laboratory experiments to facilitate concentration measurements in the H3K4me1 peptide, was removed. A simulation of the wild-type CW42-H3K4me1 complex without the C-terminal coil (residues G922 to A928) was also performed to evaluate the relevance of this part that was missing in the X-Ray structure. Finally, to compare the complex's flexibility, we performed an MD simulation of the NMR structure containing the E917A mutation and without the C-terminal coil. In this manner, the X-Ray and NMR structures could be compared directly with respect to flexibility, as both simulations contained the same number of atoms. The mutant was prepared using Chimera from the wild-type CW42-H3K4me1 structure without the C-terminal coil. For a graphical representation of these setups and how they related to X-ray and NMR protein structure, please see Fig. 8 in the main document.

Each system was solvated with TIP3 water molecules and neutralized with chloride and potassium ions using CHARMM [10]. All systems were then energy minimized prior to three equilibration steps. The first equilibration step consisted of a 100 ps water simulation, performed while keeping all complex coordinates fixed. Then, the backbone of the whole complex and the side chains of the residues W865, W874, I915, L919, I921, Q923, all in contact with the residue K4me1 of the H3K4me1 peptide in our structures, were constrained in a 500 ps equilibration. Since the X-Ray E917A mutant (PDB ID: 5YVX) contains only residues from E862 to I921, we did not apply the constraint on the Q923 side chain in the second equilibration step in this case. Finally, the systems were equilibrated for 500 ps in the NPT ensemble without any constraints prior to a 50 ns NPT simulation for the production phase. After equilibration, each simulation was done at 300 K with a time step of 1 fs. In all cases, we used Langevin dynamics with a temperature damping coefficient of 1 and the Langevin piston method with a pressure of 1 atm (oscillation period: 200 fs, damping timescale: 100 fs) [11]. Coordinates of systems were written every 2 ps in the trajectory files. The last 46 ns of each simulation were then analyzed using VMD [12]. The Root Mean Square Deviation (RMSD) and the Root Mean Square Fluctuation (RMSF, using the C_α atoms) were calculated throughout

the simulations to evaluate the complex stability and its local flexibility, respectively. The RMSD evolution was calculated using the initial structure as a reference, taking into account only the backbone atoms.

MD simulations of the CW42 *apo* form: The first conformation of the *apo* CW42 NMR structure (PDB ID: 2L7P) was used for MD simulations. To allow comparisons between the CW42 *apo* form simulations and those obtained for the CW42-H3K4me1 complex, we removed the first 14 residues and the last 9 residues in the CW42 *apo* form. The protocol used for the MD simulations was the same as described above except for the second step of equilibration. In this step, only the backbone was constrained in the CW42 *apo* forms.

Table S1

Selected distances between the centres of mass of amino acids (aa) of the binding site and K4me1. The table shows distances in the representative NMR structure before (initial structure) and during MD simulation (mean distance ± sd).

Amino Acid	Distance in initial structure (Å)	Distance in MD simulation (Å)
I915 – L919	5.80	5.59 ± 0.21
I915 – W865	6.56	6.74 ± 0.32
I915 – W874	7.66	6.29 ± 0.42
L919 – W865	9.04	8.82 ± 0.82
L919 – W874	11.4	10.28 ± 0.87
W865 – W874	5.87	6.46 ± 0.27

Table S2

List of NMR experiments used in this study.

Type of NMR experiment	Pulse sequence (Bruker designation)	Spectral width in each dimension (SW), ppm	Number of complex points in each dimension (TD)	Number of scans per experiment (NS)	Sample
2D ^{1}H-^{1}H filtered TOCSY	dipsi2gpphwgxf	16 x 14	4096 x 100	32	^{15}N-CW42-ARTKme1QTARY
2D ^{1}H-^{1}H filtered NOESY	noesygpphwgxf.2	16 x 14	4096 x 100	32	
2D ^{1}H-^{15}N HSQC	hsqcetf3gpsi	16 x 35	2048 x 128	32	CW42-A$^{(15N,13C)}$RTKme1QTA$^{(15N,13C)}$RY; CW42-AR$^{(15N, 13C)}$TKme1QTARY; CW42-ARTKme1QTAR$^{(15N, 13C)}$Y
2D ^{1}H-^{13}C - HSQC	hsqcetgpsi	13 x 80	1024 x 128	2	
3D ^{15}N-NOESYHSQC	noesyhsqcf3gpwg3d	14 x 35 x 14	2048 x 40 x 118	32	
3D ^{13}C-NOESYHSQC	noesyhsqcetgp3d	14 x 80 x 14	2048 x 64 x 128	16	
3D ^{13}C-NOESYHSQC	noesyhsqcetgp3d	14 x 80 x 14	2048 x 64 x 128	16	^{15}N, ^{13}C-CW42-ARTKme1QTARY
3D ^{15}N-NOESYHSQC	noesyhsqcf3gpwg3d	14 x 35 x 14	2048 x 40 x 128	16	
2D ^{1}H-^{13}C – aromatic HSQC	hsqcetgpsi	13 x 100	1024 x 256	2	
3D ^{13}C- aromatic NOESYHSQC*	noesyhsqcetgp3d	11 x 50 x 12	2048 x 64 x 176	8	
3D ^{15}N-NOESYHSQC filter-edited	noesyhsqcf3gpwgx13d	14 x 35 x 14	2048 x 64 x 128	32	
3D ^{13}C-NOESYHSQC filter-edited	noesyhsqcgpwgx13d	15 x 80 x 12	2048 x 32 x 64	24	
3D ^{13}C- aromatic NOESYHSQC filter-edited	noesyhsqcgpwgx13d	15 x 40 x 12	2048 x 32 x 64	24	
2D ^{1}H-^{15}N HSQC (for ^{1}H-^{15}N NOE)	hsqcnoef3gpsi	16 x 35	2048 x 1024	8	^{15}N-CW42-ARTKme1QTARY
2D ^{1}H-^{15}N HSQC (for R$_1$)	hsqct1etf3gpsi3d	16 x 35	2048 x 256	8	
2D ^{1}H-^{15}N HSQC (for R$_2$)	hsqct2etf3gpsi3d	16 x 35	2048 x 256	8	
2D ^{1}H-^{15}N HSQC (for CPMG RD)	hsqcrexetf3gpsitc3d.4	16 x 35	2048 x 128	8	^{15}N-CW42, ^{15}N-CW42-ARTKme1QTARY

*Acquired at 600 MHz; all other experiments acquired at 850 MHz.

Table S3

Primers used for Ligation Independent Cloning and site-directed mutagenesis of CW42. FW = forward, RV = reverse. Primers for making the CWs constructs are given in Hoppmann *et al*.

	Forward	Reverse
CW33 (residues 854-923)	5'-CAGGGACCCGGTACTATTGAAGATAGCTATTCC	5'-GGCACCAGAGCGTTACTGTCCTATGCCCAACTCTTC
CW37 (residues 858-923)	5'-CAGGGACCCGGTAGCTATTCCACAGAGAGT	5'-GGCACCAGAGCGTTACTGTCCTATGCCCAACTCTTC
CW42 (residues 861-928)	5'-CAGGGACCCGGTACAGAGAGTGCCTGGGTT	5'-GGCACCAGAGCGTTATGCATCTGCTTCATCCTG
CW42_D886A	5'-CAATTGcCGAGAGCTCTAGATGGATCTG	5'-ATCCTACAACAGAAGCAGGTATTCG
CW42_S907P	5'-CAAAAcCTCAAGAGATGTCAAATGAAGAAATTAAG	5'-GcAATCCGCAAATCTTTTATCTGAG
CW42_S907G	5'-CAAAAggTCAAGAGATGTCAAATGAAGAAATTAATGAAGAG	5'-AGcAATCCGCAAATCTTTTATCTGAG
CW42_Q908E	5'-ATCTgAAGAGATGTCAAATGAAGAAATTAATGAAG	5'-TTTGAGCAATCCGCAAATCTTTTATC
CW42_Q923A	5'-GGGCATAGGAgcGGATGAAGCAG	5'-AACTCTTCATTAATTTCTTCATTTG

Table S4

Model choice and output summary from NESSY for the unbound state of CW42. The following resulting parameters were extracted: R_2 – effective transverse relaxation rate at infinite υ_{CPMG}; **Kex** – chemical/conformational exchange constant; **Rex** – exchange contribution to transverse relaxation; **pb** – population of the second state (where pa and pb are the populations of a two-state models with the pa is a major conformation, pa+pb=1); **dw** – chemical shift differences between states; **dG and dG*** – free energy extracted for the corresponding models; **Chi2** – a measure of goodness of fit for the selected model.

Res	Model	R_2 [1/s]	kex [1/s]	Rex [rad/s]	pb	dw [ppm]	dG at 298 K [kj/mol]	dG* at 298 K [kj/mol]	Chi2
861	2	7.14±3.56	1136.28±196.50	2.72±0.18					16.19
862	2	7.08±4.05	1095.87±138.49	4.07±0.21					23.24
863	2	6.83±4.62	2244.75±347.90	3.53±0.17					30.32
864	2	6.17±3.76	3854.45±1242.90	2.98±0.60					23.45
865	1	6.48±5.63						45.36	
866	1	7.23±3.96						35.73	
867	2	6.95±4.82	500±58.17	3.84±0.24					16.97
868	1	7.55±4.71						40.57	
869	1	6.36±4.58						62.05	
870	2	5±3.35	5703.38±782.76	4.38±0.42					24.63
871	1	6.92±3.57						21.04	
872	2	7.39±4.66	738.62±120.21	2.80±0.23					19.50
873	1	8.04±3.65						133.40	
874	2	6.63±3.51	2209.00±304.64	3.32±0.15					50.47
875	1	7.15±4.13						33.67	
876	1	8.74±4.78						52.44	
877	2	7.59±4.65	1492.79±282.44	3.03±0.20					34.88
878									
879	1	7.70±5.07						48.24	
880	1	9.26±3.17						25.43	
881	2	6.90±4.58	4943.76±2115.95	3.68±1.20					15.61
882	2	6.06±3.85	888.81±73.82	2.79±0.11					14.08
883	2	7.43±7.10	2185.81±445.06	3.47±0.26					31.86
884	1	7.36±3.34						13.01	
885	2	5.84±3.30	635.14±73.61	3.18±0.21					18.26
886	3	7.88±6.80	728.25±412.04	1.89±0.33	0.005±0.122	0.124±0.068	12.75	69.43	81.27
887	1	11.88±14.12						86.12	
888	1	6.86±4.09						35.54	
889	7	6.03±3.03	681.80±140.99	4.48±0.45	0.049±0.037	0.051±0.005	7.30	64.24	45.71
890	1	7.32±5.99						51.48	
891	2	6.08±3.84	2757.21±1428.67	2.09±0.49					28.25
892	1	6.29±5.16						16.90	
893	1	6.70±4.43						38.61	

894	1	6.90±4.54						27.17	
895	1	7.40±4.69						18.16	
896	2	6.54±4.84	531.40±99.19	2.66±0.23					13.67
897	1	8.43±4.49						31.52	
898	2	5.14±4.38	8062.12±1552.57	8.63±1.41					46.92
899	1	6.85±6.23						62.04	
900	1	7.76±4.34						17.16	
901	1	6.98±4.36						31.71	
902	1	6.79±3.91						15.51	
903	1	7.47±4.70						21.67	
904	1	7.35±4.87						28.27	
905	1	8.37±5.48						38.95	
906	1	8.35±3.97						31.51	
907	3	16.29±4.29	10.54±5.35	2.21±1.12	0.299±3.08E-08	0.814±0.248	2.09	70.13	386.74
908	2	6.67±5.03	808.58±54.47	2.63±0.09					17.59
909	1	8.80±4.68						86.94	
910	3	9.74±16.58	18.40±71.91	3.86±0.63	0.299±0.022	1.224±0.106	2.09	68.75	140.54
911	1	7.14±4.50						20.61	
912	1	9.34±3.26						28.01	
913	1	7.09±5.39						52.90	
914	1	7.26±4.13						39.55	
915	1	5.97±5.11						55.32	
916	1	6.80±5.07						33.10	
917	1	7.68±4.93						35.47	
918	1	6.12±4.75						29.13	
919	2	8.05±7.05	1712.08±341.75	2.85±0.19					15.57
920	2	4.74±4.24	910.27±95.11	3.13±0.14					15.91
921	1	5.34±4.15						28.71	
922	1	5.63±2.14						15.85	
923	1	4.12±2.47						37.96	
924	1	3.98±1.61						55.88	
925	1	2.81±1.73						62.69	
926	1	2.38±1.31						127.19	
927	1	2.59±0.85						104.97	
928	1	0.08±0.89						204.51	

Table S5

Model choice and output summary from NESSY for the bound state of CW42. Model choice and output summary from NESSY for the unbound state of CW42. The following resulting parameters were extracted: R_2 – effective transverse relaxation rate at infinite υ_{CPMG}; **Kex** – chemical/ conformational exchange constant; **Rex** – exchange contribution to transverse relaxation; **pb** – population of the second state (where pa and pb are the populations of a two-state models with the pa is a major conformation, pa+pb=1); **dw** – chemical shift differences between states; **dG and dG*** – free energy extracted for the corresponding models; **Chi2** – a measure of goodness of fit for the selected model.

Res	Model	R₂ [1/s]	kex [1/s]	Rex [rad/s]	pb	dw [ppm]	dG at 298 K [kj/mol]	dG* at 298 K [kj/mol]	Chi2
861	1	11.88±2.55							37.96
862	1	11.97±2.36							29.93
863	2	15.54±3.89	500.0±23.21	1.37±0.03					12.01
864	3	15.85±4.40	231.22±52.82	1.98±0.24	0.299±0.12	0.01±0.01	2.09	62.48	13.42
865	2	17.06±4.50	1943.63±314.26	1.13±0.06					21.85
866	2	16.61±4.58	2087.35±377.01	1.14±0.07					16.67
867	1	16.76±4.60							35.66
868	3	16.14±4.04	15.79±43.12	1.15±0.83	0.080±0.07	0.04±0.01	6.03	72.39	34.44
869	1	15.74±4.47							26.14
870	1	16.75±4.42							19.28
871	1	15.78±4.22							14.35
872	1	17.72±4.90							25.90
873	1	18.63±4.98							25.45
874	1	17.09±4.75							47.84
875	2	15.95±4.29	2260.32±248.72	1.89±0.07					28.31
876	2	16.01±3.93	1943.94±324.19	1.33±0.07					21.55
877	1	17.84±4.78							25.48
878									
879	1	16.00±4.16							21.80
880	2	16.76±3.70	760.72±200.56	0.76±0.08					19.41
881	1	19.48±5.59							35.14
882	1	16.73±4.76							25.75
883	1	17.03±4.59							36.91
884	1	16.53±4.21							16.43
885	1	16.48±4.14							14.75
886	2	16.51±4.42	1461.86±208.63	1.04±0.05					17.04
887	2	25.83±7.91	6382.31±3070.48	2.22±0.91					27.13
888	1	17.46±4.19							11.52
889	1	16.42±4.09							31.37
890	1	27.18±8.33							45.42
891	1	15.65±3.87							20.99
892	1	16.85±4.92							16.36
893	1	15.74±4.14							25.61

|									
894	3	15.51±4.07	2999.99±181.70	0.90±0.06	0.003±0.06	0.23±0.09	13.77	66.92	12.58
895	1	16.84±4.62							28.08
896	1	17.49±4.91							28.21
897	1	17.81±4.76							21.21
898	1	16.76±4.28							38.76
899	1	15.49±4.02							22.92
900	1	17.73±4.80							14.52
901	1	7.51±4.10							18.91
902	1	6.91±4.37							21.23
903	1	7.93±4.51							16.56
904	1	7.79±4.37							12.77
905	1	7.93±4.80							19.00
906	1	8.09±4.52							17.29
907	2	5.0±3.13	8363.99±2605.55	11.05±2.67					158.75
908	1	8.45±4.86							67.91
909	1	8.89±5.25							40.75
910	2	6.13±4.81	5769.98±649.21	9.84±0.89					22.21
911	2	7.87±4.69	2705.62±462.15	2.20±0.15					14.18
912	1	8.40±4.70							40.05
913	1	7.95±4.78							29.63
914	1	7.58±4.75							29.81
915	1	7.20±5.43							25.07
916	1	7.32±4.96							32.45
917	1	8.00±4.75							24.39
918	1	7.32±5.11							30.71
919	2	7.44±4.70	853.88±51.99	2.00±0.05					23.17
920	3	7.10±4.35	65.81±36.15	6.25±1.96	0.196±0.09	0.01±0.009	3.48	66.64	17.39
921	7	7.02±5.42	241.31±18.34	0.52±0.05	0.002±0.0002	-0.13±0.011	14.91	74.31	50.08
922	3	6.65±3.74	214.40±16.09	4.48±0.29	0.299±0.0000001	0.013±0.0001	2.09	62.66	13.43
923	1	5.63±2.72							12.72
924	1	5.10±2.96							20.93
925	1	4.74±2.34							47.19
926	1	3.91±1.97							24.75
927	1	4.43±1.41							54.12
928	1	1.83±0.84							121.95
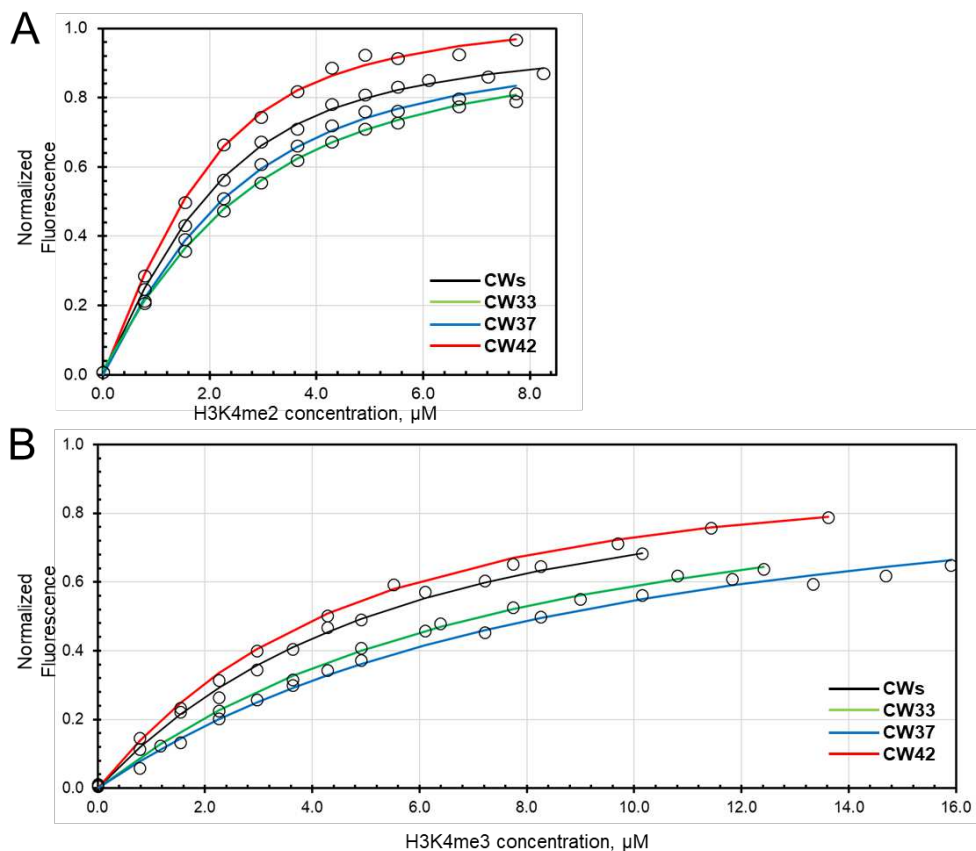
Fig. S1

K_d determinations of CWs(—),CW42(—),CW37(—) and CW33(—) for H3K4me2 and H3K4me3. Normalized ΔFluorescence intensity values plotted against ligand concentrations (0.0 - 8.2 µM for H3K4me2, 0.0 – 15.9 µM for H3K4me3). Protein concentrations were constant throughout any one titration, but could vary somewhat from construct to construct (always within 2.0 - 2.4 µM). Three parallels were performed for each set of lignad-construct combination. The data were fitted using non-linear least-square methods to Eq. 1, yielding three K_d values in each instance. The K_d values given in the main document, Fig. 1D, are averages of these. The errors quoted in Fig. 1 (main document) are one standard deviation, n = 3. Although the data does not plateau completely, the determined K_ds were well able to differentiate between the construct-lignad interactions, and were used to select CW42 for further work.
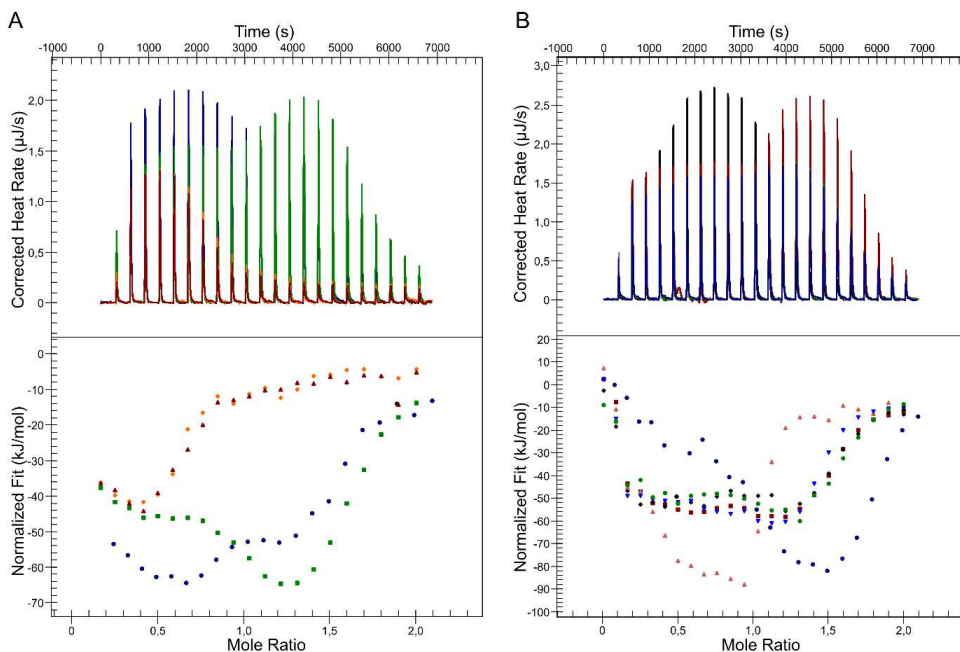
Fig. S2

Loss of specific binding caused by mutations in the H3-mimicking peptide and the post-α1 loop. A) CW42 Q923A mutant interacting with the H3K4me1 ligand. (Top) Selection of ITC heat flows in responses to titration of ligand in the protein solution. (Bottom) Normalized integrated peak areas. 3 out of 4 parallels performed are shown. B) Selection of ITC binding, CW42:H3K4me1 titrated with the T6A-modified peptide. (Top) Selection of ITC heat flows in responses to titration of ligand in the protein solution. (Bottom) Normalized integrated peak areas. 7 out of 12 attempts at acquiring reproducible isotherms are shown.
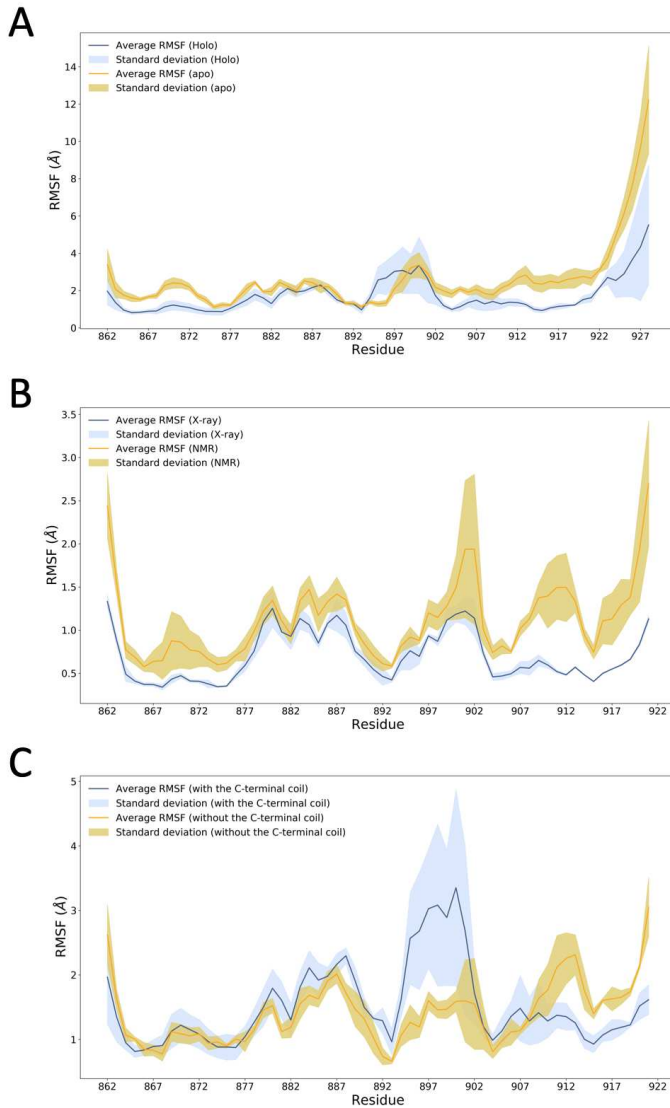
Fig. S3

 Average RMSF and standard deviation of the CW constructs assessed by MD. (A) Wild type NMR *holo* and *apo* states, (B) the crystal structure (X-ray) and the modified NMR structure (NMR structure with the E917A mutation and without the C-terminal coil), (C) the wild-type NMR CW42:H3K4me1 complex with and without the C-terminal coil.

References for Supplementary Information

1   Dobrovolska O, Bril'kov M, Odegard-Fougner O, Aasland R & Halskau O (2018) (1)H, (13)C, and (15)N resonance assignments of CW domain of the N-methyltransferase ASHH2 free and bound to the mono-, di- and tri-methylated histone H3 tail peptides, *Biomol NMR Assign* doi:10.1007/s12104-018-9811-x.

2   Keller R L J (2005) *Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment.*

3   Guntert P (2004) Automated NMR structure calculation with CYANA, *Methods Mol Biol* 278, 353-78 doi:10.1385/1-59259-809-9:353.

4   Shen Y & Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks, *J Biomol NMR* 56, 227-41 doi:10.1007/s10858-013-9741-y.

5   Pettersen E F, Goddard T D, Huang C C, Couch G S, Greenblatt D M, Meng E C & Ferrin T E (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput Chem* 25, 1605-12 doi:10.1002/jcc.20084.

6   Meiboom S & Gill D (1958) Modified Spin-Echo Method for Measuring Nuclear Relaxation Times, *Rev Sci Instrum* 29, 688-691 doi:Doi 10.1063/1.1716296.

7   Lipari G & Szabo A (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity, *J Am Chem Soc* 104, 4546-4559 doi:DOI 10.1021/ja00381a009.

8   Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K & Schulten K (1999) NAMD2: Greater scalability for parallel molecular dynamics, *J Comput Phys* 151, 283-312 doi:DOI 10.1006/jcph.1999.6201.

9   Best R B, Zhu X, Shim J, Lopes P E M, Mittal J, Feig M & MacKerell A D (2012) Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles, *J Chem Theory Comput* 8, 3257-3273 doi:10.1021/ct300400x.

10 Brooks B R, Brooks C L, Mackerell A D, Nilsson L, Petrella R J, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner A R, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor R W, Post C B, Pu J Z, Schaefer M, Tidor B, Venable R M, Woodcock H L, Wu X, Yang W, York D M & Karplus M (2009) CHARMM: The Biomolecular Simulation Program, *Journal of Computational Chemistry* 30, 1545-1614 doi:10.1002/jcc.21287.

11 Feller S E, Zhang Y H, Pastor R W & Brooks B R (1995) Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method, *J Chem Phys* 103, 4613-4621 doi:Doi 10.1063/1.470648.

12 Humphrey W, Dalke A & Schulten K (1996) VMD: Visual molecular dynamics, *J Mol Graph Model* 14, 33-38 doi:Doi 10.1016/0263-7855(96)00018-5.

uib.no