A COMPUTATIONAL DESIGN STRATEGY FOR DISCOVERING THE PROTEIN TARGETS OF HISTONE LYSINE METHYLTRANSFERASES

A thesis submitted for the degree of Doctor of Philosophy

of Imperial College London by

Diego Alonso Martínez

Department of Chemical Engineering

Imperial College London

London SW7 2AZ

United Kingdom

2017

To my parents,

José Manuel and María Jesús

Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Declaration of Originality

I hereby declare that the work presented here is my own and all references made to other work have been appropriately acknowledged and cited in the text.

© 2017 Diego Alonso Martínez

"Deep in the sea, all molecules repeat the patterns of one another till complex new ones are formed. They make others like themselves… and a new dance starts.

Growing in size and complexity... living things, masses of atoms, DNA, protein… dancing a pattern ever more intricate.

Out of the cradle onto the dry land... here it is standing… atoms with consciousness… matter with curiosity.

Stands at the sea… wonders at wondering… I… a universe of atoms… an atom in the universe."

Richard Feynman (1918-1988)

Acknowledgements

I would like to acknowledge my supervisor Dr Peter DiMaggio for giving me the opportunity of working in his lab and for his support, being available most of the times I wanted to have a chat with him. I would also like to thank the members of the DiMaggio lab for our extensive conversations throughout the years. I also cannot forget about the great MEng students I had the pleasure to teach and supervise, who helped me confirm my passion for learning and teaching. In addition, many thanks to Imperial College London for awarding me the President's PhD Scholarship and for funding my placement at MIT.

I would also like to acknowledge some collaborators I managed to meet during these years: Professor Bruce Tidor (MIT) and the members of his lab for the insightful talks during my placement there; Dr Matthew Fuchter for his insights about chemical biology; Fabio Fisher (Imperial) for his help on protein purification; Juan Eiros-Zamora (Imperial), Dr Ornella Di Pietro (Imperial) and Dr Shozeb Haider (UCL) for the talks about simulating biomolecules and Professor Luis Aragon-Alcaide (Imperial) for offering me the opportunity of working in his group and broaden my skills in performing simulations. In addition, I would like to thank Professor Amparo Galindo for her priceless support and advice.

During my PhD years, I had the chance to meet some of the best people I have ever met. These are the ChemEng PhDs. Not only our trips to Copenhagen, Dublin, Bristol, Manchester, Liverpool (even Southend-on-Sea!) were great, but most importantly the everyday lunches, dinners, walks and conversations where we supported each other during this challenge called PhD. I would also like to thank my friends back in my hometown, my colleagues at the University of Valladolid and my Erasmus friends for not letting the distance stop our friendship, and to my family for supporting me in the bad moments I had to overcome, in particular to my aunts that happen to be both called Nieves.

Finally, the most important acknowledgements. I would like to thank my mother Maria Jesús Martínez Rodríguez, for pushing myself to overcome my limits, showing me the value of effort and constancy and being the definition of the term "strength", and my father José Manuel Alonso Gutiérrez for making me believe in myself and teaching me that nothing is impossible, always with a smile in his face.

You were truly a great father, teacher and friend. An example to follow.

Abstract

The activity of lysine methyltransferases (KMTs) is crucial for the regulation of chromatin-related processes, such as gene expression. Since all KMTs catalyse the transfer of a methyl group from the same SAM cofactor, no current technology is able to directly assess which KMT has performed each observed methylation in a cellular context. This information would be particularly informative in the development of more specific treatments for epigenetically-regulated diseases, such as cancer. To address this technology gap, this work aims at developing a "methylome" profiling assay based on engineering synthetic SAM cofactors specific towards a given KMT.

Our proposed computational design strategy takes aligned crystal structures of KMTs as input to identify unique cavities in wild type KMTs that could accommodate complementary synthetic "bumps" on SAM. SAM analogues with sterically matching "bumps" are then designed, and molecular dynamics simulations are utilised to predict their binding mode and affinity with the target KMT to rank the designs. The highest-ranked SAM analogues are synthesised and include an isotopically-heavy methyl donor group (CD_3), which retains the biological-consistency of cellular methylation and renders new methylation events detectable via MS. The ability of the KMT to utilise the synthesised SAM analogues as a cofactor for methylation is finally assessed using *in vitro* assays and quantified via LC-MS/MS proteomics.

The results of this work identified bumped SAM analogues that were selective to certain KMTs lacking a post-SET domain. Specifically, a natural pocket was identified in the SET domain of SETD7 that does not exist in post-SET containing KMTs, such as G9a. Complementary bumped SAM analogues were designed and synthesised for experimental validation. A computational screening approach using docking, molecular dynamics and binding affinity estimation revealed that the binding mode and affinity for some of these bumped SAM analogues were comparable to the one of native SAM for SETD7, but not for G9a. Subsequent *in vitro* experiments using recombinant G9a and SETD7 with our engineered SAM analogues and nucleosome substrates revealed cofactor selectivity towards SETD7. These results were consistent with the predictions of the computational screening approach, which supports the feasibility of our proposed design methodology.

Table of contents

ABSTRACT 1

TABLE OF CONTENTS 2

LIST OF TABLES 7

LIST OF FIGURES 8

NOMENCLATURE AND ABBREVIATIONS 14

CHAPTER 1 INTRODUCTION TO EPIGENETICS 21

1.1. The nucleosome and chromatin 22

1.2. Epigenetic control of gene expression and its effect on disease 24

1.2.1. DNA methylation 24

1.2.2. Histone post-translational modifications 26

CHAPTER 2 REVIEW OF HISTONE METHYLATION 31

2.1. Histone methylation 31

2.2. Histone lysine methyltransferases 32

2.2.1. G9a 35

2.2.2. SETD7 39

2.2.3. DOT1L 40

2.3. Histone and non-histone targets of PMTs 41

2.3.1. G9a substrates 44

2.3.2. SETD7 substrates 45

2.3.3. DOT1L substrates 45

CHAPTER 3 EPIGENETIC PROFILING METHODS: BACKGROUND AND MOTIVATION 49

3.1. Substrate methylation detection methods 50

3.1.1. Radiometric quantification 52

3.1.2. Antibody-based detection 53
3.1.3. Mass spectrometry 54
3.1.3.1. Top-down MS 55
3.1.3.2. Middle-down MS 56
3.1.3.3. Bottom-up MS 56

3.2. SAH detection methods 58

3.3. Bump-and-hole approach 59
3.3.1. Bump-and-hole approach applied to kinases 61
3.3.2. Bump-and-hole approach applied to protein methyltransferases (PMTs) 63

3.4. Motivation and approach 68

CHAPTER 4 DESIGN OF SAM COFACTOR ANALOGUES SELECTIVE TO ENDOGENOUS KMTS 71

4.1. KMT target identification methods 73
4.1.1. MS-based methods 73
4.1.2. Chemical biology-based methods 75
4.1.3. Technical challenges and opportunities 76

4.2. Methodologies 77
4.2.1. Overview of proposed design methodology 77
4.2.1.1. Computational identification of natural KMT pockets 78
4.2.2. Materials and methods 80
4.2.2.1. G9a expression 80
4.2.2.2. G9a purification 82
4.2.2.3. Protein gel analysis 83
4.2.2.4. G9a in-solution digestion and MS identification 84
4.2.2.5. *In vitro* methyltransferase assay 85
4.2.2.6. Mass spectrometry workflow for histones 86
4.2.2.7. Targeted LC-MS method to characterise histone PTMs 88
4.2.2.8. Estimation of the abundance of heavy methylation 89

4.3. Results and discussion 90
4.3.1. Computational validation: DOT1L “hole” 90
4.3.2. SETD7 vs a post-SET domain-containing KMT (G9a) 92
4.3.2.1. Identification of structural differences unique to SETD7 92
4.3.2.2. G9a purification and characterization 93
4.3.2.3. SAM cofactor analogues design and synthesis 95

4.3.2.4. *In vitro* assays with SETD7 and SAM cofactor analogues 96
4.3.2.5. *In vitro* assays with G9a and SAM cofactor analogues 97

4.4. Conclusions and future work 99

CHAPTER 5 *IN SILICO* SCREENING OF ENGINEERED KMT-SAM SYSTEMS 101

5.1. Introduction to molecular dynamics 103
5.1.1. The basis of molecular dynamics simulations 103
5.1.2. Methods to calculate the binding affinity 105
5.1.2.1. Free energy perturbation (FEP) 106
5.1.2.2. Thermodynamic integration (TI) 107
5.1.2.3. Docking 108
5.1.2.4. Linear interaction energy (LIE) 109
5.1.2.5. Molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) 110
5.1.3. Molecular dynamics simulations of KMTs 111
5.1.3.1. MD simulations of SETD7 and G9a-like protein 112
5.1.3.2. MD of PMTs to generate binding affinity estimates 113
5.1.4. Technical challenges and opportunities 114

5.2. Methodologies 116
5.2.1. Pre-processing 119
5.2.1.1. Pre-processing the cofactor structures 119
5.2.1.2. Pre-processing the KMT structures 119
5.2.1.3. Preparing the PDB file of the cofactor-protein complex 120
5.2.2. Parametrization of the complexes 120
5.2.2.1. Parametrization of the SAM cofactor analogues 120
5.2.2.2. Parametrization of the G9a complex 120
5.2.2.3. Parametrization of the complexes of SETD7 with LEaP 122
5.2.3. Molecular dynamics simulations 123
5.2.4. CPPTRAJ post-analysis of the MD trajectories 124
5.2.4.1. RMSD Graphs 124
5.2.4.2. Distances and angles 124
5.2.4.3. Clustering of MD trajectories 125
5.2.5. MMPBSA calculation of binding affinities 126
5.2.5.1. Generating prmtop files for the binding affinity estimation 126
5.2.5.2. Running MMPBSA.py 127

5.3. Results and discussion 128
5.3.1. Pre-processing results 128
5.3.1.1. Protonation state of the SAM analogues major microspecies 128

5.3.1.2. Atom types and partial charges of the SAM analogues 128
5.3.1.3. Parametrization of the ZnCys motif 131
5.3.1.4. Prmtop and inpcrd files for the solvated enzyme-cofactor complexes 132
5.3.2. RMSD evolution of the protein backbone 133
5.3.3. Distances and angles to enable methyl transfer 135
5.3.3.1. G9a distances and angles analysis 136
5.3.3.2. SETD7 distances and angles analysis 139
5.3.4. Representative structures of the major clusters 142
5.3.5. MM-PBSA calculation of the binding affinities 144

5.4. Conclusions and future work 146

CHAPTER 6 A BUMP-AND-HOLE APPROACH TO IDENTIFY G9A'S TARGETS 149

6.1. Introduction 151
6.1.1. Introduction to G9a 151
6.1.2. Non-histone targets of G9a and methods for non-histone target identification 152
6.1.3. MD simulations for the bump-and-hole approach 153
6.1.4. Technical challenges and opportunities 154

6.2. Methodologies 155
6.2.1. Rosetta structure prediction 155
6.2.1.1. Sequence alignment 156
6.2.1.2. Model generation 156
6.2.1.3. Refinement of the structure using the Rosetta full-atom energy function. 156
6.2.1.4. Clustering and model selection 157
6.2.2. Proposed KMT design strategy 157
6.2.2.1. Identification of candidate G9a residues for mutation 158
6.2.2.2. Computational screening of mG9a candidates 160
6.2.3. Experimental materials and methods 163

6.3. Results and discussion 164
6.3.1. G9a structural characterization for a bump-and-hole design 164
6.3.1.1. Bonding interactions of G9a with SAM 164
6.3.1.2. Suitability of SAM design atoms and complementary G9a mutations 167
6.3.1.3. Assessment of G9a mutations to fit C2-modified SAM analogues 168
6.3.1.4. Candidate groups to incorporate on the C2 atom of SAM 170
6.3.1.5. Structural integrity of mutations to the post-SET domain: the ZnCys motif. 172
6.3.1.6. Proposed mutations of G9a Q1169 174
6.3.2. Computational screening of proposed mG9a candidates using Rosetta 176
6.3.2.1. Negative control: mutations to conserved residues 176

6.3.2.2. Rosetta folding predictions for proposed mG9a candidates 178
6.3.3. MD simulations of G9a Q1169A with SAM and SAM analogues 180
6.3.3.1. G9a Q1169A distances and angles analysis 181
6.3.3.2. Representative structure of the major cluster 185
6.3.3.3. MM-PBSA calculation of the binding affinities 186
6.3.4. Experimental validation 188

6.4. Conclusions and future work 190

CHAPTER 7 CONCLUSIONS AND FUTURE WORK 193

7.1. Conclusions 193

7.2. Implications and future work 197

BIBLIOGRAPHY 201

APPENDICES 219

List of tables

Table 2-1. Sites of histone 3 lysine methylation and their relationship with activating or repressing chromatin according to their methylation state [58-61] 32
Table 2-2. Histone lysine methyltransferases according to their target residues within histones H1, H3 and H4 and the degree of methylation they catalyse [55,67,68] 34
Table 5-1. Energetic analysis for the SETD7-SAM analogue complexes as obtained by MM–PBSA calculations. 144
Table 6-1. Main applications within Rosetta [300,305]. 155
Table 6-2. Design objectives of the mG9a-SAM analogue complex and their influence on the analysis 158
Table 6-3. Identified hydrogen bonds between S-Adenosyl-homocysteine and G9a. BB: backbone. SC: side chain 165
Table 6-4. Possible SAH atoms to modify and analysis of the closest G9a residues to mutate. 167
Table 6-5. Energetic analysis for the G9a Q1169A-SAM analogue complexes as obtained by MM–PBSA calculations. 187
Table A-1. Copyright permissions. 227

List of figures

Figure 1-1. Basic structure of the nucleosome (PDB accession code: 5AV6). Histone H3 is depicted in red, H4 in yellow, H2A in cyan, H2B in blue and DNA in grey. Histone H3 lysines K4, K9, K27 and K36 (located at the histone H3 tails) and K79 (located at the core region) are also represented as orange sticks. 22
Figure 1-2. Chromatin structure. Adapted from Jansen et al. [12]. 23
Figure 1-3. Role of CpG islands in transcriptional regulation. When DNA is not methylated (white circles) on promoter CpG islands, transcription occurs. However, when they are methylated (black circles), transcription is blocked and the gene is silenced. 25
Figure 1-4. Main histone PTM sites. Me: monomethylation; me2: dimethylation; me3: trimethylation; ac: acetylation; fo: formylation; oh: hydroxylation [37]. 27
Figure 1-5. Types of epigenetic regulators and examples [42]. 28
Figure 2-1. Mechanism of histone lysine methylation by lysine methyltransferases [42]. 33
Figure 2-2. Chemical mechanism for the methylation of a histone lysine [64]. 33
Figure 2-3. Phylogenetic tree of the human HKMTs [65]. G9a is highlighted in green, SETD7 in blue and DOT1L in orange. The phylogeny is based on a multiple sequence alignment of the methyltransferase domain including the N-SET, Pre_SET, SET, I-SET, and Post-SET motifs for the SET-domain containing HKMTs. 35
Figure 2-4. A) G9a domains. SAH is represented with yellow sticks and the Zn^{2+} ions are represented as cyan spheres. B) G9a electrostatic potential (+5 kcal/electron units = blue, -5 kcal/electron units = red), with SAH represented with yellow sticks. Adapted from Wu et al. [65]. 36
Figure 2-5. A) The helix-turn-helix-β-turn structure of the ankyrin repeats in GLP. B) H3K9me2 bound to GLP in a partial hydrophobic cage. GLP is represented in green, while the histone H3 peptide is represented as grey sticks [73]. 37
Figure 2-6. Schematic catalytic mechanism of SETD7 [40]. 39
Figure 2-7. Comparison of the aligned structures of SETD7 (green, PDB:1XQH) and G9a (cyan, PDB: 2O8J) and the location of their crucial residues. The structure of SAH bound to SETD7 is shown as a reference (magenta sticks). A: SETD7's Y245 and G9a's Y1067. B: SETD7's Y305 and G9a's P1121. A: SETD7's Y335 and G9a's Y1154. 40
Figure 2-8. Structure of DOT1L's catalytic core (green) bound to SAM (magenta sticks). 41
Figure 2-9. Methylation network between lysine methyltransferases (KMT) and their substrates, where each of the lines indicates an interaction. Adapted from Biggar & Li [93]. 43
Figure 3-1. Summary of protein methyltransferases activity assays and chemicals used in them [140]. Reprinted with permission from Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. ACS Chem. Biol. 7, 443–63 (2012). Copyright 2012 American Chemical Society. 51

Figure 3-2. SPA-based HTS approach. The β particles emitted by the radioactive groups are received by the SPA beads attached to the biotin–tagged substrates, emitting light that can be detected and related to the methylation state. Adapted from Ibáñez [139]. 52
Figure 3-3. Illustrative model of LANCE Ultra and AlphaLISA epigenetic assays [144]. Reprinted with permission from Gauthier, N. et al. Development of homogeneous nonradioactive methyltransferase and demethylase assays targeting histone H3 lysine 4. J. Biomol. Screen. 17, 49–58 (2012). Copyright 2012 Society for Laboratory Automation and Screening. 54
Figure 3-4. Simplified schemes of the bottom-up, middle-down, and top-down MS strategies to analyse histone PTMs [147]. 55
Figure 3-5. Overview of mass spectrometry assays for proteomic studies [4]. 58
Figure 3-6. Differences in requirements and applications for chemical probes and drugs [168]. 59
Figure 3-7. Schematic representation of the "bump-and-hole" approach. 60
Figure 3-8. ATP chemical structure. 61
Figure 3-9. Schematic ATP binding site for the majority of protein kinases [171]. 61
Figure 3-10. A) Native ATP. B) N6-(cyclopentyl) ATP analogue. 63
Figure 3-11. SAM chemical structure. 63
Figure 3-12. SAM analogues acting as cofactor alternatives for PMTs, with Hey-SAM highlighted in green [140]. Reprinted with permission from Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. ACS Chem. Biol. 7, 443–63 (2012). Copyright 2012 American Chemical Society. 65
Figure 3-13. Scheme of the chemoenzymatic strategy to profile genome-wide chromatin modifications via the formation of the S-alkyl analogues at *in vivo* conditions [137]. Reprinted with permission from Wang, R. et al. Profiling Genome-Wide Chromatin Methylation with Engineered Posttranslation Apparatus within Living Cells. J. Am. Chem. Soc. 135, 1048−1056 (2013). Copyright 2013 American Chemical Society. 66
Figure 3-14. General overview of the approach to develop a MS-compatible method to profile the targets of different KMTs. 68
Figure 4-1. Mesh discretization of the coordinate space and grid occupancy of every atom. Steric clashes between the KMT and the inserted atoms are shown in yellow. 79
Figure 4-2. Structural differences between the SAM binding pockets of DOT1L (a) and G9a (b). These differences could enable the differential binding of SAM analogues when inserting groups at SAM N7 position. G9a residues H1113 and W1159 that clash with carbon atoms inserted at SAM's N7 position (represented as spheres) are shown in yellow. 90
Figure 4-3. Structural differences between the SAM binding pockets of SETD7 (a) and G9a (b). These differences enable the differential binding of SAM analogues when inserting groups at SAM C2 position. G9a residues T1167 and Q1169 clashing with carbon atoms inserted at SAM's C2 position (represented as spheres) are shown in yellow 92
Figure 4-4. Spectrum obtained during the FPLC purification of G9a. The fractions collected appear in red at the bottom of the figure. 94

Figure 4-5. SDS-PAGE gel electrophoresis analysis of even fractions of the expressed G9a after FPLC. The added band (in green) indicates the MW expected for G9a (34.5 kDa). 94
Figure 4-6. Proposed C2-substituted SAM cofactor analogues to be specific towards SETD7 vs G9a..... 95
Figure 4-7. SETD7 H3K4me1 proteoform abundance vs relative abundance of heavy methylation for different SETD7 and SAM concentrations. 96
Figure 4-8. SETD7 H3K4me1 proteoform abundance vs relative abundance of heavy methylation for different SAM analogues. The error bars indicate the standard deviation of the measurements. 97
Figure 4-9. G9a peptide proteoform abundances vs relative abundance of heavy methylation with different SAM analogues. Un = unmethylated. 98
Figure 5-1. Scheme of force field interactions. b represents the distance, θ the angle and ϕ the dihedral. Covalent bonds are shown with solid lines, whereas non-bonded interactions are shown with a dashed line. Adapted from Ponder et al. [221]. 103
Figure 5-2. Thermodynamic cycle used to calculate the relative binding free energy difference, $\Delta\Delta G^{0}_{Bind}$, between two different ligands A and B, shown in blue and red respectively. $\Delta G^{0}_{Bind, A}$ and $\Delta G^{0}_{Bind, B}$ correspond to the binding free energies of each of the ligands, while $\Delta G^{0}_{TI, solvated}$ and $\Delta G^{0}_{TI, bound}$ are the variations in free energy occurring when transforming the ligands in the bound and solvated (or unbound) state, respectively. Adapted with permission from Steinbrecher, T., Case, D. A. & Labahn, A. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. J. Med. Chem. 49, 1837–1844 (2006). Copyright 2006 American Chemical Society[226]. 106
Figure 5-3. Methodology followed for the pre-processing (red), parametrization (green), MD simulations (blue) and post-processing (orange) of the cofactor-KMT pairs under study. All the programs that are external to the Amber16 suite appear underlined. 118
Figure 5-4. Distance (d) between the oxygen of the hydroxyl group of a G9a/SETD7 catalytic tyrosine with the carbon of the methyl transfer group and angle (α) formed between these two atoms and the sulphur atom of SAM. 125
Figure 5-5. Relative abundance (RA) and protonation state prediction of the major microspecies of native SAM, Pip-SAM, MPip-SAM, TPhe-SAM and Phe-SAM at pH = 7.4 and T = 310K. 129
Figure 5-6. Amber atom types (A) and partial charges (B) for each of the atoms of SAM, Pip-SAM, MPip-SAM, TPhe-SAM and Phe-SAM. 130
Figure 5-7. G9a's Zn motif parametrised by MCPB.py using the bonded model. The Zn atom is shown in grey, the S atoms of the CYS surrounding the Zn in yellow and G9a in green. Native SAM is shown in black for reference. 131
Figure 5-8. Primary cell of the simulation of endogenous G9a with native SAM. G9a is shown in green, SAM in blue, Na^{+} counter ions in purple, Zn atom in grey and water molecules in red. 133
Figure 5-9. Root-mean-square-deviation (RMSD, Å) for the backbone atoms of G9a and SETD7 with regards to the initial frame for the production trajectory 134
Figure 5-10. Experimental distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 in G9a (a) or Y254 in SETD7 (b). 135

Figure 5-11. A) Evolution of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 during 100 ns of simulation. B) Probability distribution function of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 for the last 80 ns of the simulation. 137
Figure 5-12. A) Evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 during 100 ns of simulation. B) Probability distribution function of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 for the last 80 ns of the simulation. 138
Figure 5-13. A) Evolution of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 during 100 ns of simulation. B) Probability distribution function of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 for the last 80 ns of the simulation. 140
Figure 5-14. A) Evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 during 100 ns of simulation. B) Probability distribution function of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 for the last 80 ns of the simulation. 141
Figure 5-15. Differences in the relative positioning of Pip-SAM (cyan) to a) G9a's Y1067 (green) and b) SETD7's Y245 (magenta) based on their respective representative structures obtained from the MD production trajectories. Both the angle and the distance between Pip-SAM's methyl group (depicted in black) and the catalytic tyrosine are not adequate for G9a, but for SETD7 they resemble the values obtained with native SAM. 142
Figure 5-16. Comparative of the interactions of SAM (a) or piperidine-SAM (b) with SETD7. 143
Figure 6-1. Description of a hydrogen bond. D: donor, H: hydrogen, Ac: acceptor, An: antecedent of the acceptor, d: distance donor-acceptor, maximum 3.9 Å. 159
Figure 6-2. Methodology followed for the identification of candidate G9a residues for mutation (red) and computational screening of mG9a candidates (green). All the programs that are external to the Rosetta 3.5 suite appear underlined. 162
Figure 6-3. Detail of G9a's SET domain (green surface) bound to SAH (magenta sticks), both from PDB: 2O8J, and H3K9me2 (cyan sticks) from PDB: 2RFI. 165
Figure 6-4. Hydrogen bonding interactions between SAH (yellow sticks) and G9a (green sticks) from two viewpoints. Generated with PyMOL using PDB: 2O8J. 166
Figure 6-5. G9a's sub-domains within the SET domain shown with SAH (red sticks) and the possible atoms to insert groups at. 166
Figure 6-6. Statistical significance of the frequency of selected G9a residues located closer to each SAM atom candidate using pLogo. N(fg) indicates the number of aligned foreground sequences and

n(bg) the number of aligned background sequences used to generate the image, respectively. The red horizontal bars on the pLogo correspond to P = 0.05 169
Figure 6-7. G9a's SAM binding pocket and hydrophobic channel residues according to their conservation state. Red: high conservation; white: average conservation; blue: low conservation. G9a's Q1169 residue and post-SET domain are indicated. 170
Figure 6-8. Insertion of groups at SAM's C2 position (magenta sticks) to analyse clashes with G9a's residues (green). Left: surface view. Right: sticks view 171
Figure 6-9. ZnCys motif of G9a. a) Front view. B) Side view. SAH is represented as blue sticks, Zn^{2+} as a grey sphere, G9a's pseudoknot C1115 as orange sticks and G9a's post-SET domain C1168, C1170 and C1175 as magenta sticks. 173
Figure 6-10. Analysis of the presence of a ZnCys motif in the post-SET domain of lysine and arginine methyltransferases based on available crystal structures. Adapted from a phylogenetic tree by Wu et al. [65]. 173
Figure 6-11. The ZnCys motif is frequent in H3K9 KMTs. G9a is shown in green, GLP in cyan, SUV39H2 in magenta and ASH1L in light grey. The C2 position of SAM is indicated, as well as the Q1169 residue of G9a and its equivalent residues in the other KMTs (depicted as sticks). 175
Figure 6-12. Conservation of the ZnCys motif in human H3K9 methyltransferases. The cysteine residues involved in the ZnCys motif are highlighted in orange. The mG9a candidate Q1169 and its equivalent residues in other KMTs are highlighted in blue 175
Figure 6-13. G9a's post-SET domain fold is altered (RMSD = 2.203 Å) when C1168, C1170 and C1175 are mutated to glycine. Endogenous G9a is depicted in cyan, the model of mG9a in pink and SAH in yellow. 177
Figure 6-14. Surface representation of native SAH bound to endogenous G9a (a) and the same SAH structure superposed on an aligned model of C1168G, C1170G and C1175G G9a (b). The mutations in the highly-conserved cysteine residues strongly alter the binding pocket, possibly resulting in a loss of G9a's activity. 178
Figure 6-15. G9a's post-SET domain conserves its fold (RMSD = 0.938) when Q1169 (cyan sticks) is mutated to glycine (green sticks). Endogenous G9a is depicted in cyan, the model of Q1169G in green and SAH in yellow 179
Figure 6-16. Surface representation of C2 Phe-SAH bound to endogenous G9a (a) and the same C2 Phe-SAH structure superposed on an aligned model of G9a Q1169G (b). While the phenyl group causes a steric clash with endogenous G9a, the Q1169G mutation generates an additional space that could suffice to accommodate the phenyl bump. 179
Figure 6-17. Root-mean-square-deviation (RMSD, Å) for the backbone atoms of G9a Q1169A with regards to the initial frame for the production trajectory. 181
Figure 6-18. A) Evolution of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 during 100 ns of simulation. B) Probability distribution function of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 for the last 80 ns of the simulation 183

Figure 6-19. A) Evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 during 100 ns of simulation. B) Probability distribution function of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 for the last 80 ns of the simulation. 184
Figure 6-20. Differences in the relative positioning of Phe-SAM (cyan sticks) to a) G9a's Y1067 (green sticks) and b) G9a Q1169A's Y1067 (orange sticks). 185
Figure 6-21. The interactions of Phe-SAM with G9a Q1169A (a) appear mostly preserved when compared to the interactions between SAH and endogenous G9a (b), except for the residues in the post-SET domain. 186
Figure 6-22. SDS-Page gel showing G9a Q1169A after FPLC purification. 188
Figure A-1. First minimization with 2 kcal/mol energy restraints 219
Figure A-2. Second minimization without energy restraints. 219
Figure A-3. First heating for 500 ps from 0 to 100K with 10 kcal/mol energy restraints on the complex at constant volume. 220
Figure A-4. Second heating for 500 ps from 100K to 310K without energy restraints at constant pressure. 221
Figure A-5. Production run for 50 ns at 310K at constant pressure. 222
Figure A-6. CPPTRAJ input to perform the clustering of MD trajectories. 223
Figure A-7. A) MATLAB code to assign conservation values based on the frequency of each residue at each position in a multiple sequence alignment according to the BLOSUM62 matrix. B) Finder function definition. 224
Figure A-8. BLOSUM62 matrix. 225
Figure A-9. Example of Rosetta homology modelling script. 225
Figure A-10. Example of clustering algorithm operation in Rosetta. 226

Nomenclature and abbreviations

ACIN1/ACINUS	Apoptotic chromatin condensation inducer in the nucleus
ALS	Amyotrophic lateral sclerosis
AM1-BCC	Semi-empirical charge model with bond charge correction
AQUA	Absolute quantification
ASH1L	ASH1 like histone lysine methyltransferase
ATP	Adenosine triphosphate
Azide-A-DSBSO	Azide-tagged, acid-cleavable disuccinimidyl bis-sulfoxide
BAH	Bromo-adjacent homology domain
BB	Backbone
BDRG	Biotin-aspartate-Rink-glycine reagent
BET	Bromo- and extra-terminal domain
BLOSUM62	Blocks of amino acid substitution matrix 62
BMP	Bone morphogenetic proteins signaling
Brd4	Bromodomain-containing protein 4
c-Src	Proto-oncogene tyrosine-protein kinase Src
C18	Octadecyl carbon chain
Cα	Alpha carbon
CARM1	Coactivator-associated arginine methyltransferase
CASP	Critical assessment of protein structure prediction
CD_3	Deuterated methyl group
CD_3BrSAM	Deuterated 7-bromo-deaza-SAM
CDYL / CDYL1	Chromodomain Y-like protein
Cfp1	CXXC finger protein 1
CGAL	Computational geometry algorithms library
CH_3	Methyl group
ChIP	Chromatin immunoprecipitation
ChIP - seq	Chromatin immunoprecipitation sequencing
ChIP-on-chip	Chromatin immunoprecipitation with DNA microarray
CID	Collision induced dissociation
CliEn-seq	Clickable chromatin enrichment with parallel DNA sequencing
CPPTRAJ	C++ process trajectory
crd	Coordinate
CTCL	Cutaneous T cell lymphoma
Da	Dalton
DE52	Preswollen microgranular diethylaminoethyl cellulose
DKS	Dynamic kinetic stability
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DNMT1	DNA (cytosine-5)-methyltransferase 1
DOT1L	DOT1 like histone lysine methyltransferase

DP1	Transcription factor Dp-1
dsRNA	Double-stranded RNA
dt	Time step
DTT	Dithiothreitol
E2F6	E2F transcription factor 6
E_{bnd}	Bond, angle and dihedral energy
ECD	Electron capture dissociation
E_{el}	Electrostatic energy
EHMT1	Euchromatic histone lysine methyltransferase 1
EHMT2	Euchromatic histone lysine methyltransferase 2
EPR	Electron paramagnetic resonance
ERCC6	DNA excision repair protein ERCC-6
ESBTL	Easy structural biology template library
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
EuHMT	Euchromatic histone methyltransferase
E_{vdw}	Van der Waals energy
EZH1	Histone-lysine N-methyltransferase EZH1
EZH2	Histone-lysine N-methyltransferase EZH2
F	Force acting on an atom
FDA	US Food and Drug Administration
FEP	Free energy perturbation
FEW	Free energy workflow
FF	Force field
FPLC	Fast protein liquid chromatography
G9a	Euchromatic histone-lysine N-methyltransferase 2
GAFF	General AMBER force field
GB	Generalised Born
G_l	Binding energy contribution of the ligand
GLP	Euchromatic histone-lysine N-methyltransferase 1
G_{np}	Non-polar contribution to the solvation free energy
G_p	Binding energy contribution of the receptor
G_{pl}	Binding energy contribution of the receptor-ligand complex
G_{pol}	Polar contribution to the solvation free energy
GPU	Graphics processing unit
H	Enthalpy
HCMV	Human cytomegalovirus
HDAC	Histone deacetylase
HDAC1	Histone deacetylase 1
HELA	Henrietta Lacks immortal cell line
Hey-SAM	(E)-hex-2-en-5-ynyl SAM
His-tag	Polyhistidine-tag
HKMT	Histone lysine methyltransferase

hmSILAC	Heavy methyl stable isotope labeling by amino acids in cell culture
HMT	Histone methyltransferase
HP1	Heterochromatin protein 1
HPLC	High-performance liquid chromatography
HTS	High throughput screening
I-BET/JQ1	Small-molecule inhibitors of BET bromodomains
I-SET	Insert SET domain
I-TASSER	Iterative threading assembly refinement
IA	Iodoacetamide
IPF1/PDX-1	Pancreatic and Duodenal Homeobox 1
IPTG	Isopropyl β-D-1-thiogalactopyranoside
JAK-STAT	Janus kinase/signal transducers and activators of transcription
k_{cat}	Turnover number
k_{cat}/K_m	Specificity constant
kDa	Kilodalton
KLF12	Krueppel-like factor 12
K_m	Michaelis constant
KMT buffer	Buffer for lysine methyltransferases
KMTs	Lysine methyltransferases
(KR)-(STA)K motif	Consensus motif for SET7/9-dependent methylation
L3MBTL1	Lethal (3) malignant brain tumor-like protein
LANCE	Homogeneous time-resolved fluorescence quenching assay
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography tandem-mass spectrometry
LEaP	Link, edit and parm
LIE	Linear interaction energy
MALDI-MS	Matrix-assisted laser desorption/ionization mass spectrometry
MAPK	Mitogen-activated protein kinase
MAT	Methionine adenosyltransferase
MAX	Max DNA binding protein
MBD	Methyl-CpG-binding domain proteins
MCPB	Metal center parameter builder
MD	Molecular dynamics
MDS	Myelodysplastic Syndrome
Me	Methyl
MGA	MAX Gene-Associated Protein
MLL	Histone-lysine N-methyltransferase 2A
MLL2	Histone-lysine N-methyltransferase 2D
MLL3	Lysine N-methyltransferase 2C
MLL4	Myeloid/lymphoid or mixed-lineage leukemia 4
MM	Molecular mechanics
MM-PBSA	Molecular mechanics Poisson-Boltzmann surface area
MoA	Mechanism of action

MPip-SAM	Methylpiperazine-SAM
mRNA	Messenger RNA
MS	Mass spectrometry
MSn	Multistage tandem mass spectrometry
MTAN	5'-Methylthioadenosine/S-adenosylhomocysteine (MTA/SAH) nucleosidase
MW	Molecular weight
MWCO	Molecular weight cut-off
MyoD	Myogenic differentiation protein 1
N-SET	N-terminal SET domain
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGPS	Néstor-Guillermo progeria syndrome
Ni-NTA	Nickel-nitrilotriacetic acid
NMR	Nuclear magnetic resonance
NPT	Isothermal–isobaric ensemble
NSD	Nuclear receptor SET domain-containing methyltransferases
NSD1	Nuclear receptor binding SET domain protein 1
NSD2	Nuclear receptor binding SET domain protein 2
NSD3	Nuclear receptor binding SET domain protein 3
NVE	Microcanonical ensemble
OD	Optical density
OD600	Optical density of a sample at a wavelength of 600 nm
p53	Tumor protein p53
Paf1	RNA polymerase II-associated factor 1 homolog
PBS	Phosphate-buffered saline
PCNA	Proliferating cell nuclear antigen
PCR	Polymerase chain reaction
PDB	Protein data bank
PES	Polyethersulfone
PHD	Plant homeodomain finger domain
Phe-SAM	Phenyl-SAM
PHF20	PHD finger protein 20
Pip-SAM	Piperidine-SAM
PKCζ	Protein kinase C, zeta
PMSF	Phenylmethanesulfonyl fluoride
PMTs	Protein methyltransferases
Post-SET	Post-SET domain
PPIs	Protein–protein interactions
PRC2	Polycomb repressive complex 2
PRDM2	PR domain zinc finger protein 2
Pre-SET	Pre-SET domain
PRMD9	PR domain zinc finger protein 9
PRMT	Protein arginine methyltransferase
PRMT1	Protein arginine methyltransferase 1

ProSeAM	Se-adenosyl-L-selenomethionine
PTM	Post-translational modification
PWM	Position weight matrix
$\mathbf{q}_i$	Partial charge
QM	Quantum mechanics
Rad6/Bre1	E2 ubiquitin-conjugating enzyme / E3 ubiquitin ligase pathway
RELA	Transcription factor p65
RESP	Restrained electrostatic potential
$\mathbf{r}_i$	Position vector
RMSD	Root mean square deviation
Rmt1	Protein arginine N-methyltransferase 1
RNA	Ribonucleic acid
RNAi	RNA interference
RT	Room temperature
RT-PCR	Reverse transcription polymerase chain reaction
S	Entropy
SAAMs	S-alkyl analogues of methionine
SAH	S-Adenosyl-L-homocysteine
SAM	S-Adenosyl methionine
SC	Side chain
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SET	Su(var)3-9, Enhancer-of-zeste and Trithorax domain
SET7/9	Histone-lysine N-methyltransferase SETD7
SETD1A	Histone-lysine N-methyltransferase SETD1A
SETD1B	Histone-lysine N-methyltransferase SETD1B
SETD2	Histone-lysine N-methyltransferase SETD2
SETD6	N-lysine methyltransferase SETD6
SETD7	Histone-lysine N-methyltransferase SETD7
SETD8	N-lysine methyltransferase KMT5A
SETDB1	Histone-lysine N-methyltransferase SETDB1
SETMAR	Histone-lysine N-methyltransferase SETMAR
SGC	Structural Genomics Consortium
SHP	Small heterodimer partner
shRNA	Short hairpin RNA
SILAC	Stable isotope labelling by amino acids in cell culture
siRNA	Small interfering RNA
SLE	Systemic lupus erythematosus
SMYD2	N-lysine methyltransferase SMYD2
SMYD3	Histone-lysine N-methyltransferase SMYD3
SOB	Super optimal broth
SOC	Super optimal broth with catabolite repression
SPA	Scintillation proximity assay
SRA	SET- and RING-associated domain

Suv39h	Subgroup of SET domain-containing molecules
SUV39H1	Histone-lysine N-methyltransferase SUV39H1
SUV39H2	Histone-lysine N-methyltransferase SUV39H2
SUV420H1	Histone-lysine N-methyltransferase KMT5B
SUV420H2	Histone-lysine N-methyltransferase KMT5C
T	Temperature
TI	Thermodynamic integration
$\mathbf{t_o}$	Initial time
TP53BP1	Tumor suppressor p53-binding protein 1
TPhe-SAM	Thiophenyl-SAM
TR-FRET	Time resolved fluorescence resonance energy transfer
U	Potential energy function
$\mathbf{U_{elec}}$	Boltzmann electrostatic interaction energy
UNC0638	G9a inhibitor UNC0638
UNC0642	G9a and GLP inhibitor UNC0642
UPLC-MS/MS	Ultrahigh-performance liquid chromatography tandem mass spectrometry
UV	Ultraviolet
$\mathbf{U_{vdw}}$	Boltzmann van der Waals interaction energy
vi	Velocity vector
vSET	Viral SET domain protein
Wiz	Protein Wiz
WNT	Wingless-related integration site
XL-MS	Protein cross-linking coupled with mass spectrometry
$\mathbf{\Delta E_{el}}$	Electrostatic energy change
$\mathbf{\Delta E_{np}}$	Non-polar solvation energy change
$\mathbf{\Delta E_{pol}}$	Polar solvation energy change
$\mathbf{\Delta E_{vdw}}$	Van der Waals energy change
$\mathbf{\Delta G^0_{Bind}}$	Binding free energy of a ligand
$\mathbf{\Delta G^0_{TI,\ bound}}$	Variation in free energy when transforming ligands in the bound state.
$\mathbf{\Delta G^0_{TI,\ solvated}}$	Variation in free energy when transforming ligands in the solvated state
$\mathbf{\Delta G_{bind}}$	Total binding energy
$\mathbf{\Delta G_{gas}}$	Total gas phase free energy
$\mathbf{\Delta G_{solv}}$	Total solvation free energy
$\mathbf{\Delta\Delta G^0_{Bind}}$	Relative binding free energy between two systems
$\mathbf{\Delta\Delta G_{MMPBSA}}$	Difference of binding energies compared to SAM
$\boldsymbol{\phi}$	Angle of right-handed rotation around N-CA bond
$\boldsymbol{\chi}$	Side chain dihedral angle
$\boldsymbol{\psi}$	Angle of right-handed rotation around CA-C bond
$\boldsymbol{\omega}$	Angle of right-handed rotation about C-N bond

Chapter 1

Introduction to epigenetics

As opposed to the concept of genetics, the branch of science studying DNA heredity and subsequent phenotypic modifications, the definition of epigenetics is still evolving and has done so since it was first coined by Conrad Waddington more than half a century ago as "the branch of biology which studies the causal interactions between genes and their products, which brings the phenotype into being" [1]. The definition of epigenetics has evolved in parallel with the knowledge gained about the molecular mechanisms regulating gene expression in eukaryotes. The fact that the DNA present in all somatic cells of an organism is the same but gene expression patterns differ greatly between different cell types led to the most commonly accepted definition of epigenetics, being: "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [2]. More recently, an updated definition of epigenetics was proposed by Berger et al.: "an epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" [3]. Epigenetic modifications are crucial in the mechanism of cell differentiation [4] as these modifications explain why cells evolve into different phenotypes even when their underlying genome is still the same [5]. For a further understanding of this concept we will focus in establishing the molecular basis of epigenetics.

1.1. The nucleosome and chromatin

It is widely known that eukaryotic cells store DNA in the form of chromosomes within their nucleus. These chromosomes are not only made up of DNA, but also contain proteins and RNA. Chromatin can be defined as the complex of DNA, RNA and proteins that provide maintenance and regulatory functions to the dynamic DNA structure [6]. A set of these chromosomal proteins regulate the expression of the genes contained in the DNA. Histones are basic and hydrophilic proteins that give structural support to the DNA. The nucleosome is the basic structure and repeating unit of chromatin, consisting of a segment of approximately 147 base pairs of DNA wrapped around a histone octamer [7]. The DNA chain is wrapped 1.65 turns around the octamer, establishing 14 contact points with it [8]. The basic structure of the nucleosome is shown at Figure 1-1.

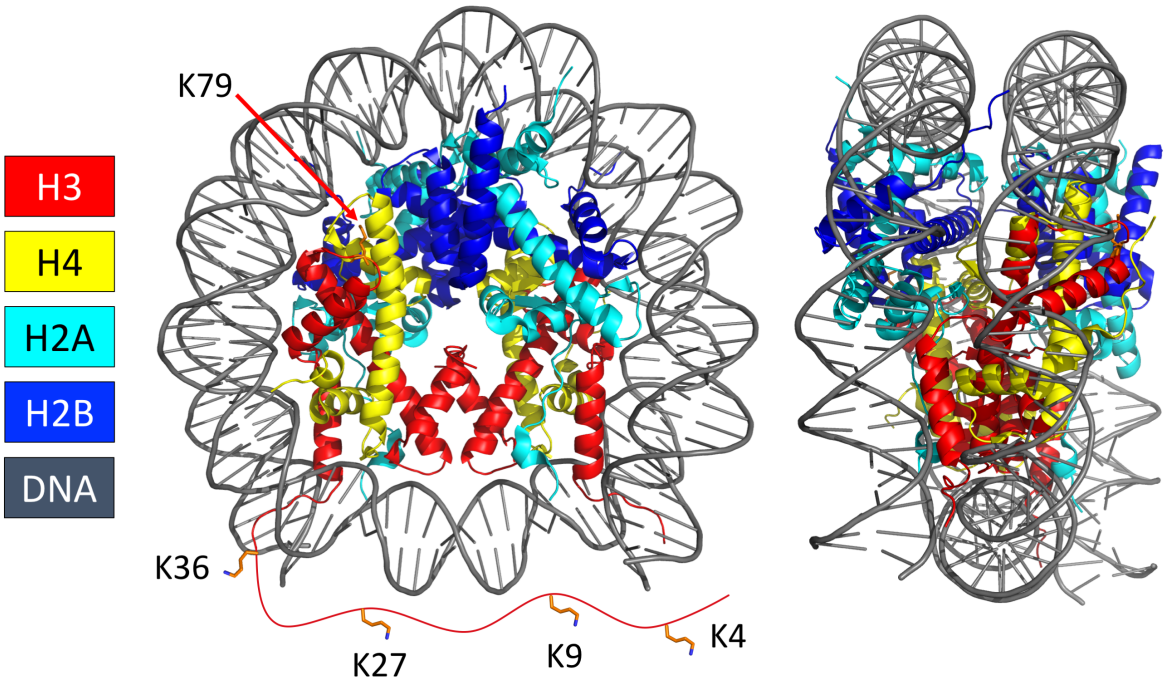


Figure 1-1. Basic structure of the nucleosome (PDB accession code: 5AV6). Histone H3 is depicted in red, H4 in yellow, H2A in cyan, H2B in blue and DNA in grey. Histone H3 lysines K4, K9, K27 and K36 (located at the histone H3 tails) and K79 (located at the core region) are also represented as orange sticks.

Chromatin structure is crucial for the correct regulation of the cell. Without this organised structure, long DNA structures would be unable to fit within the compact cell nucleus [9]. Each octamer is composed by two copies of each of the following histone proteins: H3, H4, H2A and H2B. H3 and H4 appear as a H3-H4 tetramer, whereas H2A and H2B appear as dimers.

In addition to these proteins, the linker histone H1 binds the entry and exit points of the DNA, which ensures that the DNA is retained forming the higher order structures of chromatin [10]. Histone H1 is also important because it joins together the different nucleosomes to create the 30-nm chromatin fibre solenoid structure that increases the DNA packing ratio [11]. The different chromatin structure levels from the DNA chain to the chromosome with the approximate width of each state can be seen at Figure 1-2.

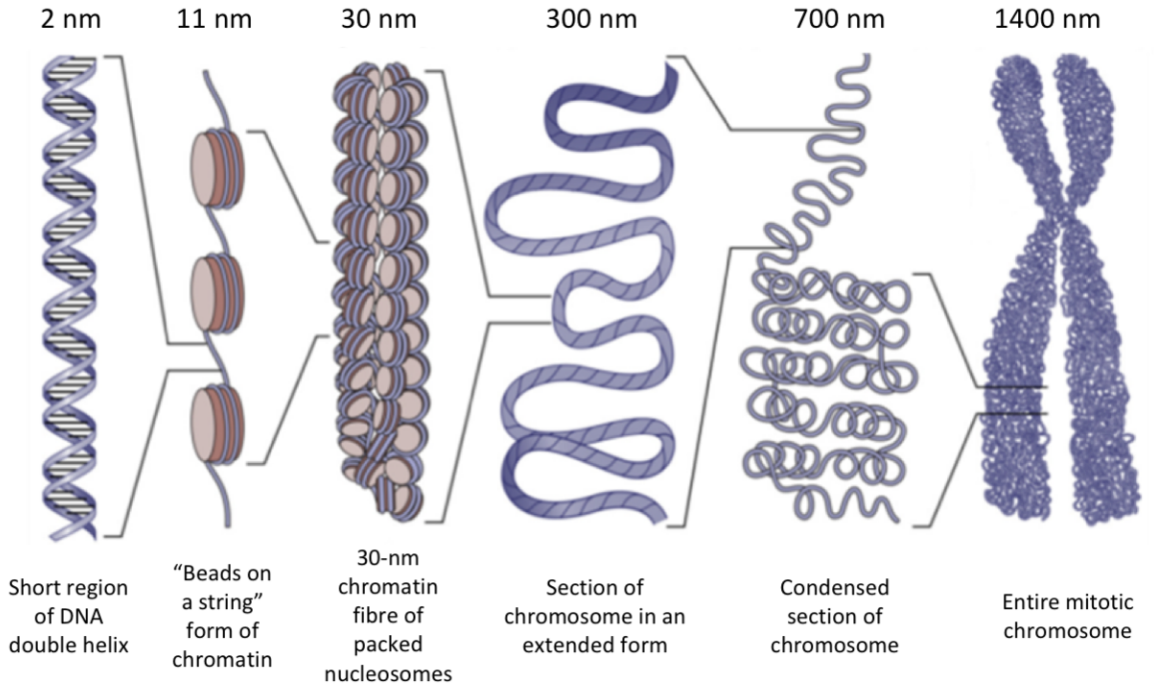


Figure 1-2. Chromatin structure. Adapted from Jansen et al. [12].

For transcription to occur, chromatin must be in an open configuration that enables access to RNA polymerases, transcription factors and other regulatory proteins [13]. According to their transcriptional states, chromatin can be classified in two different groups: euchromatin, a slightly agglomerated form of chromatin which usually undergoes active transcription; and heterochromatin, a more tightly packed form of chromatin that is transcriptionally inactive [14]. Heterochromatic chromatin is so densely packaged that DNA fragments are sterically confined. In this way, many protein factors associated with transcription are not able to access the DNA fragments and their DNA remains unexpressed. Therefore, chromatin appears as a dynamic structure which not only offers a structural function but it is also important in the regulation of transcription [15]. The protein complexes that modify the location of DNA around the nucleosomes are called nucleosome remodeling factors [16].

1.2. Epigenetic control of gene expression and its effect on disease

Epigenetic mechanisms work as an additional layer of regulation than solely DNA inheritance, and additionally stabilise gene expression programmes that lead to different cell types. The development of the epigenetics field is tied to advancements in techniques such as chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) that have allowed for the study of epigenetic marks in both healthy and diseased cells and tissues [17]. Epigenetic marks show a special clinical interest due to their reversible nature, as opposed to genetic changes (mutations). Epigenetic modifications can be categorised into DNA methylation and histone modifications. There are at least four distinct DNA modifications and 16 types of histone modifications [18], which are crucial for the regulation of every DNA-based process, such as replication, DNA repair, condensation and transcription.

1.2.1. DNA methylation

DNA methylation occurs at cytosine nucleotides and is the most studied epigenetic modification in humans. This modification occurs mainly at CpG dinucleotides that are found principally in the named CpG islands, formally described by Gardiner-Garden and Frommer as sections of the DNA with more than 200 bases with a content of G+C of at least 50% and with a ratio of CpG frequencies of at least 0.6 [19]. The methylation of DNA at the CpG islands is commonly related to gene silencing [20]; this is well seen, for example, in the genomic imprinting of the genes, where the hypermethylation of one of the two parental alleles leads to the expression of only the other [21]. A simplified diagram describing the function of the CpG island methylation is shown in Figure 1-3.

DNA methylation is not exclusive from CpG islands but also occurs in regions that are close (~2 kb) to them, called CpG island shores, being strongly related with transcriptional inactivation [22]. Besides, cytosine nucleotides methylation can also be found in CHG and CHH sites [23], where H can be adenine (A), cytosine (C) or thymine (T). DNA methylation can also be related to transcriptional activations occurring at gene bodies, preventing transcription at incorrect sites. If gene bodies are demethylated, transcription is initiated at

erroneous sites, leading to disease [24]. DNA methylation also acts as a key factor in preserving chromosomal stability by hypermethylation of CpGs found in repetitive sequences [25].

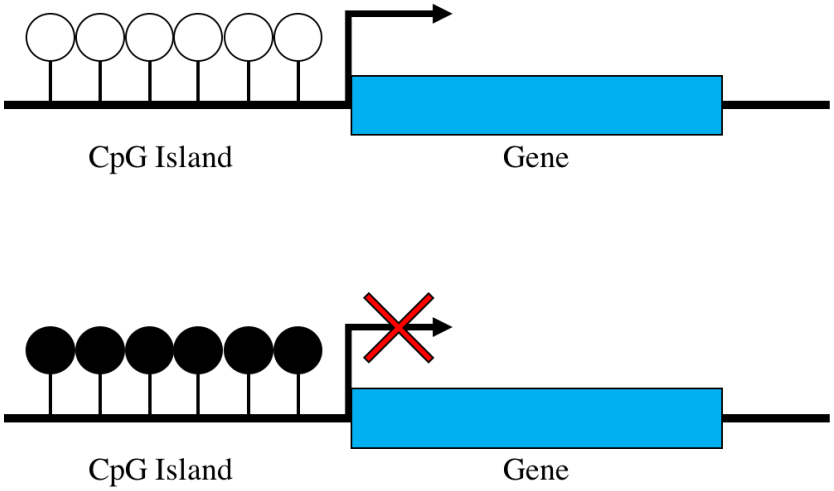


Figure 1-3. Role of CpG islands in transcriptional regulation. When DNA is not methylated (white circles) on promoter CpG islands, transcription occurs. However, when they are methylated (black circles), transcription is blocked and the gene is silenced.

The mechanisms of DNA methylation are varied; methylated DNA can promote the binding of methyl-CpG-binding domain (MBD) proteins, that are themselves able to recruit other epigenetic modifiers such as histone modifiers or chromatin remodelling complexes [26]. Methylated DNA can also act by preventing the binding of transcription factors to their target sites [27]. As opposed to methylated DNA, the unmethylated CpG islands promote gene expression by recruiting Cfp1, a protein that associates with the histone methyltransferase SETD1A which catalyses histone H3K4 trimethylation, associated with active, uncondensed euchromatin [28].

DNA methylation is regulated by the enzymes of the DNMT family. These enzymes catalyse the transfer of a methyl group from S-adenosyl methionine to DNA. Among these, DNMT1 is the most abundant in the cell and it is mostly transcribed during the S phase of the cell cycle, when DNA is replicated. Its main function is to perform the methylation of hemimethylated sites [29].

Cancer cells present an important global loss of DNA methylation in 5-methyl cytosines [30]. The global activating hypomethylation promotes chromosomal instability, translocations and gene disruption among other outcomes [31]. At the same time, repressing hypermethylation is found at CpG islands of specific promoters. These hypermethylations are tumour specific and the reason why they happen in some CpG islands and not others is unclear [22]. It has been

found that most of the aberrant DNA methylation in cancer occurs at the CpG island shores, resulting in transcriptional repression [32]. Other diseases related to DNA methylation changes are neurological, such as Alzheimer's, where CpG island repressing hypermethylation occurs [33], and multiple sclerosis, where CpG island activating hypomethylation occurs [34]. Certain autoimmune diseases such as rheumathoid arthritis, showing both CpG island repressing hypermethylation and activating hypomethylation sites, and systemic lupus erythematosus (SLE), presenting CpG island activating hypomethylation, also appear related to DNA methylation events [35].

1.2.2. Histone post-translational modifications

The histones which constitute the octamer in the nucleosome are formed by a globular domain and a histone tail. Both domains undergo post-translational modifications (PTMs). Many of the histone residues exposed to PTMs are in the tail domains [36,37]. PTMs affect chromatin by breaking inner contacts, changing the higher order structure in which chromatin is organised, altering the noncovalent interactions inside and in between nucleosomes; PTMs might also act by altering the binding of chromatin modifying enzymes to chromatin, which regulate the manipulation of DNA and histones [5]. The "histone code" is the hypothesis that states that the transcription of the genetic material is partially regulated by multiple PTMs on the nucleosomal histones. Both DNA methylation and histone PTMs constitute the epigenetic code. An essential concept of the histone code is that the histone modifications recruit effector proteins through specific protein domains that are able to recognise these marks, which modify the chromatin structure to activate or stop transcription [9,38,39].

The different types of histone modifications including the affected residues and nomenclature in parentheses are : acetylation (K-ac), methylation (K-me1, K-me2, K-me3, R-me1, R-me2s-symmetrical (each of the two available nitrogen atoms have a methyl group), R-me2a-asymmetrical (both methyl groups are bound to the same nitrogen atom)), phosphorylation (S-ph,T-ph,Y-ph), ubiquitylation (K-ub), sumoylation (K-su), ADP ribosylation (E-ar), deimination (R>Cit), proline isomerisation (P-cis <>P-trans), crotonylation (K-cr), propionylation (K-pr), butyrylation (K-bu), formylation (K-fo), hydroxylation (Y-oh) and O-GlcNAcylation (S-GlcNAc, T-GlcNAc) [18].

According to Tan [37], there are at least 130 PTM sites on human histones, and the function of many of these remain understudied. Further investigation about these sites could help understanding the vast diversity of biological processes in higher eukaryotes. Figure 1-4 shows the sites for some of the more abundant and widely studied PTMs on histones.

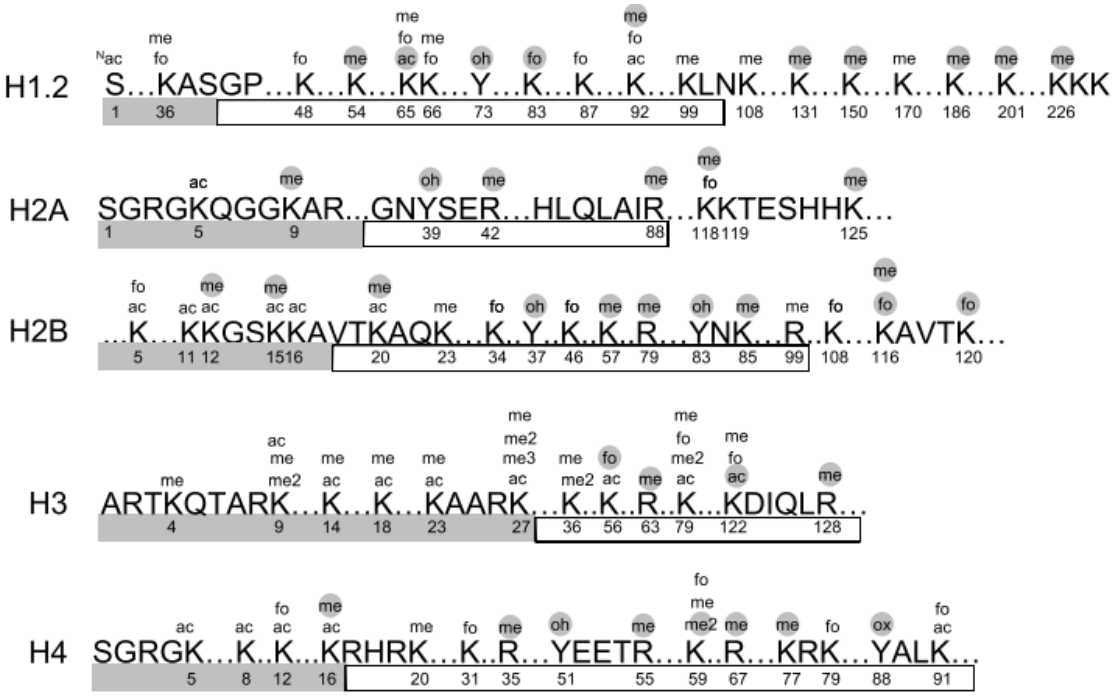


Figure 1-4. Main histone PTM sites. Me: monomethylation; me2: dimethylation; me3: trimethylation; ac: acetylation; fo: formylation; oh: hydroxylation [37].

Histone modifications are involved in DNA repair, DNA replication, alternative splicing, chromosome condensation and transcriptional regulation [40]. Histone modifications are controlled by epigenetic regulators that can be classified as "erasers", which remove the marks and prepare the histone for other modifications; "writers", which catalyse the transfer of the PTMs to histones or DNA; and "readers", which have specialised domains that bind these PTMs and carry some chromatin-modifying activity. Figure 1-5 shows some examples of each of these epigenetic regulators.

One of the main reasons why it is interesting to study histone modifications is because of their critical role in cancer. Cancer cells consistently show a global decrease of monoacetylated H4K16 mediated by HDACs that have been found to be overexpressed or mutated depending on the tumour type [41].

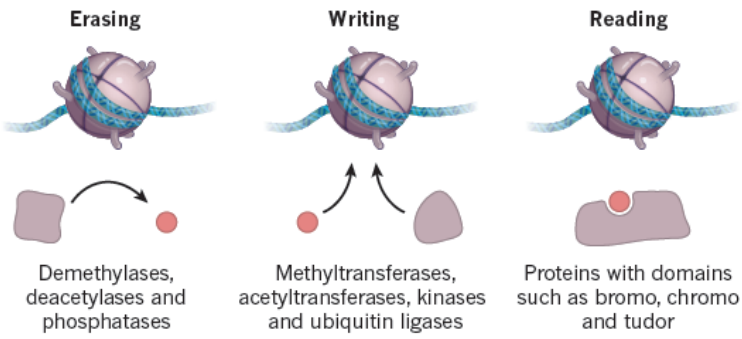


Figure 1-5. Types of epigenetic regulators and examples [42].

Besides acetylation, histone methylation seems to play a major role in cancer as well, with a global loss of the activating mark H3K4me3 and the repressive H3K20me3 due to the aberrant expression of histone methyltransferases or demethylases [41]. Some examples of methyltransferases involved in cancer are DOT1L, the main H3K79 methyltransferase, that induces an euchromatic state enabling the expression of tumour suppressor genes [43], and MLL, whose presence in leukaemia leads to aberrant H3K79 and H3K4 methylation patterns [44].

Histone modifications also play a role on neurological disorders such as histone hypoacetylation in amyotrophic lateral sclerosis (ALS) [45]. Besides, histone H3K9 hypertrimethylation has been found to be related to Huntington's disease [46]. Histone modifications also seem to be involved in autoimmune diseases; for example, in type 1 diabetes lymphocytes but not monocytes have increased H3K9me2 methylation states [47]. Epigenetic modifications such as DNA methylation and post-translational modifications of histones have also recently appeared to be one of the most conserved hallmarks in aging; cellular reprogramming to pluripotency has been achieved *in vitro*, this showing promise of ameliorating age-related phenotypes by epigenetic cellular reprogramming, still not demonstrated at the organismal level [48].

Research in model organisms such as *Neurospora crassa* and *Arabidopsis thaliana* has demonstrated an extensive cross-talk between DNA methylation and histone modifications [49,50]. The recognition of methylation sites by histone methylation readers (PHDs, chromodomains and BAH domains) and readers of DNA methylation such as SRA, CXXC and methyl-CpG-binding domain (MBD) [51] are necessary elements for this cross-talk. For example, non-methylated DNA in CpG islands promotes the formation of H3K4 and H3K27 trimethylation marks and excludes H3K36 methylation [28].

Due to their reversible nature, epigenetic marks are attractive targets for the pharmaceutical industry. The US Food and Drug Administration (FDA) approved the first commercially available epigenetic drugs decitabine (DNA methyltransferase inhibitor) and vorinostat (histone deacetylase inhibitor) for the treatment of human cancers in 2006 [52]. These epigenetic inhibitors are effective in the selective elimination of cancer cells due to a more fragile chromatin structure compared to healthy cells when combined to other techniques such as radiation therapy [53]. As noted by Jones, there is a general trend in the pharmaceutical industry towards epigenetic therapies, with above 30 different epigenetic drugs currently in clinical trials by companies such as Novartis, Merck & Co, GSK or Bayer Corporation [54], with promise to combine epigenetic drugs with other therapies for approaches in personalised medicine.

Chapter 2

Review of histone methylation

2.1. Histone methylation

Histone methylation is one of the most important post-translational modifications. It affects crucial biological processes such as cell-cycle regulation, DNA damage, stress response, development and differentiation [55], and it is one of the most extensively studied epigenetic modifications. The methylation of histones takes place on the side chains of lysine, arginine and histidine (to a lower extent) residues. Unlike other common PTMs such as phosphorylation and acetylation, methylation does not alter the overall charge of the histone. Lysines can be mono-, di- or tri- methylated [56], arginine can be monomethylated or dimethylated, symmetrically or asymmetrically [57], and histidines appear monomethylated [55]. The focus of this section will be lysine methylation, as it is the best characterised and the focus of this work.

Even though methylation occurs on a variety of histone residues, the best known are those occurring on histone H3 at the positions K4, K9, K27, K36, K79, and on the histone H4 at K20 position. Some of these histone lysine methylations are related with active euchromatic genes (H3K4, H3K36, H3K79), whereas others are associated with non-active heterochromatic regions (H3K9, H3K27, H4K20) [18]. Furthermore, the state of chromatin is

not only related with the residue position but also with the degree of methylation, as seen in Table 2-1.

Table 2-1. Sites of histone 3 lysine methylation and their relationship with activating or repressing chromatin according to their methylation state [58-61].

Site of lysine methylation	Activation	Repression
H3K4	me1, me2, me3	-
H3K9	me1	me2, me3
H3K27	me1	me2, me3
H3K36	me1, me3	-
H3K79	me3	-

As shown in Table 2-1, the effect on chromatin organisation depends on the methylation state of histone H3. In order to understand the underlying processes that lead to the different degrees of methylation, it is important to characterise the specific enzyme(s) catalysing the transfer of a methyl mark to a histone residue in order to understand the undergoing processes leading to a specific methylation event. These enzymes (called enzyme methyltransferases) form a complex group of chromatin regulators.

2.2. Histone lysine methyltransferases

Histone lysine methyltransferases (HKMTs) are enzymes catalysing the transfer of a methyl donor group from the cofactor S-Adenosyl methionine (SAM) to a histone lysine target, producing a methylated histone lysine and S-Adenosyl homocysteine (SAH). A simplified mechanism is shown in Figure 2-1.

All methyltransferases use the same cofactor (SAM) to perform the methylation, hence the challenge in detecting the individual activity of each methyltransferase. To date, three families of these enzymes have been described: Su(var)3-9 Enhancer-of-zeste and Trithorax (SET) domain-containing lysine methyltransferases (where the catalysis takes place in a SET domain), non-SET domain-containing lysine methyltransferases and arginine

methyltransferases. The only lysine methyltransferases not containing SET domain are the DOT1-like proteins [55].

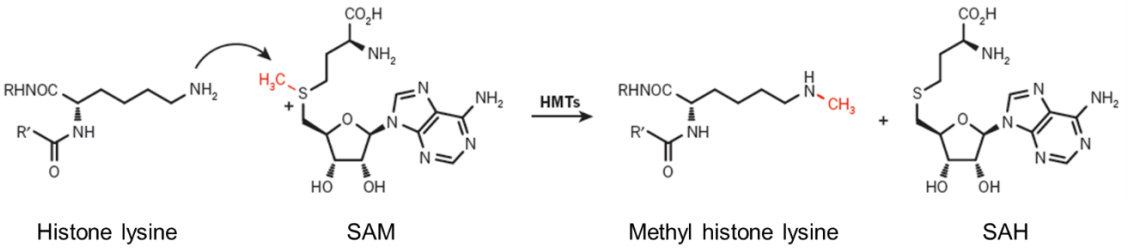


Figure 2-1. Mechanism of histone lysine methylation by lysine methyltransferases [42].

The histone lysine methylation mechanism requires prior nitrogen deprotonation to produce a lone pair of electrons prior nucleophilic attack [62]. This deprotonation is essential, as the lone pair of electrons (nucleophile of the reaction) is only available in the basic form of the target lysine, the site where the methyl group is transferred from SAM. In addition, for HKMTs that di- and trimethylate lysine substrates, the multiple rounds of methylation proceed processively (catalysing consecutive reactions without releasing the histone substrate), such that after each round of methylation, the lysine side chain becomes positively charged again and needs to be deprotonated for the next round of methylation [63]. The methylation mechanism is shown in Figure 2-2.

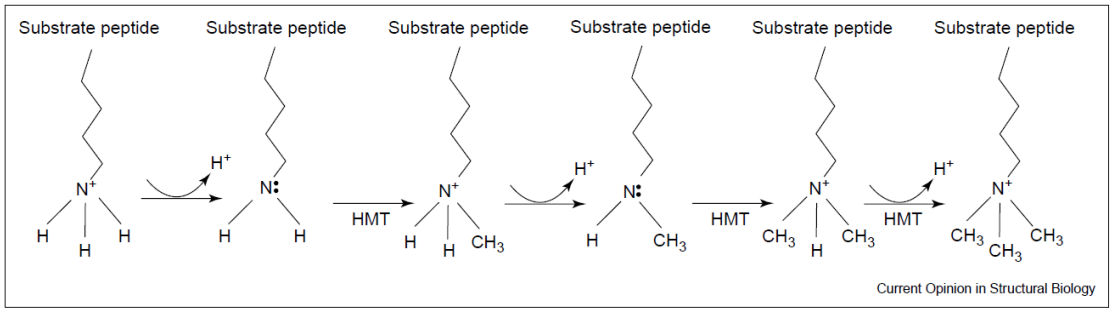


Figure 2-2. Chemical mechanism for the methylation of a histone lysine [64].

SET domain-containing HKMTs can specifically mono-, di- or trimethylate their substrate lysine. For example, G9a can only mono- and dimethylate a substrate lysine due to the presence of the Y1067 residue, which positions the dimethylamine in a conformation that inhibits the transfer of a third methyl group to the substrate lysine; this was verified by experiments with G9a Y1067F mutant showing that it can trimethylate histone H3K9. In addition, if residue F1152 is mutated to tyrosine, the mutant G9a can only monomethylate H3K9 [65]. In a similar way, a Y305F mutation in SETD7 converts it from a

monomethyltransferase to a dimethyltransferase [62]. Therefore, methylation state specificity seems to be inversely proportional to the number of tyrosine residues located in the volume surrounding the methyl accepting nitrogen of the histone lysine.

Table 2-2 shows some of the most well-known HKMTs and their related histone methylation targets. This diversity of methylation targets depending on the HKMT is due to differences in the residues present in the substrate recognition domains of the HKMT and the hydrogen bonds formed between these and the different histone residues [66].

Table 2-2. Histone lysine methyltransferases according to their target residues within histones H1, H3 and H4 and the degree of methylation they catalyse [55,67,68].

	me3	me2	me1
H1.2K187		G9a GLP	G9a GLP
H1.4K26		G9a GLP	G9a GLP
H3K4	SETD1A SETD1B ASH1L MLL MLL2 MLL3 MLL4 SMYD3 PRMD9	SETD1A SETD1B MLL MLL2 MLL3 MLL4 SMYD3	SETD1A SETD1B ASH1L§ MLL MLL2 MLL3 MLL4 SETD7
H3K9	SUV39H1 SUV39H2 SETDB1 PRDM2	G9a SUV39H1 SUV39H2 SETDB1 GLP PRDM2	G9a SETDB1 GLP PRDM2
H3K27	EZH2 EZH1	G9a EZH2 EZH1	G9a
H3K36	SETD2	NSD3 NSD2 NSD1 SMYD2 SETD2	SETD2 NSD3 NSD2 NSD1
H3K79	DOT1L	DOT1L	DOT1L
H4K20	SUV420H1 SUV420H2	SUV420H1 SUV420H2	SETD8

The following sections will characterise the relationship between methyltransferase structure and function by focusing on three case studies relevant to this thesis: G9a, SETD7 and DOT1L. The evolutionary relationship of these enzymes with other methyltransferases is shown in Figure 2-3.

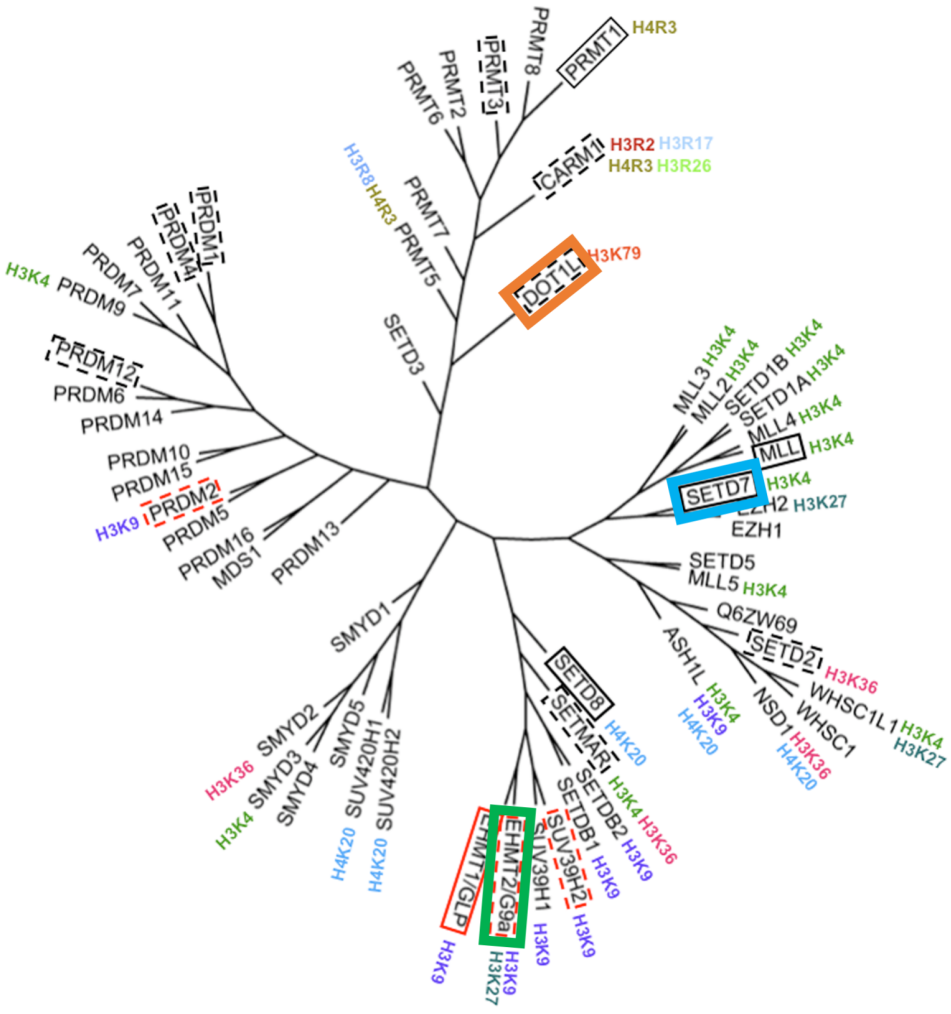


Figure 2-3. Phylogenetic tree of the human HKMTs [65]. G9a is highlighted in green, SETD7 in blue and DOT1L in orange. The phylogeny is based on a multiple sequence alignment of the methyltransferase domain including the N-SET, Pre_SET, SET, I-SET, and Post-SET motifs for the SET-domain containing HKMTs.

2.2.1. G9a

G9a (also named EHMT2) is a member of Suv39h subgroup of the SET domain-containing enzymes and part of the subfamily of H3K9 methyltransferases. Together with GLP (also known as EHMT1), they are the most important HKMTs to methylate and dimethylate the H3K9 position on histone H3 [69]. G9a is comprised of 1210 amino acids with an atomic mass

of 132,370 Da [70]. It shares with other Suv39 related proteins, such as GLP and SUV39H2, a common fold (or arrangement of the secondary structure elements of the protein relative to each other in space) which is composed of a conserved SET domain (residues 1036-1163 for G9a) and a variable insert SET domain (I-SET), which are bordered by a Pre- (residues 972-1035 for G9a) and a Post-SET (residues 1164-1180 for G9a) domain. The I-SET can be defined as a "rigid peptide docking platform" for the histone H3 tail [65], composed of a helix followed by a 2-stranded β-sheet connected by loops. One of these strands establishes 2 evolutionarily conserved hydrogen bonds with the H3K9 substrate, which might be responsible for the lysine specificity and orientation for proper methylation, while also providing interactions with other neighbouring histone residues [65]. The mobile post-SET domain of G9a consists of a partially folded α-helix that seals around the histone substrate, contributing to the peptide binding, including also a ZnCys motif. The post-SET domain has three conformations: flexible state (when no co-factor or peptide is bound), loose state (when only SAM or SAH is bound) or rigid state (when both cofactor and substrate are bound, allowing the I-SET domain to close onto the substrate) [65]. These domains are shown graphically for G9a in Figure 2-4A.

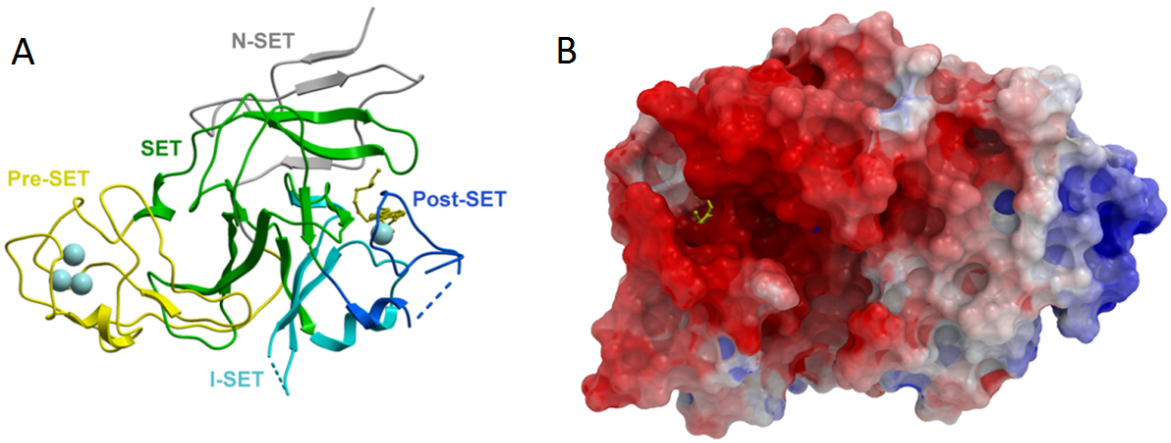


Figure 2-4. A) G9a domains. SAH is represented with yellow sticks and the Zn^{2+} ions are represented as cyan spheres. B) G9a electrostatic potential (+5 kcal/electron units = blue, -5 kcal/electron units = red), with SAH represented with yellow sticks. Adapted from Wu et al. [65].

Other special features that characterise these enzymes are: the presence of a pseudoknot secondary structure alongside the catalytic site, the fact that cofactor (e.g. SAM) and substrate (e.g. histone residue H3K9) binding areas join together at the methylation site, and the presence of a narrow and highly selective lysine docking hydrophobic channel [65,71]. There

is also an N-SET region that folds around the core SET domain and is located N-terminal to the Pre-SET [72]. In addition, both GLP and G9a share a common domain called ankyrin repeat (residues 730-965 for G9a) which acts as methyl-lysine binding module, binding preferentially to H3K9me1 or H3K9me2 peptides. With this helix-turn-helix-β-turn structure, both G9a and GLP can identify the methylation state of the histone. The basis for recognition of the methylated lysine has been analysed for GLP, finding that the H3K9me2 residue is bound in a partial hydrophobic cage formed by three tryptophans (Trp839, Trp844 and Trp877) and one acidic residue (Glu847)[73]. The basis for this substrate recognition is shown in Figure 2-5.

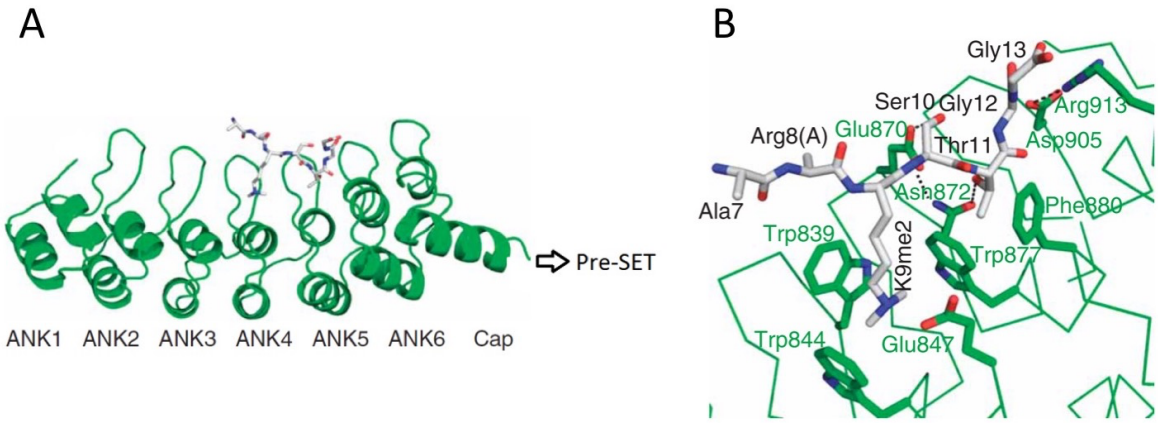


Figure 2-5. A) The helix-turn-helix-β-turn structure of the ankyrin repeats in GLP. B) H3K9me2 bound to GLP in a partial hydrophobic cage. GLP is represented in green, while the histone H3 peptide is represented as grey sticks [73].

The H3K9 lysine is located inside a channel that is formed by the I-SET and the Post-SET domains, which provide interactions with the histone by both backbone and side-chain forces. Factors such as the shape and the width of this channel are key to increase the selectivity for a specific substrate, in this case H3K9, and also to determine the number of methyl groups that are transferred from SAM to the histone. Once the first methylation has taken place, the channel geometry restricts the rotation around the H3K9 C-N side chain bond. If the rotation cannot take place, the histone lysine is hidden from other SAM molecules and a second methylation is restricted [64].

For G9a, there are two reported residues that are critical for catalysis: tyrosine Y1154, whose hydroxyl group is vital to maintain catalytic activity, and tyrosine Y1067, which establishes hydrogen bonds with the lysine methyl nitrogen acceptor, thus preserving the orientation of dimethyl-amine from receiving further methyl groups, thereby avoiding a trimethylation

state [40]. In general, the greater the number of tyrosines around the methyl-accepting nitrogen, the lower the methylation state. The histone lysine hydrophobic channel of G9a is essentially electronegative to ensure attraction to electropositive histone peptides (abundant in lysines and arginines) via long range electrostatic interactions. G9a electrostatic potential is shown in Figure 2-4B.

The SET domain of GLP shares 80% of its sequence with the SET domain of G9a [74]. Not only are these the most important enzymes for the mono- and dimethylation of H3K9, but also exist principally as a G9a-GLP heterodimer which appears to be the functional H3K9 methyltransferase *in vivo* [75]. The formation of the G9a-GLP complex depends on their SET domains. Additionally, both G9a and GLP can create homodimers [76]. A possible explanation for the predominance of the G9a-GLP heterodimer was proposed by Shinkai et al. [75], based on the fact that G9a also interacts with the multi-zinc-finger-containing molecule Wiz. Wiz recognises and binds the homodimer or heterodimer structure of G9a and GLP SET domains, but most stably interacts with the G9a–GLP heterodimer. This fact provides a mechanistic explanation about why the G9a–GLP heterodimer is the dominant intracellular form.

Another important interaction of the G9a/GLP heterodimer is with the Polycomb Repressive Complex 2 (PRC2), one of the main complexes associated with epigenetic silencing through methylation of the H3K27 position. PRC2 interacts with G9a/GLP both physically and functionally, and the activity of G9a is related to the PRC2 recruitment to a certain number of its target genes. This interaction has been observed both *in vitro* and in cells [77]. Histone lysine methyltransferases Suv39h1, G9a, GLP and SETDB1 can be associated in a multimeric protein complex, determined by eliminating the presence of Suv39h1 or G9a in cells, resulting in the destabilisation of the rest of HKMTs. Fritsch et al. [78] proposed that it was also possible that not all the HKMTs of this complex perform their activities within it, but their binding is instead associated with the recruitment of other proteins. Other molecules found to be associated with G9a were the small heterodimer partner (SHP) and DNA methyltransferase-1 (Dnmt1), but many other relations remain elusive due to the limitations in the techniques used for this analysis [79,80].

2.2.2. SETD7

SETD7 is a SET domain-containing histone lysine mono-methyltransferase of 366 residues in length whose main function is the monomethylation of histone H3K4 [81]. In a similar way to G9a, SAM and the peptide substrate are situated on opposite faces of the SET domain, with the histone peptide passing through a narrow channel in order to reach the cofactor binding surface, where the amine group of H3K4 can accept the methyl donor group of SAM [62]. The C-terminal segment of SETD7 (residues 337-366) form a β-hairpin followed by an α-helix (based on PDB: 1O9S).

Residues Y335 and Y337 are important in the formation of the binding channel. The lysine amine is forming hydrogen bonds with the hydroxyl group of Y245 and to a water molecule that itself hydrogen-bonds to the hydroxyl group of Y305 [40]. Once the SAM molecule is properly aligned, and the methyl group of the cofactor is in close proximity with the de-protonated nitrogen of the substrate lysine, a nucleophilic attack occurs, resulting in the methylation of the H3K4 lysine and the appearance of SAH as a by-product [82]. The catalytic mechanism is shown in Figure 2-6.

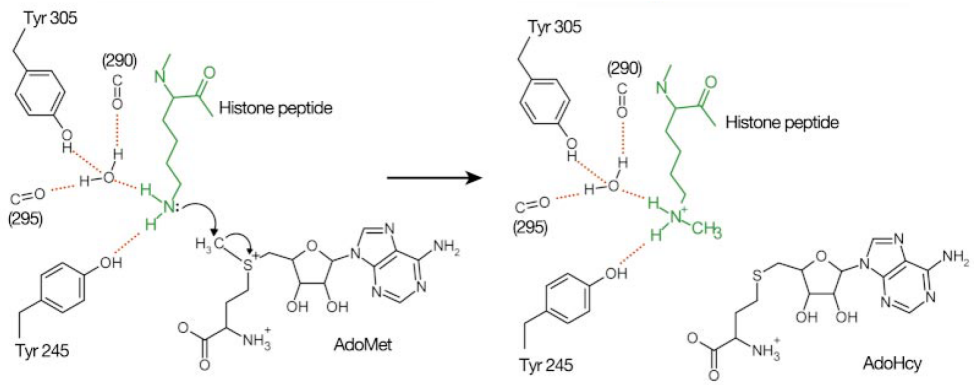


Figure 2-6. Schematic catalytic mechanism of SETD7 [40].

The reason why SETD7 is a monomethylase can be found in its crystallographic structure: both the Y245 and the water molecule make favourable interactions that stabilise the rotamer after monomethylation, and do not enable a second methylation to take place. In G9a, Y305 is substituted by a proline (P1121), generating a cavity that can accommodate the dimethyl modification [40]. The structure of SETD7 compared to that of G9a and the most important residues in both are shown in Figure 2-7.

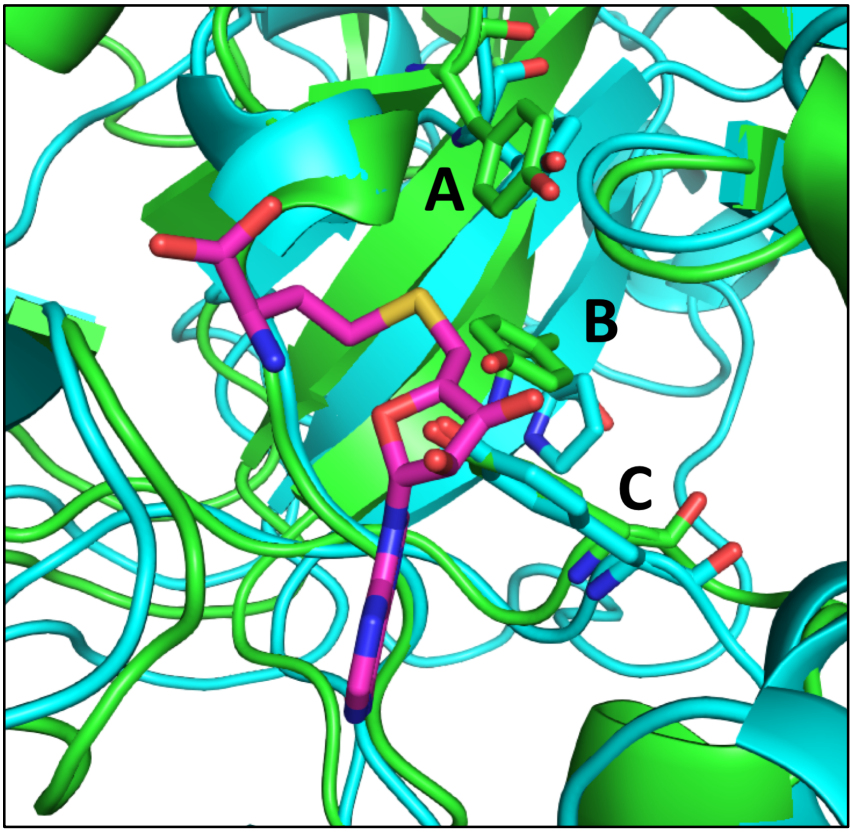


Figure 2-7. Comparison of the aligned structures of SETD7 (green, PDB:1XQH) and G9a (cyan, PDB: 2O8J) and the location of their crucial residues. The structure of SAH bound to SETD7 is shown as a reference (magenta sticks). A: SETD7's Y245 and G9a's Y1067. B: SETD7's Y305 and G9a's P1121. A: SETD7's Y335 and G9a's Y1154.

2.2.3. DOT1L

DOT1L is a histone lysine methyltransferase of 1537-1739 residues in length (depending on the isoform) whose main function is the mono, di and trimethylation of histone H3K79 [43]. As opposed to other methyltransferases such as G9a or SETD7, DOT1L does not contain a catalytic SET domain, and is more closely related to the arginine methyltransferase (RMT) class [83]. The catalytic core of DOT1L is elongated and consists of an N-terminal domain with five α-helices and two pairs of short β strand hairpins. On the other side, the C terminal domain consists of seven-strand central β sheets (named (β5–β11) and five α-helices. The SAM molecule is situated between the loop L-EF and the C-terminal ends of β5 and β8. Five DOT1L residues make hydrogen bonds with SAM: D161, Q168, E186, T139 and D222. All this structural information was first solved by Min et al. [84]. DOT1L also contains a lysine binding channel that leads to the methyl group of SAM. As there is no available crystallographic structure of DOT1L with the substrate lysine bound, it is not currently clear

how the substrate lysine gets deprotonated for the methyl transfer [85]. The structure of DOT1L's catalytic core bound to SAM is shown in Figure 2-8.

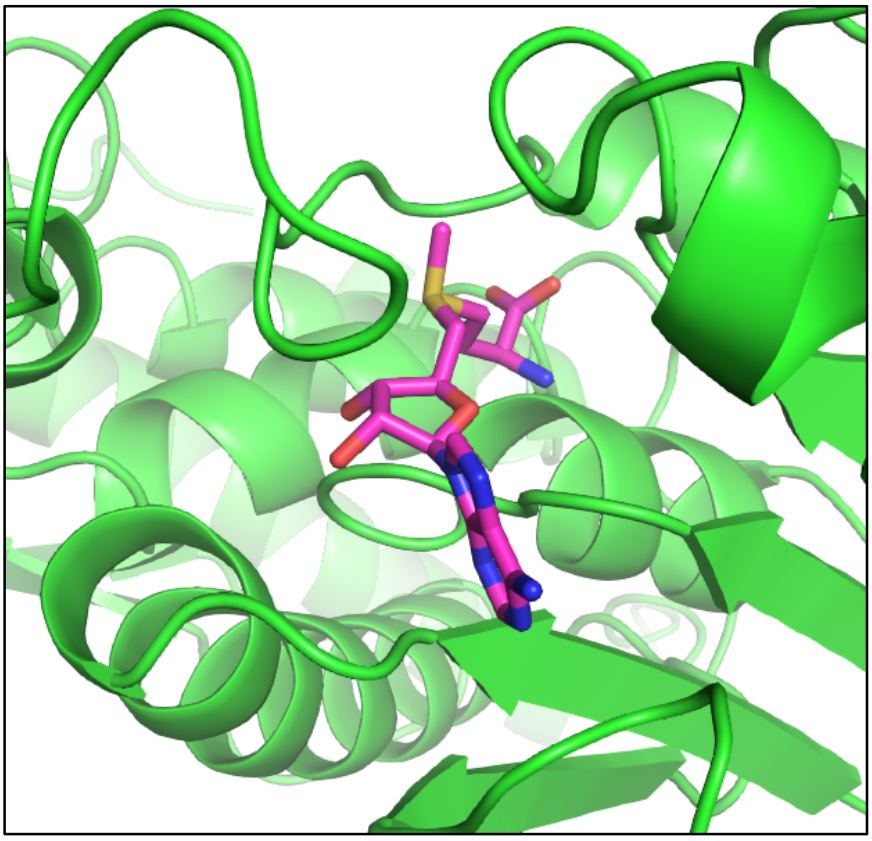


Figure 2-8. Structure of DOT1L's catalytic core (green) bound to SAM (magenta sticks).

The shapes of both SAM binding pocket and histone lysine channel are completely different from those of G9a or SETD7. Besides, the conformation adopted by SAM in DOT1L's pocket is different from SET domain-containing methyltransferases and enables a more direct design of inhibitors being specific towards this enzyme. For example, this SAM conformational difference guided the design of Bromo-deaza-SAH by Yu et al., an inhibitor that can be further exploited towards the development of clinical candidates against mixed lineage leukemia [86].

2.3. Histone and non-histone targets of PMTs

Biological processes require a constant interaction between the different elements of the cell. Signal transduction, which is implemented by specific protein-protein interactions, can be altered by post-translational modifications (PTMs), such as the methylation of Lys and Arg or the phosphorylation of Thr, Ser and Tyr. The study of protein methylation has

traditionally been linked to histone methylation, changing the physical properties of the chromatin leading to activation (monomethylation of H3K4, H3K36 and H3K79) or repression of gene expression (di- and trimethylation on H3K9, H3K27 and H4K20) [87]. This is the reason why they were initially named histone lysine methyltransferases (HKMTs). However, with the discovery of novel targets for these enzymes they were renamed as lysine methyltransferases (KMTs) [88]. The non-histone substrates of KMTs include transcription factors, epigenetic regulators, tumour suppressors, nuclear receptors, and chaperone proteins [89]. The methylation signals emerging from histones are integrated with those of non-histone proteins to extend the regulation of cellular functions, in what seems to be a "PTM" code instead of the more established histone code [90].

The process of methylation is commonly represented as a writer-reader-eraser model as it was shown in Figure 1-5: the methyltransferase (writer) methylates a specific substrate on its Lys or Arg residues, leading to changes that alter the protein structure, activity, localization or the interaction of the substrate with an effector protein (reader) that contains a methyl binding domain; the methylation can then get removed by a demethylase (eraser) [91]. Methylation has traditionally been considered as a very stable modification, but recent studies show that it can be highly dynamic, for example in the active histone methylation by MLL3 and its effect on circadian clock genes [92].

There are at least 1005 Lys methylation sites in 974 human proteins and 2159 Arg methylation sites in 896 human proteins, mostly derived from MS analysis with a majority of non-histone targets [93]. Due to their abundance, these modifications seem to play an important role in the regulation of the cell; however, the function of most of these modifications is unknown [94] and it has only started to be understood. The most studied modifications are those of enzymes that can be readily treated by biochemical assays such as SETD7 or G9a, with SETD7 found to methylate at least other 30 proteins and G9a another 17 including its own automethylation [93]. Analyzing the human methylation network generated by Biggar and Li [93], histones appear as hubs (or high-degree nodes) of this network as epigenetic studies have traditionally focused on histone modifications, while less is known about other (non-histone) targets. Part of the current knowledge about the PTM code network can be found in Figure 2-9.

PTMs are indeed not isolated events, but highly interconnected marks that interact with each other to reach a biological outcome. This crosstalk can happen between different

methylation marks or between methylation and other types of modifications such as phosphorylation, ubiquitylation and acetylation.

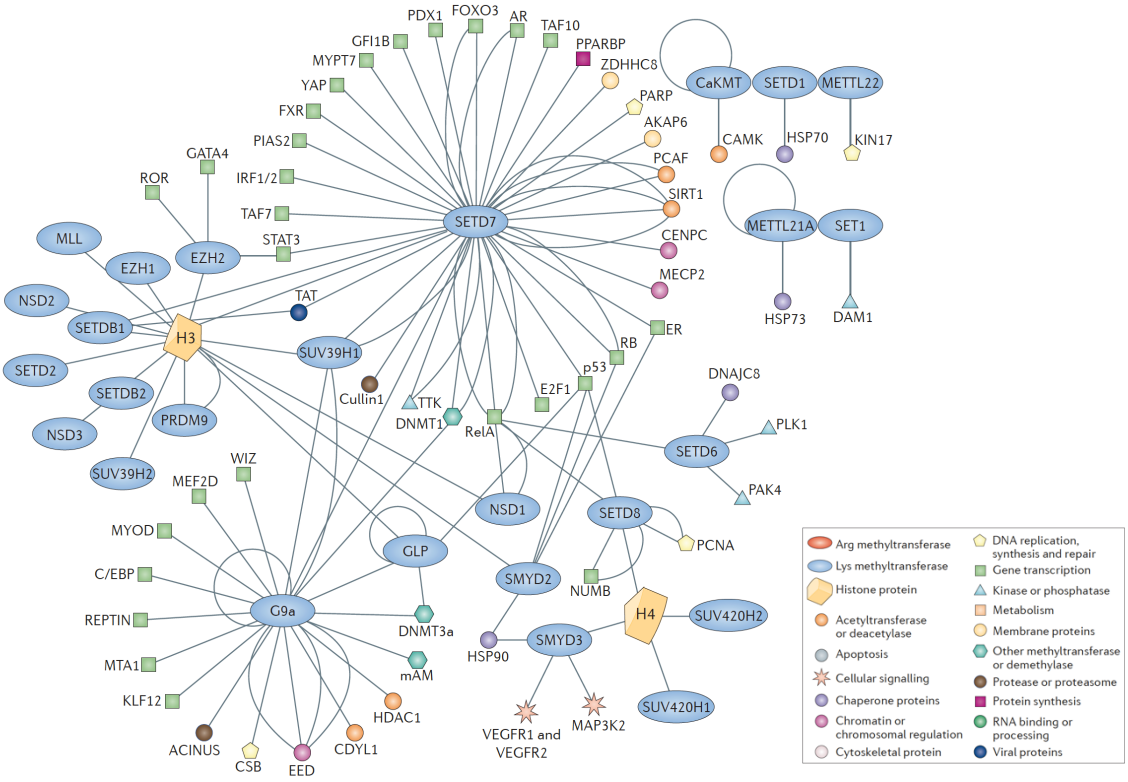


Figure 2-9. Methylation network between lysine methyltransferases (KMT) and their substrates, where each of the lines indicates an interaction. Adapted from Biggar & Li [93].

An example of methylation-methylation crosstalk is the regulation of the transcription factor p53 by the dynamic methylation of a series of lysine residues. For example, the monomethylation of p53 at K370 by SMYD2 represses its transcriptional activity, but the dimethylation of the same site enables the binding of the tumor suppressor p53-binding protein 1 (TP53BP1) by its Tudor domains, leading to transcriptional activation [95]. Besides, the dimethylation of the nearby K372 inhibits the methylation of K370 by SMYD2, therefore promoting gene expression [96]. In addition, K382 monomethylation promotes the binding of the tumour-like protein L3MBTL1 leading to transcriptional repression of p53, whereas K382 dimethylation promotes the binding of PHF20 leading to transcriptional activation [97]. The cross-talk of these chemical modifications is a process where the dynamic kinetic stability (DKS) characteristic of life [98] seems to emerge towards a higher complexity.

Protein methylation holds a broad influence on essential biological functions such as chromatin structure remodelling; DNA replication, synthesis and repair; protein synthesis;

RNA metabolism; cell cycle progression; apoptosis; gene transcription and signal transduction [93]. Protein methylation is a critical component in the regulation of cell signalling pathways such as Hippo [99], JAK-STAT [100], MAPK [101], WNT [102] and BMP [103] by regulating either the activity of the modified protein or the recruitment of other regulatory proteins to the receptor.

2.3.1. G9a substrates

G9a is a histone methyltransferase which can mono- and dimethylate lysine 9 of histone H3, resulting in H3K9me1 and H3K9me2, respectively [75]. This histone methylation acts as a scaffold for the subsequent recruitment of HP1 proteins, highly conserved proteins associated with important functions in the cell such as gene repression by heterochromatin formation or transcriptional activation [104]. Other important histone modifications connected to this enzyme are histone 3 lysine 56 monomethylation (H3K56me1), which enhances the interaction between the H3 and proliferating cell nuclear antigens (PCNA) and also regulates DNA replication [105]; histone 3 lysine 27 monomethylation [67], which might act as a controller for the subsequent di- and tri- methylation by the Polycomb Repressive Complex 2 (PRC2) [67,77] ; and histone H1, for example at the H1.4.K26 position *in vivo* and *in vitro*, providing a recognition site for the chromatin binding proteins HP1 and L3MBTL1 [68,106]. G9a seems targeted to histone H3 by DNA binding proteins such as E2F6, MGA, MAX and DP1 [75].

Apart from histones, G9a's targets include itself (automethylation of G9a's K239, which is located in a G9a's motif (ARKT) similar to histone H3K9 regions (ARKS)) and proteins such as CDYL at K135, WIZ at K305, ACIN1 at K654, Reptin at K67, p53 at K372, DNMT1, HDAC1, ERCC6, and KLF12[80]. The automethylation of G9a does not halt or enhance the enzymatic activity *in vitro*. The motif at which G9a automethylates (ARKT) resembles the one of the histone H3K9 region (ARKS), and it is identical to other motifs in EuHMT and mAM. This automethylation at the ARKT motif creates a binding site on G9a for the HP1 class of proteins, and the mutation of the lysine in the motif to K239A substantially reduces the co-localisation [104,107]. Therefore, it can be concluded that the non-histone protein methylation generates binding sites used in protein-protein interactions.

2.3.2. SETD7 substrates

As commented previously, SETD7 is a methyltransferase known for the specific monomethylation of histone H3 at the K4 position, being this an important tag for transcriptional activation [81]. Some of the genes that are activated by this mechanism are collagenase [108] and insulin, where SETD7 is recruited by IPF1/PDX-1 to the insulin promoter [109].

SETD7 is also known to have non-histone targets. For instance, it methylates the K372 of p53/TP53, stabilizing it and leading to an increased activity of this tumour suppressor [110]; SETD7 also monomethylates K189 of TAF10, increasing the affinity of this protein towards RNA polymerase II [111] and TAF7 [112] by recognising the (KR)(STA)K motif in the substrate proteins [113]. Another known target of SETD7 is DNMT1, methylating it at K142 position leading to DNMT1 degradation via proteasome pathways [112]. In total, SETD7 has been found to methylate at least 30 non-histone proteins [93]. Together with G9a, SETD7 is one of the main focal points of methyltransferase-substrate networks, as could be seen at Figure 2-9.

2.3.3. DOT1L substrates

DOT1L is, as previously mentioned, an H3K79 methyltransferase that does not possess a SET domain, but instead binds SAM in a similar way to class I arginine methyltransferases [114]. As opposed to other histone residues known to be methylated, the H3K79 residue is not located on the H3 tail, but on the globular, surface-exposed domain of H3 [115]. Therefore, H3K79 methylation is able to regulate transcription at both euchromatin and heterochromatin [114], being essential for mammalian development and heterochromatin structure [116]. Another remarkable feature of DOT1L is that it is the only H3K79 methylase that can catalyse mono-, di-, and trimethylation non-processively [117], such that the probability of an H3K79me1 mark becoming H3K79me2 and an H3K79me2 mark becoming H3K79me3 is the same. However, it is known that the different H3K79 methylation states are related to different transcriptional outcomes [118]. Therefore, it is possible that its enzymatic activity is regulated by other proteins such as such as Rad6/Bre1 and the Paf1 complex in yeast and equivalent proteins in humans, or by a corresponding H3K79 demethylase [43].

DOT1L is not only critical for transcription, but also participates in other processes such as DNA damage response and cell cycle regulation [43]. In addition, the mammalian DOT1L gene is essential for functions such as cardiac function, embryogenesis and hematopoiesis [114], and is involved in MLL-related leukemogenesis [119]. Aging (damage accumulation and loss of cellular homeostasis) [116] is the main risk factor for the principal human pathologies [120]. The methylation of H3K79 has been proven to change after histone deposition, with methylation being accumulated on H3K79 on aging histones, relating cell-cycle length to changes in nucleosomal modifications and acting as a timer mechanism [121]. DOT1L has also been found to be a relevant effector of the role of the transcriptional protein complex NF-kB role in cellular reprogramming [122]. It has been found that providing inhibitors specific to DOT1L increased the reprogramming efficiency in Néstor-Guillermo progeria syndrome (NGPS) fibroblasts, suggesting that DOT1L could be an excellent target for rejuvenation therapies [123].

For all these reasons, DOT1L has emerged as a novel therapeutic target through the development of specific inhibitors such as Bromo-deaza-SAH [86] or phenoxyacetamide derivatives [124]. Research interest in DOT1L inhibitor discovery has been encouraged by a straightforward design, as the binding mode and pockets of SAM and the histone lysine to DOT1L are different from SET domain methyltransferases. However, the design of molecules specific towards a particular set of SET domain methyltransferases such as G9a has traditionally relied on high throughput screening [125] approaches due to methyltransferase non-specificity. This is one of the reasons why finding structural differences between different KMTs becomes so important, not only for the development of new inhibitors but also for generating specific cofactors able to be utilised by some methyltransferases but not others; this could provide mechanistic insights into cancer emergence and progression. The development of novel drugs targeting Lys or Arg methyltransferases or demethylases seems like an attractive therapeutic approach; however, in order to develop targeted treatments, it would be necessary to understand not only the sites that are methylated, but also which is each specific enzyme performing each particular action in the cell.

The delineation of the methylproteome and its implications on different disorders has just started when compared to the high number of identified phosphorylated sites. In order to develop drugs for diseases where epigenetic modifications are involved, it is necessary to

understand the exact mechanims and cross-talk between methylation marks and the proteins involved in their writing or removal. There is a need for more developed epigenetic methylation profiling methods to elucidate the network complexity of writers, readers and erasers of the PTM code. The latest advances in both molecular and cellular experimental approaches, mass spectrometry and proteomic bioinformatic analyses show promise for a rapid advancement of this field.

This project aims to perform an unambiguous identification of specific methylation targets of SET domain-containing KMTs. KMT methylation regulates signalling pathways, tumour development, gene expression and cancer, among other processes [126]. Most assays employed to identify non-histone KMTs methylation events are currently based on screening tools that not being optimised for specific sets of enzymes. The design of specific cofactors could help elucidating the functions of these marks by providing a detailed description of the methylation targets of each methyltransferase and therefore contributing to further understand the PTM code.

Chapter 3

Epigenetic profiling methods: background and motivation

The focus of this project is profiling post-translational modifications, more specifically the substrates of lysine methyltransferases (KMTs), epigenetic enzymes playing a crucial role in the regulation of gene expression, cellular reprogramming and response to the environment by methylating histone and other protein substrates.

Proteomics is widely defined as the large-scale research of proteins, principally their sequence and function within living organisms [127]. The complete group of proteins that can be expressed by a given genome in a living thing receives the name of proteome [128]. The proteome of an organism represents a vast collection of proteins which interact with different kinds of molecules through complex pathways [129–131]. Explaining the mechanisms of these networks is crucial not only to understand the homeostasis of an organism but also to find new ways to discern the pathologic mechanisms and propose adequate treatments.

Even though the central dogma of molecular biology states that the DNA contains all the instructions needed to develop a full organism, proteomics studies about epigenetic marks, protein-protein interactions or post-translational modifications (PTMs) have shown that the phenotype is also regulated by epigenetic modifications [132,133]. This fact considerably expands

the amount of information contained in the human genome, and may explain the greater complexity in higher eukaryotes compared with other organisms, even though their genetic compositions are similar [134].

A more precise study of the proteome is fundamental to expand our knowledge about many diseases, such as cancer. The collection and analysis of the proteome data is far more complex than that of the genome as it varies depending on the cell cycle stage and cellular location where the data is measured. However, tools such as mass spectrometry (MS) and recent advances in bioinformatics are overcoming those barriers [132,135]. These new techniques to decipher the proteome operation have an enormous biomedical potential; for example, they could be used in the development of drugs targeting epigenetic regulators, which could modify diseased cells back to their healthy phenotype. In order to achieve that goal, further characterisation of the residues undergoing post-translational modifications and how they affect cellular pathways is required.

3.1. Substrate methylation detection methods

The complexity of biological systems makes necessary the development of functional tools to analyse the contribution of each element to organismal homeostasis. Chemical reporters in combination with bioorthogonal labelling assays can be used to analyse proteins, nucleic acids, glycans, lipids or other metabolites *in vitro*, in cells or even in an entire organism [136]. Even though many assays have been designed to decipher the specific methylome contribution of a given KMT, to date there is no biologically consistent cell-based model available. Among the obstacles that prevent proper identification of lysine methylation targets is the fact that some redundancy in KMT activity exists, meaning that there is more than one type of KMT that can methylate a specific histone site. For example, H3K9 site can be methylated by SUV39H1, SUV39H2, SETDB1, PRDM2, GLP and G9a [55], so an analysis of the methylated residues, using mass spectrometry or other techniques, show the cumulative activity of all these KMTs, but not one of a particular enzyme.

As mentioned in the previous chapter, all KMTs except DOT1L contain a SET domain that is responsible for their catalytic activity, and this conservation makes the activity of these enzymes more difficult to differentiate. The fact that KMTs can methylate substrates other

than histones, such as transcription factors and chromatin modifiers, makes it increasingly important to find the individual contribution of a given KMT [137]. To achieve this, current KMT activity assays involve the quantification of the methylated histone products or the reaction by-product SAH, as seen in Figure 3-1.

The reason to analyse product appearance and not reactant depletion is the slow enzymatic turnover of these protein methyltransferases (PMTs) [138,139]. The methylated protein can be directly analysed using mass spectrometry, radiometric or antibody-based methods. On the other hand, SAH can be directly quantified using mass spectrometry or anti-SAH antibody methods, or using enzymatic colorimetric assays. The next sections will summarise the most important methods and provide a critical comparison between them. Some of the existing assays for profiling the activity of KMTs are shown in Figure 3-1.

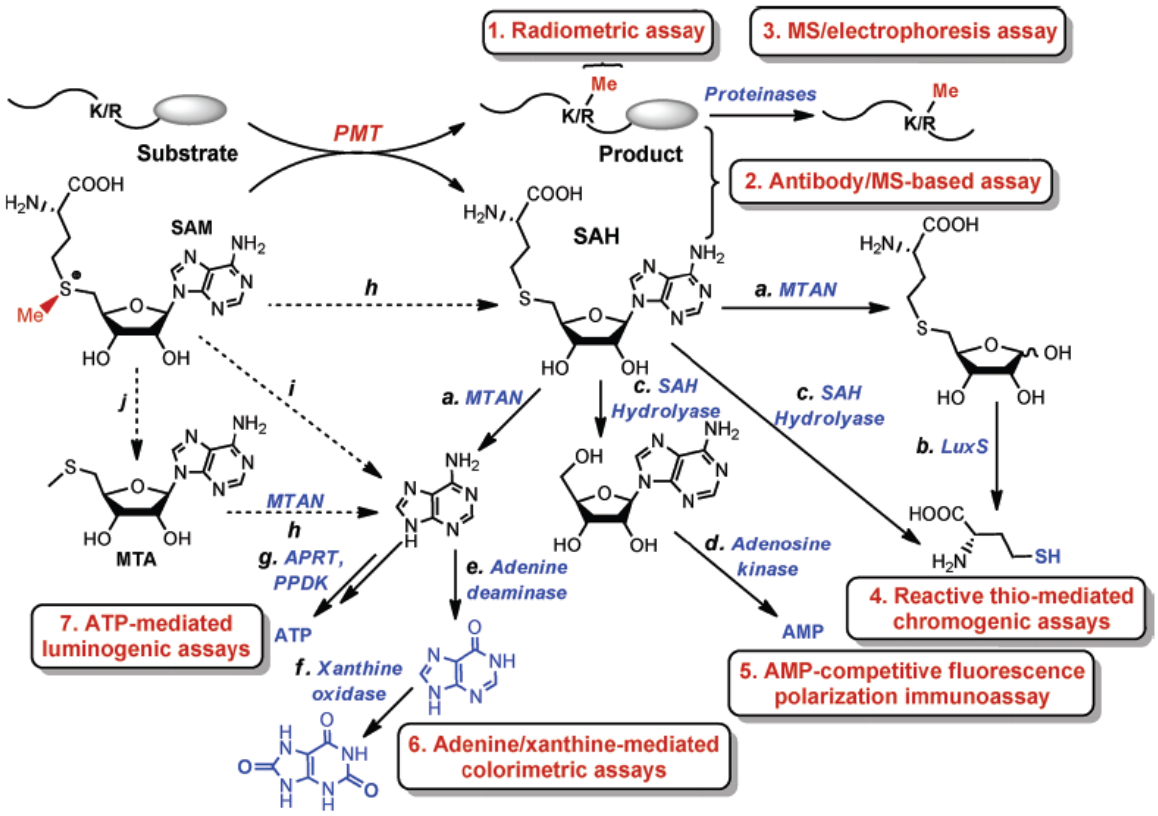


Figure 3-1. Summary of protein methyltransferases activity assays and chemicals used in them [140]. Reprinted with permission from Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. ACS Chem. Biol. 7, 443–63 (2012). Copyright 2012 American Chemical Society.

3.1.1. Radiometric quantification

In this assay, a radiolabelled methyl group coming from [Me-^{3}H] or [Me-^{14}C] SAM is introduced into a substrate target. Once the unreacted SAM is removed, the methylations can be quantified using techniques such as liquid scintillation counting or autoradiography [141]. To separate the unreacted SAM from the sample, the most established assay is using phosphocellulose filter paper to bind the protein products, then washing and scintillation counting or SDS-PAGE separation, and then autoradiography or gel extraction/scintillation counting. Even though these assays are consistent and with relative low complexity, they are also arduous, requiring many washing and transferring steps, and time consuming, requiring between 6 hours and 7 days [142].

To overcome the limitations of basic radiometric methods, a Scintillation Proximity Assay (SPA) is used. In this procedure, the already biotinylated peptide substrate is then radiolabeled with [Me-^{3}H]-labelled SAM. The biotinylated residues bind to streptavidin-conjugated SPA beads, and the protein is immobilised. The radiolabeled residues emit β-particles, and so the SPA beads emit light due to their proximity from the residues. The SPA signals can then be easily detected via scintillation counting. The main advantage of this method is that it does not require the separation of the unreacted SAM, therefore it can be readily adaptable for High Throughput Screening (HTS), saving time and effort to future researchers in this field [143]. A scheme of this method is shown in Figure 3-2.

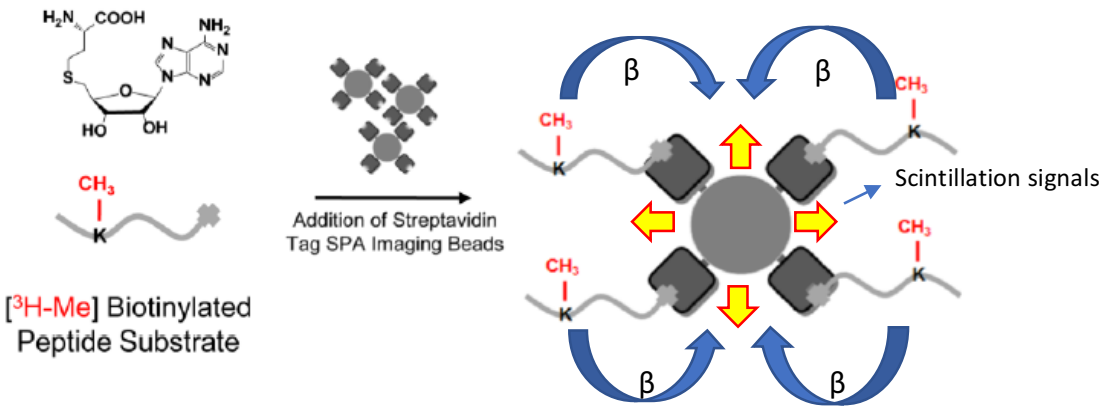


Figure 3-2. SPA-based HTS approach. The β particles emitted by the radioactive groups are received by the SPA beads attached to the biotin–tagged substrates, emitting light that can be detected and related to the methylation state. Adapted from Ibáñez [139].

Although the previously reviewed radiometric methods are adequate to assess when methylation is occurring, they cannot establish the exact position of these methylation marks nor the degree of methylation. Therefore, radiometric methods should be used to validate new KMT activities but not to analyse the complexity of the PTM code.

3.1.2. Antibody-based detection

In order to overcome radiometric assay limitations such as the use of radioactive chemicals and the fact that they can only show methylation activity but not the degree of methylation, antibody-based assays were developed based on monoclonal or polyclonal antibodies that recognise certain epitopes and allow the analysis of the methylation sites using Western blot, CHIP, CHIP-on-chip and CHIP-Seq techniques [61]. Together with other techniques such as AlphaScreen, AlphaLISA, LanthaScreen or LANCE Ultra, this assay can be used in homogeneous KMTs activity assays [144].

All the antibody-based homogeneous KMTs activity assays share a comparable principle: pairing the KMT substrate and an anti-methyl-lysine antibody with donor and acceptor colouring agents. The binding of the methylated protein with the antibody makes the donor and acceptor dyes closer to each other, and then the excitation of the donor dye causes the release of singlet oxygen for AlphaScreen and AlphaLISA or time resolved fluorescence resonance energy transfer (TR-FRET) for LANCE Ultra and LanthaScreen [140], which excite the acceptor making it release emissions that can be related to the level of substrate modification.

Some successful examples of the antibody-based detection technique are reported by Quinn [145], using the AlphaScreen technology with a polyclonal anti-methyl-H3-K9 antibody to evaluate H3K9 methylation, and by Gauthier, using AlphaLISA to research the SET7/9 H3K4 methylation, and also verified the LANCE Ultra technology using a europium-labelled antimethyllysine antibody [144]. A representation of both AlphaLISA and LANCE Ultra assays is shown in Figure 3-3. Even though these assays are more specific and adaptable for high throughput screening (HTS), the bottleneck of these approaches is that the specificity of the antibodies and the range of epitope concentrations need to be sufficiently defined prior to experimentation, so it is not a good approach to find new

methylation sites as this project proposes. Furthermore, as the concentrations stay within a very narrow range, these antibody-based assays are not able to estimate essential quantitative data such as k_{cat} or K_m. Another point to highlight is the fact that some enzymes can mono-, di- and tri- methylate the substrate, and antibodies should recognise all of them differently [146]. Despite all these drawbacks, this represents an adequate technique for HTS or to probe cell based methylation events.

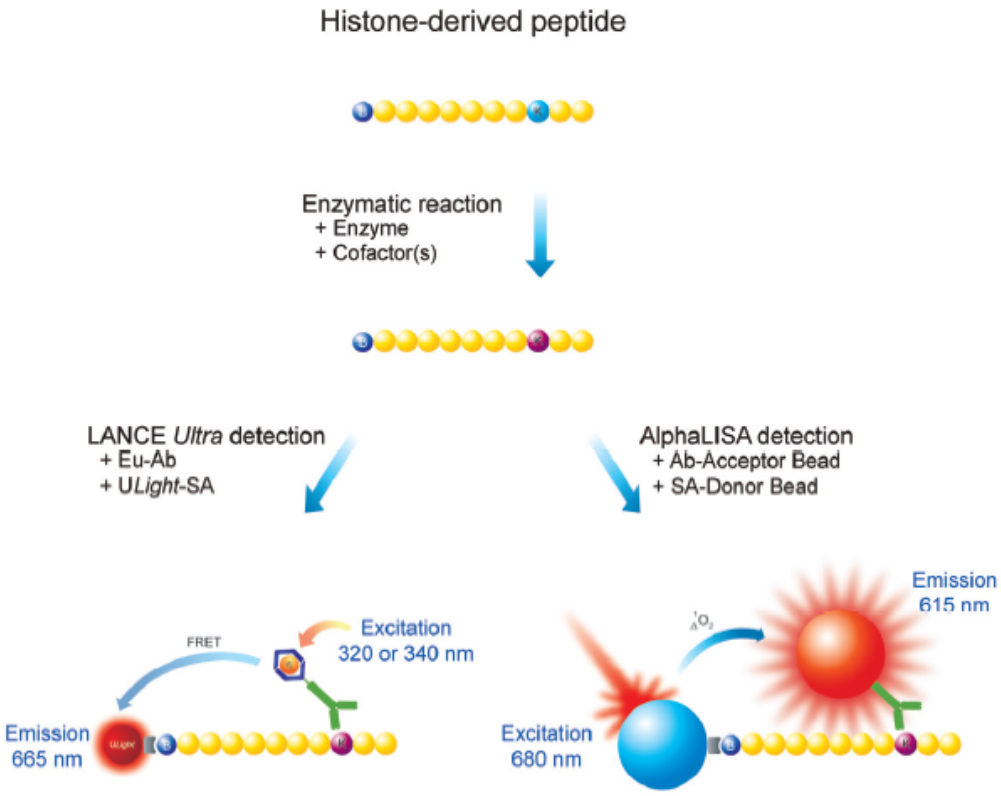


Figure 3-3. Illustrative model of LANCE Ultra and AlphaLISA epigenetic assays [144]. Reprinted with permission from Gauthier, N. et al. Development of homogeneous nonradioactive methyltransferase and demethylase assays targeting histone H3 lysine 4. J. Biomol. Screen. 17, 49–58 (2012). Copyright 2012 Society for Laboratory Automation and Screening.

3.1.3. Mass spectrometry

Mass spectrometry (MS) techniques for detecting post translational modifications can be differentiated by whether they can be directly applied to the proteins, or by whether they require substrate digestion prior to detection. The most basic classification divides the assays into three groups: bottom-up MS, in which proteins are digested using trypsin producing short peptides; middle-down MS, which uses GluC (proteinase which

preferentially cleaves peptide bonds C-terminal to glutamic acid residues) or AspN (proteinase which selectively cleaves peptide bonds N-terminal to aspartic acid residues), for H3 and H4 respectively, to produce only a partial digestion, and top-down MS, working directly with unmodified proteins. Figure 3-4 shows a schematic representation of these techniques.

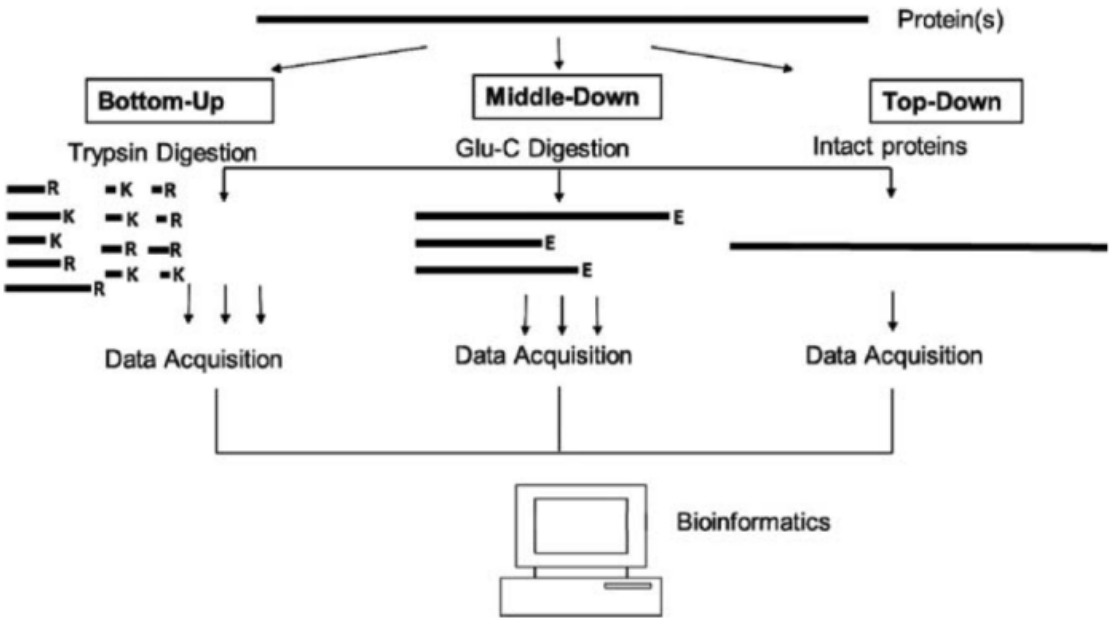


Figure 3-4. Simplified schemes of the bottom-up, middle-down, and top-down MS strategies to analyse histone PTMs [147].

3.1.3.1. Top-down MS

If small peptide substrates are used, PTM products can be directly analysed by top-down MS without requiring a digestion step before the MS analysis. The characterisation of the level of methylation is related to the equivalent mass shifts, being +14Da for monomethylation, +28Da for dimethylation and +42Da for trimethylation so information can be discerned regarding the stoichiometry of methylation. Top-down MS suits histones due to their small size and relatively large cellular abundance and can be used to detect and record accurate mass values for the molecular and fragment ions observed, to identify the most abundant proteins, to determine the isoforms present and to localise their post-translational modifications [148]. The top-down technique was combined with metabolic labelling to monitor the cell cycle dependency of H4K20 methylation [149]. One of the main advantages of this approach is that in theory all the PTMs are observed as the proteins are

kept intact [150]. Moreover, it can also differentiate between different histone variants, and there is no digestion step and all PTMs are conserved.

Top-down MS remains very limited in use due to the expensive instrumentation required, in addition to the fact that it is quite insensitive, requiring considerable amounts of sample to achieve a proper identification and sometimes leading to ambiguous PTM assignments [151]. In fact, to analyse histones properly requires a considerable effort on prefractionation [152], which makes this technique less applicable for high throughput screening. As the analysis of the histone code requires a great amount of biological sampling to be further understood, top-down MS appears limited for histone PTM analysis.

3.1.3.2. Middle-down MS

Even though the majority of current proteomic MS experiments are developed using bottom-up MS, recent technological advances in protein fragmentation such as electron capture dissociation (ECD) and electron-transfer dissociation (ETD), based on the transfer of an unpaired low energy electron to the substrate, have consolidated the use of top down and middle-down MS [153]. These recent techniques are well suited for histone analysis than collision induced dissociation (CID), as they are especially effective when peptides are rich in basic residues [4].

The main difference between top-down and middle-down MS lies in targeted or untargeted protein digestion in order to produce peptides, commonly within the rage 3 to 9 kDa [147]. Compared to the use of trypsin in bottom-up techniques, the use of AspN or GluC in middle-down MS results in larger peptides due to the scarcity of these residues in histones compared to lysine or arginine, enabling the analysis of PTMs occurring at the same peptide while obtaining sensitivities close to the bottom-up approach [154].

3.1.3.3. Bottom-up MS

In contrast with top-down MS assays, which are limited to short proteins, bottom-up MS can be used for all types and sizes of proteins. This technique allows the mapping of methylation sites and it has successfully identified novel PTM sites on histones [6]. In this technique, proteins are first digested using enzymes such as trypsin, which cleaves the

peptide chains specifically at the C-terminal end of the basic amino acids lysine or arginine, except when they are followed by proline residue. As histone tails are rich in these residues, the trypsin usage causes very short peptides with only 3 to 4 amino acids on average, making them less compatible with MS for the following reasons: some peptides would elute with the dead volume of the chromatography and would not even be detected; even if detected, some 6 to 7 amino acid peptides are shared between a considerable amount of histone variants, complicating the assignment of a variant to each modification, and even if the assignment is achieved, it would be lost by trypsin digestion because of the small differences between their sequences [155].

Key disadvantages of bottom-up MS PTM analysis are that trypsin cleavage does not occur when the lysine or arginine residues contain a PTM such as methylation or acetylation [4] and that it is not possible to get information of PTMs happening simultaneously on 2 adjacent tryptic peptides. In order to overcome these limitations, the combination of bottom-up approaches with chemical derivatisation, in particular with propionic anhydride, can be used [156]. This approach helps reducing the trypsin-based fractionation of the peptides by adding propionyl groups to unmodified and monomethylated lysines. This blocks the peptide cleavage after lysines, generating peptides cut at arginine-terminal sites. This modification also results in more hydrophobic peptides, which are better retained during liquid chromatography [157].

One of the most important methods for peptide quantification is stable isotope labelling by amino acids in cell culture (SILAC). This introduces heavy arginine or lysine amino acids within the protein structure, which are then easily detectable using MS [158,159]. When using MS for histone PTM analysis, proteins can be marked with heavy amino acids or not, depending on the culture media used for cell growth. The proteins are then digested into peptides, separated using HPLC to be consequently electro-sprayed into the mass spectrometer. A full scan mass spectrometer records the masses of the peptides, and then the chosen peptides are isolated and fragmented, with their masses resolved by MS/MS [160]. The found masses are then used with bioinformatics algorithms for peptide identification and quantification to determine the correspondent PTMs at each position [4]. Figure 3-5 shows a broad overview of the current workflows that can be used to profile PTMs using LC-MS/MS.

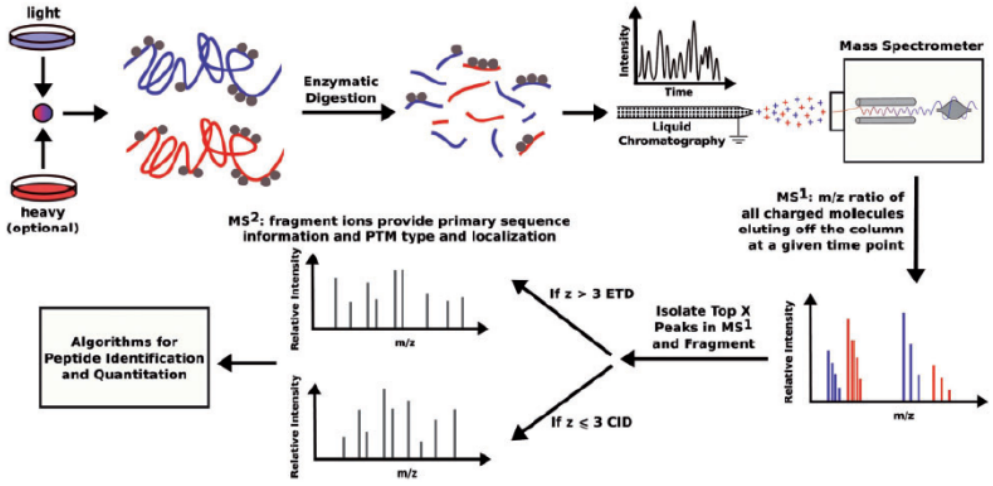


Figure 3-5. Overview of mass spectrometry assays for proteomic studies [4].

3.2. SAH detection methods

The other product that can be quantified in the KMT methylation reaction is the by-product SAH. One of the most developed and accepted methods for direct SAH detection is ultrahigh-performance liquid chromatography tandem mass spectrometry (UPLC-MS/MS) [161]. However, the challenges that arise from using this methodology for direct SAH detection are varied. For example, the detected signal can come from three other sources apart from the enzymatic activity on histones: enzymatic automethylation, SAH presence in SAM reagents and non-enzymatic pseudo-first order reaction of SAM to SAH [162]. There is another important method for directly quantifying the presence of SAH called the anti-SAH antibody assay, in which antibody competitive assays are used to measure the appearance of the by-product SAH. One of the most successful cases is the one produced by Graves et al., in which SAH is measured by competing fluorescein-SAH to bind the antibody, resulting in a decrease of fluorescent signal. Despite being successful for catechol-O-methyltransferase, the readout of SAH-based fluorescence polarisation is linear in a very narrow range of SAH concentration (from 10 nM-500nM), and so the measurements could be easily erroneous [163].

To overcome all the previously mentioned limitations of the direct quantification of SAH, KMT activity assays were developed using SAH derivatives via colorimetric assays with coupling enzymes. Some of the examples shown in a previous section (see Figure 3-1) will

be further explained. First, the fluorogenic assay for catechol-O-methyltransferase, which uses SAH hydrolases to get SAH converted into homocysteine [164], which can be then measured via free-thiol-activated dye fluorescein-cystamine-methyl red (FL-S-S-MR). Second, an ultrasensitive luminescence assay by Ibañez el at., in which SAH is transformed via the following sequence: into adenine using MTAN, then into adenosine monophosphate by adenine phosphoribosyl transferase and finally into ATP by pyruvate orthophosphate dikinase. The resulting ATP is easily detected using a luciferin/luciferase kit, being able to measure quantities as low as 0.3 pmol of SAH [165]. More recent studies have used SAH-hydrolase-mediated fluorogenic quantification with demonstrated HTS readiness [166]. Nevertheless, all coupling enzymes used provide interference that might join to the already mentioned decomposition of SAM to SAH. These methods seem more appropriate to detect and quantify the kinetics of high-turnover PMTs, but not low turnover ones, in which a protein quantification approach would result more suitable. In addition, detecting SAH does not provide any indication about the specific targets or the number of methylations per site.

3.3. Bump-and-hole approach

The bump-and-hole approach consists in introducing a mutation in an enzyme's active site with the objective of creating a "hole" so the enzyme gains selectivity towards a certain "bumped" cofactor [167]. The "bumped" cofactor is a chemical probe, that needs to be differentiated from a drug design as it requires a defined mechanism of action (MoA) and selectivity among other characteristics as it appears in Figure 3-6.

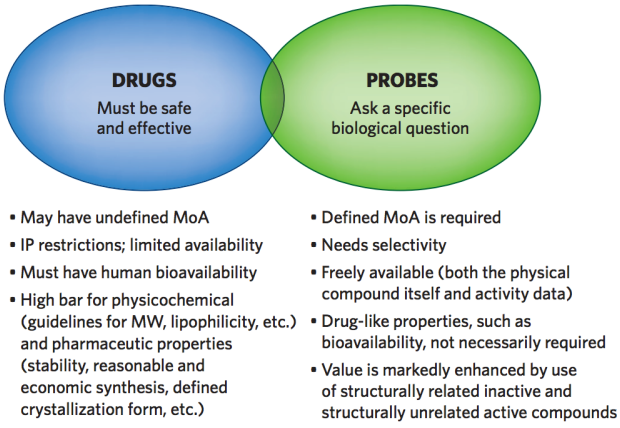


Figure 3-6. Differences in requirements and applications for chemical probes and drugs [168].

The typical strategy to generate a hole in an enzyme is a point mutation in the active site from a bigger to a smaller residue. The ideal bump-and-hole design of a mutant protein paired with an analogous cofactor follows the following criteria for the mutant enzyme: it should bind a cofactor which is not recognised by the endogenous enzyme, have reduced specificity for the endogenous cofactor and catalyse the transfer of any donor group if present. A diagram showing the basic idea of this approach is shown in Figure 3-7.

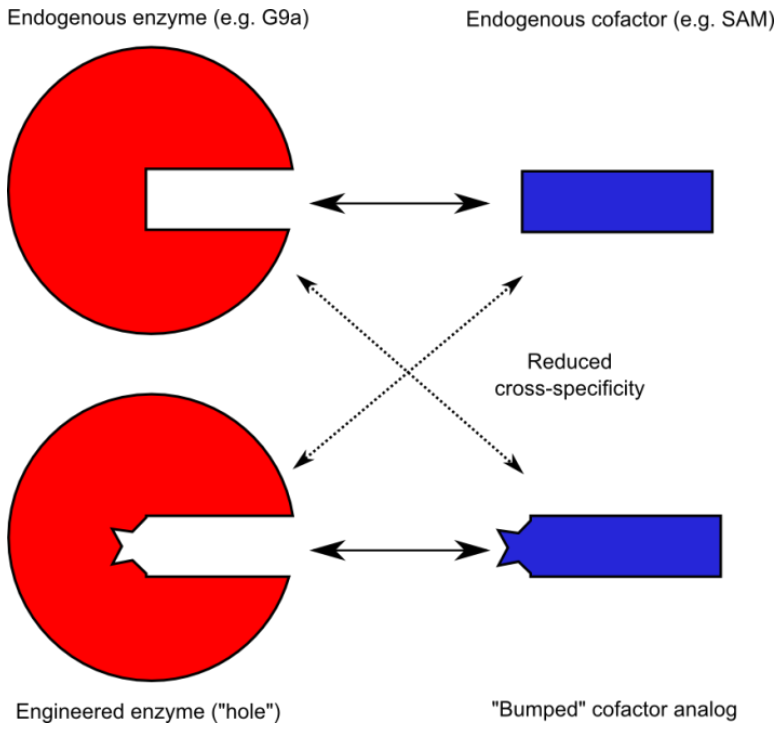


Figure 3-7. Schematic representation of the "bump-and-hole" approach.

Even when this method is mainly driven by steric considerations, the bump-and-hole approach has demonstrated its viability for the design of high specific mutant enzyme-cofactor analogue pairs [169]. For example, Ciulli's lab successfully applied the bump-and-hole approach to achieve selectivity for the second BET bromodomains of the BET family over the first bromodomains [170], doing so by mutating the leucine residue L94 of the BET bromodomain protein Brd4 to alanine and coupling this mutant Brd4 with an ethyl derivative of a small molecule inhibitor, I-BET/JQ1, to match the generated hole in Brd4. This technique is especially challenging in two points: the approach for engineering KMT-SAM systems and the appropriateness for integration into cellular models. This short review will focus on adenosine-based cofactors as the cofactor used in this project, SAM, forms part of this class. Protein kinases are the most studied enzymes that utilise adenosine-based cofactors, followed by methyltransferases.

3.3.1. Bump-and-hole approach applied to kinases

Kinases are enzymes that transfer phosphate groups from high-energy donor molecules, such as ATP, to certain substrates via phosphorylation. The structure of the ATP molecule is shown in Figure 3-8.

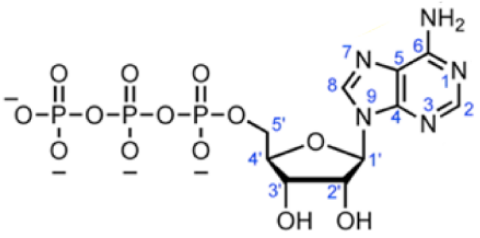


Figure 3-8. ATP chemical structure.

The considerable amount of crystallographic structures generated for the kinase-ATP system allowed a greater understanding of the binding between these 2 molecules, and thus better and more structurally rational designs of bump-and hole-modifications have emerged, allowing the *de novo* design of protein kinase inhibitors, for example [171]. The ATP binding site shares similar characteristics for most of the 518 human kinases [172]. Figure 3-9 shows a simplified representation of this acceptor site.

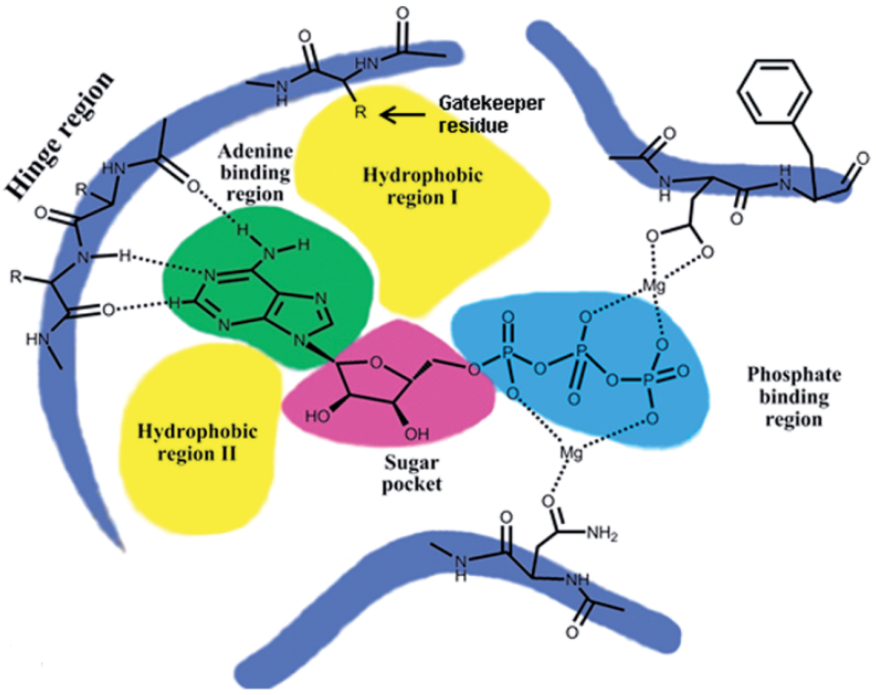


Figure 3-9. Schematic ATP binding site for the majority of protein kinases [171].

The most important recognition motif in the kinase structure is the hinge region, which provides 3 hydrogen bonds to the adenine functional group of the ATP cofactor, thus stabilising this binding and increasing its specificity. This adenine binding region is surrounded by 2 adjacent hydrophobic pockets formed by the N- and C- terminal lobes of the kinase catalytic domain, which are currently on the focus of many inhibitor design assays as they are more variable between the individual kinases than on the hinge region [173]. As can be seen in Figure 3-9, there is a "gatekeeper" residue in the hinge region of kinases that acts as an activator and increases the specificity for ATP while decreasing the specificity for other, similar molecules. Consequently, the modification of this residue to a sterically less restricting residue such as glycine or alanine would allow the creation of bump-and-hole pairs. For instance, a T338G mutation in c-Src with a N_6-(benzyl) ATP demonstrated specificity without changing the structure of the binding site of the kinase [174]. These approaches have been extended to *in-vivo* assays to identify the targets of specific kinases [175].

The engineering of protein kinases to selectively bind modified substrates was shown by Shah et al. 20 years ago, with the objective of identifying kinase substrates with more specificity [176]. Using the Rous sarcoma virus tyrosine kinase (v-Src) as the model system, they mutated the ATP binding site of the kinase with 2 point mutations to alanines (V323A, I338A) to create a hole in the structure so a particular ATP analogue could bind with increased specificity. They found that the ATP analogue showing the highest specificity was N6-(cyclopentyl). Analysing Lineweaver plots of the rate data, they determined the turnover number (k_{cat}), Michaelis constant (K_m) and the specificity constant (k_{cat}/K_m) for each enzyme-cofactor pair, and demonstrated that the catalytic efficiencies for the new system resembled those of the endogenous pair. The structure of the N6-(cyclopentyl) ATP analogue is compared to the structure of native ATP in Figure 3-10.

The N6-(cyclopentyl) ATP analogue presented a very low specificity for the wild type kinase, therefore validating that the approach of generating a structural distinction in the cofactor binding site of a selected kinase was appropriate to distinguish the specific activity of the modified enzyme when redundancy with other kinases exists. However, these efforts presented plenty of room for improvement, as the specificity of the engineered enzyme for native ATP was even higher than for the new analogue, and both were 50 times lower than that of the endogenous pair [176].

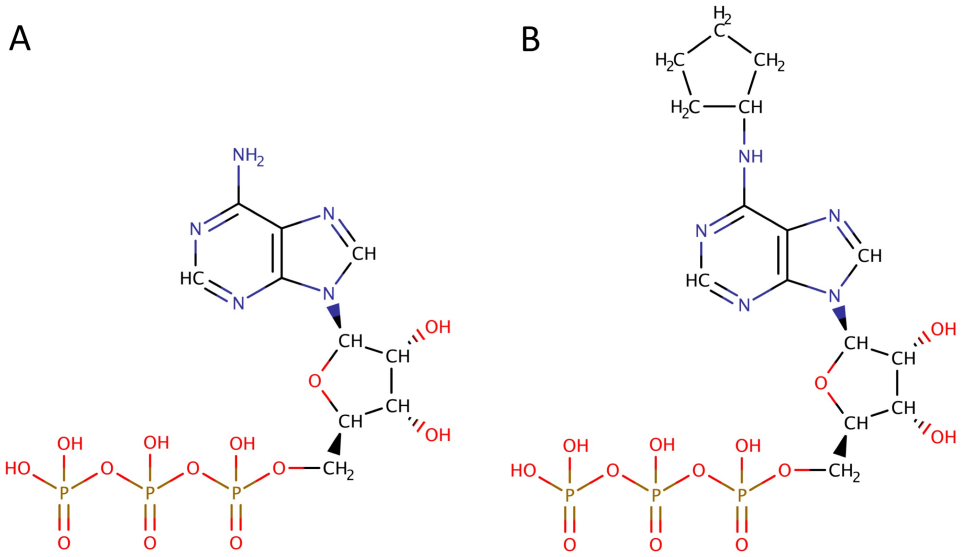


Figure 3-10. A) Native ATP. B) N6-(cyclopentyl) ATP analogue.

3.3.2. Bump-and-hole approach applied to protein methyltransferases (PMTs)

Despite the success of the bump-and-hole method for kinases, the engineering of PMTs (including KMTs) to bind and utilise SAM cofactor analogues has not been so extensively studied, probably because of the higher variance in PMTs binding pockets. The chemical structure of SAM is shown in Figure 3-11.

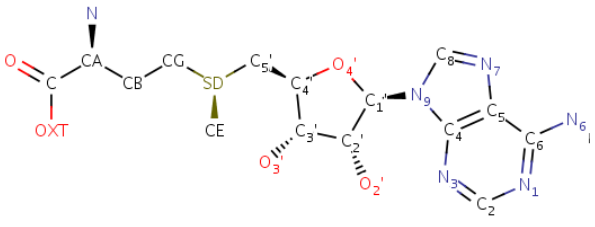


Figure 3-11. SAM chemical structure.

In an effort to develop a bump-and-hole method for PMTs, the plan of action for kinases was directly applied to methyltransferases by the Gray laboratory, using the protein arginine methyltransferase (PRMT) Rmt1 as a model system [167], as this was the only PRMT structure known at that time. To achieve this, the Gray laboratory designed a series of N6-substituted SAH analogues intended to bind Rmt1 at the cofactor binding site, identifying the mutations

in Rmt1 via structurally-guided designs which could make the selectivity for one of these non-endogenous cofactors increase. It was found that N6-naphthylmethyl and especially N6-benzyl SAH (see Figure 3-12, A) could inhibit the mutant enzyme E117G Rmt1 with a selectivity higher than 20 over the endogenous enzyme. The discovery of N6-benzyl-SAM specificity was particularly important as this analogue also shows biorthogonality *in vivo*, being utilised by the mutant over the endogenous Rmt1 with a selectivity greater than 67 [167].

Later studies by the Zhou laboratory continued the research initiated by the Gray laboratory and attempted to broaden the catalogue of SAM analogues, using vSET as the model SET domain-containing KMT [177]. They implemented a structurally-guided design of SAM analogues based on the crystal structure of endogenous vSET to identify limiting or advantageous hydrophobic interactions and hydrogen bonds. After analysing the structure, they hypothesised that mutations to L13, N119, Y109 or L116 to glycine or alanine would result in a "hole" cavity able to fit the modified cofactors onto the mutated vSET. Regarding the cofactors, apart from the already profiled N6-benzyl SAM, they introduced groups in the C2' and C3' positions and tested them with Y109G and L116A vSET. After this they synthesised the orthogonal mutant vSET/SAM analogue pairs and characterised some of these by calculating the k_{cat} and K_m. As a result, they found that the specificity (k_{cat}/K_m) of 2', 3'-dibenzyl SAM (see Figure 3-12, B) for vSET L116A was comparable to that of the endogenous pair, although the affinity of this enzyme to SAM was even higher. The small number of enzymes and cofactors analysed means that many other unknown pairs might exist with increased activity.

In addition to these studies, which were primarily limited to small mutations of SAM, the Song and the Thompson laboratories [178,179] used 5'-N-Iodoethyl/5'-Aziridine SAM analogues (see Figure 3-12, D). These cofactors have demonstrated activity for small molecule methyltransferases [180], and the application for the bump-and-hole approach was analysed using PRMT1, CARM1, DOT1L, G9a and SUV39H1. However, a crucial limitation of this approach is that aziridine SAM analogues are not stable at physiological pH, so this approach does not seem very appropriate for *in vivo* profiling.

The use of protein libraries to evaluate the substrate targets of the mutant histone methyltransferase (HMT) do not consider the structure of chromatin nor other PTMs that can be present. This is an important gap, as in order to localise their specific target, HMTs

do not only rely on a certain sequence motif, but also on the PTMs already present on the chromatin, and appear as part of multimeric complexes binding the chromatin [140]. For example, the nuclear receptor SET domain-containing (NSD) family of lysine methyltransferases, more specifically NSD2, shows a different target preference depending on the available substrate, methylating H3K36 with nucleosome substrates (i.e. histones and DNA) but H4K44 for histone octamer substrates (i.e. in the absence of DNA). Thus, the lack of cellular chromatin assay conditions might affect the specificity of any given HMT, and therefore profiling results.

The most successful profiling approach of methylome activity of PMTs so far consists of a substitution of SAM's methyl group to an alkyl group, which can be easily modified via "click" chemistry methods to identify the contribution of a given PMT [169]. Figure 3-12 shows this profiling approach (Figure 3-12, C) and the previously mentioned SAM analogues used to apply the bump-and-hole approach for protein methyltransferases.

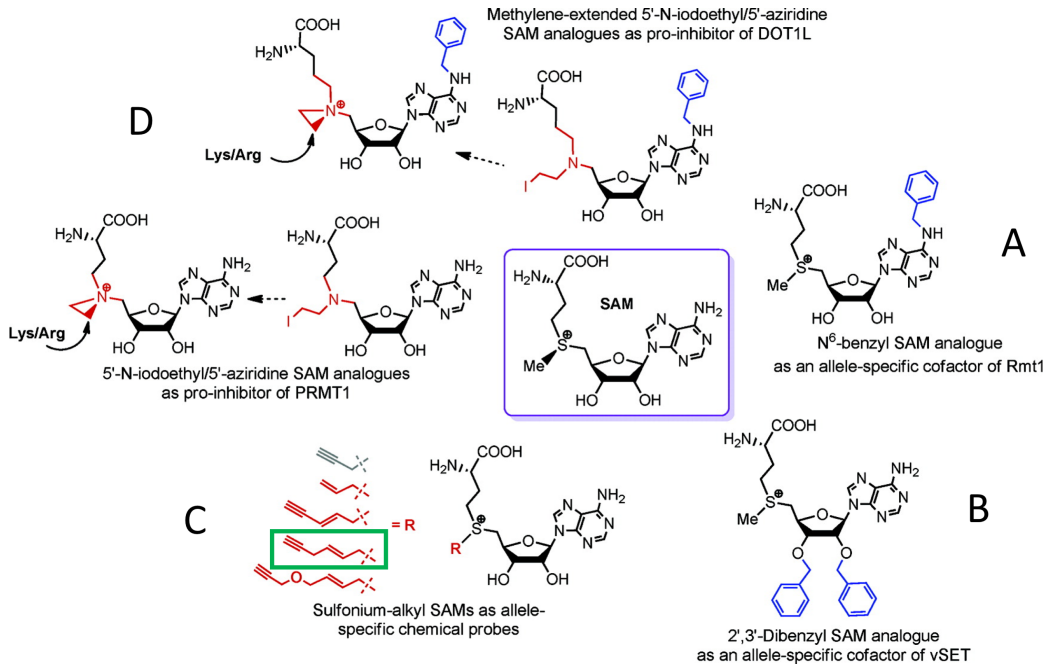


Figure 3-12. SAM analogues acting as cofactor alternatives for PMTs, with Hey-SAM highlighted in green [140]. Reprinted with permission from Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. ACS Chem. Biol. 7, 443–63 (2012). Copyright 2012 American Chemical Society.

Islam et al. at Luo's lab developed several SAM analogues and discovered that (E)-hex-2-en-5-ynyl SAM (Hey-SAM) was a suitable substrate for engineered G9a (see Figure 3-12, C), meaning that this Hey-SAM analogue paired with the mutant G9a Y1154A modified

some of the substrates of the endogenous pair with certain efficiency. The substrates are subsequently modified firstly using Cu(I)-catalysed click chemistry to allow for substrate purification and then MALDI-MS to detect alkyl modifications. Later, Wang et al. (also at Luo's lab) designed a novel technology named CliEn-seq to allow deciphering the *in vivo* activities of individual PMTs [137]. However, the SAM cofactor analogues presented low cell membrane permeability. CliEn-seq overcame this limitation by implementing an *in vivo* synthesis of SAM analogues from the cell-permeable methionine analogues using an engineered SAM synthetase, such as methionine adenosyltransferase (MAT). This approach provided *in situ* chromatin modification by the engineered lysine methyltransferases G9a and GLP, which subsequently could be analysed by "click" chemistry enrichment and DNA sequencing of the alkylated chromatins. This assay was further developed evaluating 40 human MAT mutants against 7 S-alkyl analogues of methionine (SAAMs) by Wang et al. [181]. The strategy was also applied to the protein arginine methyltransferase 1 (PRMT1) by Thomsen et al. [182]. A schematic representation of the *in vivo* approach to profile the methylation activity of PMTs compared to the native mechanism is shown in Figure 3-13.

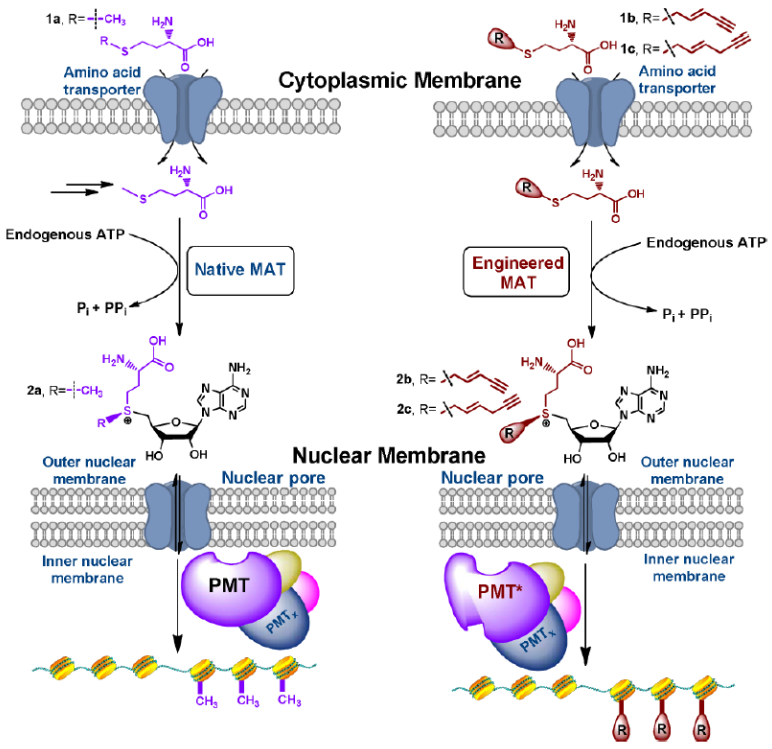


Figure 3-13. Scheme of the chemoenzymatic strategy to profile genome-wide chromatin modifications via the formation of the S-alkyl analogues at *in vivo* conditions [137]. Reprinted with permission from Wang, R. et al. Profiling Genome-Wide Chromatin Methylation with Engineered Posttranslation Apparatus within Living Cells. J. Am. Chem. Soc. 135, 1048–1056 (2013). Copyright 2013 American Chemical Society.

The aforementioned work was expanded by Wang et al. [183], who developed a different clickable SAM analogue cofactor, 4-propargyloxy-but-2-enyl SAM, which demonstrated activity when coupled with PRMT1 mutants. One pivotal limitation of this approach is that, even when the transfer rates of the alkyl groups are faster than for the endogenous PMTs, they are still slow, probably because of the steric hindrance at the SAM binding pocket. Another cofactor analysed for PMT profiling was propargylic Se-adenosyl-L-selenomethionine (ProSeAM) by Luo and colleagues [184,185], substituting a sulphur with selenium. This cofactor was reported to be able to introduce alkyl groups on both lysine and arginine residues, and the modifications were then clicked to a biotin tag to allow for enrichment.

However, there are considerable limitations that make all the S-alkyl SAM methods listed above not adequate for the profiling of the PMTs activity, particularly at *in vivo* conditions. First, the SAM analogues used by this method show basal activity for native HMTs, and thus the identity of the identified enzymes performing the marks is arguable. Second, the modified alkyl group is not structurally comparable to the methyl group, and thus it cannot be recognised by the chromodomains and other reader domains (Tudor, PHD...) of PTM-binding proteins. This means that the subsequent epigenetic purposes of these marks are lost, as they cannot be decoded by chromatin regulators. A true cell-based model should be able to discern not only the methylome targets of KMTs but also the downstream effects of these modifications. To achieve this, an understanding of the principles of protein- cofactor interactions and protein folding is required to engineer an effective bump-and-hole pair.

3.4. Motivation and approach

The main objective of this project is the development of a mass spectrometry-based technology for profiling the targets of different KMTs. Previous work in this area has involved the substitution of the methyl group by a "clickable" alkyl group [140], however this change in donor group chemistry disrupts the downstream function of a methylation event, as alkyl groups are not reported to be bound by chromodomains.Our proposed strategy is to identify existing structural differences between the catalytic SET domains of different KMTs. The goal is to search for unique "holes" in KMT structures that can accept a complementary "bump" on a synthetic SAM cofactor analogue, so that it binds selectively with a KMT or small subclass of KMTs. The corresponding bumped SAM analogues will also contain a heavy methyl donor group (e.g. CD_3) so that a cellular methylation event is biologically consistent and detectable as a mass shift by mass spectrometry.

Our approach starts with a computational comparison of the geometry of KMT structures that takes aligned KMT-cofactor structures as input and identifies pockets that are unique to one of the KMTs. The output of the program elucidates design points for incorporation of a synthetic "bump" on a matching SAM analogue. SAM analogues are then designed based on the steric properties of the binding pocket holes. Then, the candidate analogues are computationally screened for binding mode suitability using molecular dynamics. The highest-ranked SAM analogues are then synthesised and used in an *in vitro* assay to demonstrate selective utilization by the KMT of interest, where mass spectrometry (MS) proteomics is the primary readout for validating catalytic activity. Our approach is summarised in Figure 3-14.

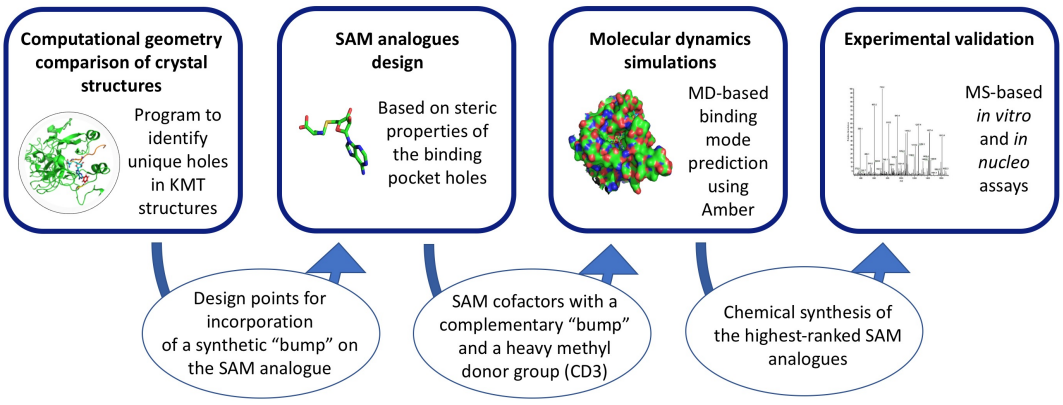


Figure 3-14. General overview of the approach to develop a MS-compatible method to profile the targets of different KMTs.

In summary, this thesis focuses on different aspects or extensions of this project distributed into three distinct chapters. Chapter 4 reports on the design of SAM cofactors which are selective to endogenous KMTs based on our computational comparison strategy for identifying unique "holes" in their structures. The proposed bumped SAM cofactor analogues are then validated experimentally for their selectivity towards SETD7 using *in vitro* assays and mass spectrometry. Chapter 5 presents an *in silico* screening tool that uses molecular dynamics simulations to rank bumped SAM cofactors for synthesis based on their compatibility with the target KMT of interest. Chapter 6 then extends this work to investigate a bump-and-hole strategy using G9a as a case-study, where bioinformatics analyses of G9a provide insight into candidate mutations that could make G9a selective towards the bumped SAM analogues described in Chapter 4. Together, this thesis presents a novel approach for the identification of KMT targets.

Chapter 4

Design of SAM cofactor analogues selective to endogenous KMTs

The epigenetic drugs and diagnostic technologies market is currently expanding, with many pharmaceutical companies focused on epigenetic targets [54]. One of the main reasons for this growth is the correlation between aging and a higher incidence of cancer and the existing rise in the elder population [120]. To address this need, governments are encouraging the development and adoption of early disease detection and diagnosis methods. For example, the Epi ProLung test for lung cancer diagnosis received a €2.77m Horizon 2020 grant from the European Commission [186].

The epigenetic drugs that are currently FDA-approved and commercially available include the DNMT inhibitors Dacogen (decitabine) by Otsuka Pharmaceutical and Vidaza (azacitidine) by Celgene for treating Myelodysplastic Syndrome (MDS) [54]. Other approved drugs include Istodax (romidepsin) from Celgene and Zolinza (vorinostat) by Merck & Co, both HDAC inhibitors to treat Cutaneous T Cell Lymphoma (CTCL) [54]. Besides, the epigenetics market is aiming to expand beyond their oncologic use, and could expand to incorporate the treatment of other diseases such as arthritis and Alzheimer [187] in the near

future. However, there exist some limitations to progress epigenetic drugs through the phase II/III of clinical trials, the main one being off-target effects [94].

It is currently problematic to assess how lysine methyltransferase (KMT) activities are modified in a specific disease state. For example, whether a specific KMT alters its methylation activity towards a specific histone or non-histone protein. KMTs catalyse the transfer of a methyl group from a S-adenosyl methionine (SAM) cofactor to a specific lysine target residue. Since all KMTs utilise the same SAM cofactor and many can methylate a given target residue, there is currently no way to directly assess which KMT performs an observed methylation event. For example, when analysing the turnover of histone methylation in MS-based cellular assays, the outcome of the analysis is the cumulative activity of all KMTs and individual enzyme contributions are not easy to discern. Being able to identify the specific targets (both histone and non-histone) of a given KMT is important because it would improve the overall understanding of the epigenetic mechanism of gene expression and chromatin-related processes within the context of disease, which could guide the development of more targeted treatments.

To address this problem, we propose a mass spectrometry-compatible technology that designs specific SAM cofactor analogues to be selective towards particular KMTs based on structural differences in their binding pockets. To achieve that, we combine an *in silico* approach to engineer synthetic cofactors selective for a given KMT, based on structural differences, with *in vitro* assays and mass spectrometry (MS) proteomics for target identification. Together, the combination of these approaches represents a novel KMT "methylome" profiling assay for identifying the targets of specific KMTs.

4.1. KMT target identification methods

4.1.1. MS-based methods

From a MS perspective, KMT target identification methods rely on the knockdown of the activity of particular KMTs by using RNA interference (RNAi) with double-stranded RNA (dsRNA) molecules, alongside hmSILAC or a general quantitative analysis to track methylation changes.

RNAi is a research tool based on the inhibition of mRNA molecules by complementary dsRNA molecules, that are typically short hairpin RNA (shRNA) or small interfering RNA (siRNA) [188]. The synthetic dsRNA can selectively suppress the genes of interest, and therefore it is currently used to assess the function of a specific protein by causing a dramatic decrease in its expression. Although this technique has been widely used [74,189], it presents some important disadvantages. Firstly, it does not totally suppress the expression of the protein as it causes a "knockdown" of the expression instead of a complete "knockout" [190]. In addition, sometimes the introduced dsRNA has a base sequence that can pair with multiple mRNA transcripts, thus reducing the expression of multiple genes (i.e. off-target effects) [191].

Stable isotope labeling with amino acids in cell culture (SILAC) is a technique used for mass spectrometry (MS)-based quantitative proteomics, based on the labelling of proteins by the incorporation of amino acids with stable isotopes of their atoms (e.g. deuterium, ^{13}C, ^{15}N). In brief, this technique consists on growing two different cell populations in the same culture media except for the fact that one of the media contains a "heavy" form of a specific amino acid. This amino acid is incorporated to the newly synthesised proteins, replacing the endogenous form after several cell divisions. The proteins from both cell populations are then combined and differentiated in a mass spectrometer based on their mass difference [192]. The properties and activity of the proteins synthesised with this method behave exactly as the non-isotopic labelled; therefore, no function-loss occurs.

Although SILAC is mainly used for quantifying protein levels, a variation of SILAC known as heavy methyl SILAC (hmSILAC [193]) is used for the identification of Lysine and Arginine methylation events. This technique consists on adding $^{13}CD_3$-Methionine to methionine-

depleted media. Once inside the cell, the $^{13}CD_3$-Methionine is metabolised into S-adenosyl Methionine (SAM). This heavy SAM is then utilised by methyltransferases to methylate histone or non-histone targets with heavy methyl groups [194]. The peptides generated this way can be identified by MS by looking for peak pairs of the heavy and light versions of the peptide.

The hmSILAC technique has been widely used to characterise methylation sites. Sites of lysine and arginine methylation have been identified combining hmSILAC with high-resolution mass spectrometry-based proteomics. Using this approach, Bremang et al. identified 501 different methylation sites on 139 unique proteins [195]. Some other studies using this approach include the study the dynamic turnover of H3K9 tri-methylation in pericentric chromatin [196] and profiling the turnover of histone lysine methylation, revealing that the rates of formation for mono-, di-, and tri-methylation events were progressively slower [197].

The hmSILAC protocols are typically combined with some method for enrichment of the methylated peptides using, for example, antibodies [195] to improve MS identification within a complex protein background. The antibodies used can be generally classified as pan-antibodies or sequence-specific antibodies [192]. Pan antibodies are sequence independent, meaning that they can identify the PTMs independently of its surrounding sequence, whereas sequence-specific antibodies can recognise the surrounding sequence and therefore can be utilised to identify specific sites. In any case, the use of histone PTM antibodies for the identification of PTMs in histones presents important issues affecting their performance: off-target recognition, influence by neighbouring PTMs, and the inability to distinguish the modification state on a specific residue (e.g., mono-, di-, or tri-methyl lysine) [198]. Therefore, all these issues also affect the enrichment of methylated peptides when using hmSILAC.

ShRNA/siRNA knockdown of a PMT enzyme in combination with hmSILAC has become a common tool to identify PTM methylation sites. For instance, O'Connor et al. [189] used this technique to study the human cytomegalovirus (HCMV) infection. More specifically, they used shRNA to create a DOT1L knockdown cell population alongside hmSILAC to confirm the histone methylation fluxes. This method enabled the discovery of a 10-fold growth defect for the knockdown cells when compared to the control cells without knockdown. More recently, Olsen et al. extended this approach to the identification of non-histone targets. The group identified a total of 1032 Kme1 sites in esophageal squamous cell

carcinoma (ESCC) cells and 1861 Kme1 sites in ESCC cells overexpressing SMYD2 using SILAC coupled with immunoaffinity enrichment of mono-methyl-lysine (Kme1) peptides and mass spectrometry. Then, they performed the shRNA-mediated knockdown of SMYD2 to find that 35 of the Kme1 sites found were potently down-regulated by this knockdown [199].

4.1.2. Chemical biology-based methods

Alongside the MS-based methods mentioned in section 4.1.1. , chemical biology efforts have been utilised to dissect the targets of KMTs. Photoreactive- and fluorescent/biotin-coupled probes based on the KMT inhibitors have been used to enrich and identify KMT methylation targets [200]. In these chemical biology-based methods, an inhibitor for a specific KMT is modified by inserting a cross-linking group (and an enrichment handle) to cross-link the KMT to its protein target, therefore revealing the substrates of the KMT. As an example, Horning et al. proposed several SAH-based photoreactive probes to profile and enrich several methyltransferases (>50) for chemical proteomic experiments on human cancer cell lysates, obtaining a remarkable specificity over other classes of proteins. They also used the SAH-based probes to screen the selectivity of MT inhibitors, leading to the discovery of a covalent inhibitor of nicotinamide N-methyltransferase (NNMT), an enzyme related to cancer [200].

Moreover, Zhu et al. designed and synthesised two photoreactive and "clickable" affinity-based probes (AfBPs) structural mimics of FED1, a previously discovered *in vitro* but not *in situ* inhibitor of DOT1L, finding that both FED1 mimics were not able to reach the cell nucleus where DOT1L performs its activity, therefore explaining why FED1 behave as a DOT1L inhibitor *in vitro* but not *in situ* [201]. This was followed by a large-scale cell-based proteome profiling followed by quantitative LC-MS/MS to identify cellular off-targets of FED1, finding that the putative ribosomal RNA methyltransferase NOP2 was a genuine off-target of the FED1 inhibitor.

As an example of a method that did not utilise photoreactive probes but just inhibitors, Ahmed at al. applied a bioinformatic approach integrating cellular, structural, biochemical and proteomic data to identify novel substrates of the lysine methyltransferase SMYD2. For confirmation, they treated the samples with a selective SMYD2 inhibitor (BAY-598). The

addition of the inhibitor abolished the methylation signal, and they used this result to prove that the methylation found was dependent on the catalytic activity of SMYD2 [202].

4.1.3. Technical challenges and opportunities

As seen in section 4.1.1. , current MS-based methods to identify KMT methylation sites have contributed to the current knowledge about the activity of these enzymes. However, they also present some important limitations. In the first place, the knockdown of the KMT of interest can perturb multiple (including unknown) biological pathways. More specifically, manipulation of a KMT can affect gene expression, which subsequently changes the protein levels (e.g. of other KMTs) and can therefore make quantitative comparisons subjective. For this reason, a methylation change cannot be directly attributed to the knockdowned KMT. Other important limitation of current MS-based methods is the technical complexity related to the immunoaffinity enrichment of methylated peptides with specific antibodies [203].

For all the reasons discussed above, current standard methods do not provide a direct technology to identify KMT targets. The work presented in this chapter addresses this problem by proposing a new KMT "methylome" profiling assay able to identify the targets of specific KMTs in three steps: *in silico* identification of structural differences in the binding pocket of KMTs to engineer selective KMT cofactor analogues, *in vitro* assays and mass spectrometry (MS) proteomics for target identification. Further details about our approach are provided in section 4.2. .

4.2. Methodologies

The proposed methodology focuses on identification of mutually exclusive holes (utilised selectively by wild type KMTs with naturally occurring holes), design and synthesis of complementary SAM cofactors and experimental validation of their specificity towards a specific enzyme.

4.2.1. Overview of proposed design methodology

Our proposed strategy consists of identifying existing structural differences between the catalytic SET domains of different KMTs that could selectively accommodate synthetic SAM cofactors. More specifically, this approach aims to identify unique "holes" in KMT structures to then design synthetic SAM cofactors with a "bump" matching the identified "hole" of the KMT so that it binds selectively to it. The matching SAM analogues also contain a heavy methyl donor group (e.g. CD_3-SAM) so that a methylation event is both biologically consistent and detectable by MS proteomics.

Regarding the computational design approach, KMT structures available in the PDB (n >= 2) are aligned based on their SAM/SAH cofactor orientation. The aligned PDB file is then used as input to a program that identifies pockets that are unique to one of the analysed KMTs. The output of the program helps assess design points for incorporation of a synthetic "bump" on the SAM analogue. The basis of this program is specified in section 4.2.1.1. .

The proposed SAM analogues are then chemically synthesised by our experimental collaborators (Katie Addison, Dr Matt Fuchter; Chemistry, Imperial College London), and the recombinant KMT enzymes are either commercially obtained or expressed and purified using His-tag / FPLC methods. The enzyme-SAM analogue activity is then assessed experimentally through *in vitro* assays. The activity that the enzymes present with SAM analogues is compared with the one with native SAM to demonstrate selective utilisation by the KMT of interest. Mass spectrometry (MS) proteomics is the primary analytical technique for assessing catalytic turnover of the SAM analogues by measuring histone methylation as an endpoint.

4.2.1.1. Computational identification of natural KMT pockets

To identify naturally-occurring pockets in KMT structures, a program was implemented in C++ that utilises the Easy Structural Biology Template Library (ESBTL). This is a C++ library that allows the reading, manipulation and exporting of protein crystal structures in the PDB format and thus provides a platform for computational structural analysis [204]. This library effectively contains an efficient PDB parser and data structures for the geometric analysis of biomolecules in C++. Briefly, the data structures in the ESBTL library are hierarchical and are organised as follows: each biological structure is represented as a "System" object, which is composed of multiple "Model" objects. From this level onwards, the hierarchical class organisation follows a general PDB data structure with each "Model" containing "Chain", "Residue" and "Atom" objects. Access to the various objects is provided using customised ESBTL iterators and class methods are available to retrieve class members.

The input to our C++ program is a PDB file containing two or more KMT-cofactor crystal structures aligned taking the adenosyl ring of SAM as reference for the alignment and an input file specifying the "design atom" on SAH (i.e. to which atom on SAH to add a "bump"). The program proceeds in four conceptual steps: (1) discretisation of the coordinate space of the PDB file into a 3D mesh comprised of small grids (e.g. 0.25 Å cubes), (2) determining the grid occupancy of every atom based on its atomic radius [205], (3) sequentially building a carbon atom chain (straight line, bond length of 1.54 Å) starting from the design atom on SAH and checking for steric clashes with KMT atoms, and (4) exporting the results in PDB format and clash statistics in text format. A visual representation of the first three points is shown in Figure 4-1. These results provide important insight for the identification of suitable design points for incorporation of a synthetic "bump" on the designed SAM analogues.

Briefly, the program uses the ESBTL classes to read and store the input PDB file, which consists of two or more methyltransferases aligned taking their SAH residues as the reference for the alignment. The ESBTL library provides customizable iterators to access the PDB information stored in the hierarchical class structures, consisting of system (which consists of one class instance since only one PDB file is read), models, chains, residues and atoms. Through this class architecture, the program can define restricted iterators which

only loop over the residues or atoms of a select KMT protein (stored as "chains"), and also loop over only the atoms of the corresponding SAH molecule (which is stored as a "residue" in the respective protein "chain").

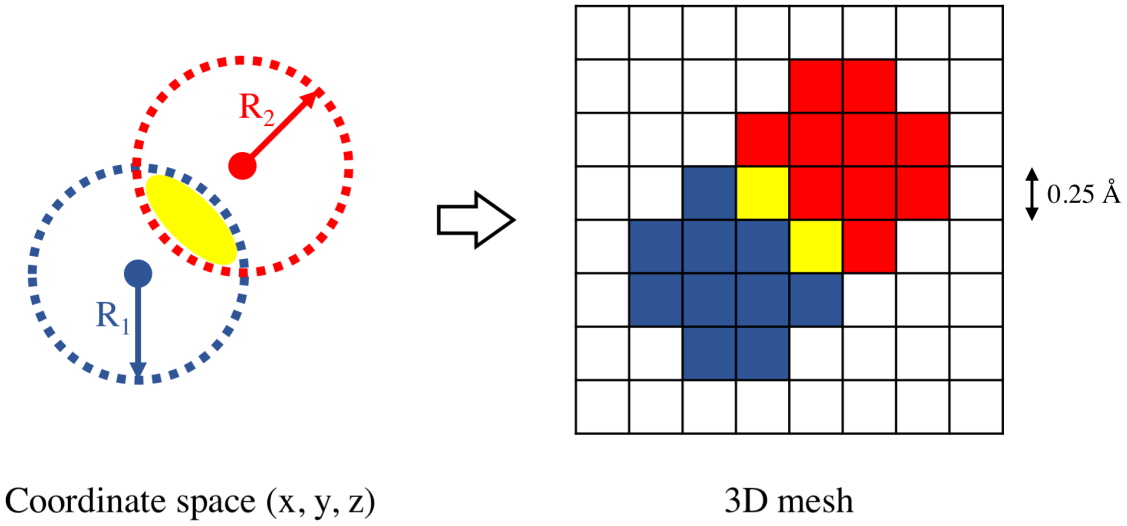


Figure 4-1. Mesh discretization of the coordinate space and grid occupancy of every atom. Steric clashes between the KMT and the inserted atoms are shown in yellow.

In the first step of discretization of the coordinate space (Figure 4-1), a 3-dimensional grid is constructed around a pre-specified radius (typically 10-20 Å) of the SAH molecule. Specifically, the x, y and z coordinates for all the protein atoms within the pre-specified radius of any SAH atom are appended to separate lists for the x-, y- and z-dimensions, respectively. The x-, y- and z-coordinate lists are then independently sorted, and the minimum and maximum coordinate value for each dimension define the overall range of the 3-dimensional grid.

Each dimension of the 3-dimensional grid is then discretised into 0.25 Å segments, which results in 0.25 x 0.25 x 0.25 Å cubes, each of which is represented by a customised C++ class. The algorithm then loops over all atoms in each of the KMT residues, computes the volume occupied by the atomic radius of that atom, and the corresponding information for the atom (i.e. atom type, residue name and number, KMT etc.) is then stored in all the 3-D cubes it occupies (see Figure 4-1 for a 2-D illustration). Storing the data in this way enables the algorithm to quickly access which atoms occupy a position in space with a resolution of 0.25x0.25x0.25 Å. Thus, when a "bump" is appended to the design atom on SAH, the program can quickly access what KMT atoms it might result in a steric clash with.

The algorithm then computes what direction to add a new covalent bond to the design atom on SAH. To compute this, a parameterised line is drawn between the neighbouring atoms that form a covalent bond on either side of the design atom. The direction to add a new covalent bond to the design atom is then determined by identifying another line segment (which we will refer to as the "search direction") that passes through the design atom and is orthogonal to this parametrised line. Given this "search direction", any atom or functional can be covalently appended to the design atom, and the program can directly query for steric clashes with the KMT structure and report the specific residues/atoms. Due to the characteristics of the calculation of the parametrised line, the current version of the program is limited to design atoms that are bound to exactly two other atoms. For simplicity, we only considered the addition of a number of carbon atoms (specified by the user) to the SAH design atom using an average carbon-carbon bond length of 1.54 Å.

The output of the program is a PDB file containing all the KMT atoms within the distance threshold of the SAH molecule, including the bumped SAH to visualise the placement of the "bump" into the design atom on SAH. In addition, the program reports the specific details of any clashes between the "bump" on SAH and the KMT atoms, where the number of cubes involved in a steric clash are proportionate to the volume of the clash (e.g. see yellow squares in Figure 4-1).

4.2.2. Materials and methods

4.2.2.1. G9a expression

Frozen cultures of BL21 *E. coli* containing G9a plasmid (EHMT2 plasmid #25503 on Addgene, with vector p28a-LIC, containing a His-tag (N terminal on insert)) in DI water and 10% glycerol were thawed. These stocks were prepared by a former member of the DiMaggio lab (Dr Ziran Da), according to the heat shock transformation protocol. This method consisted in adding 2 µL of the aforementioned plasmid to 50 µL of $CaCl_2$ competent BL21 *E. coli* cells in a 1.5 mL Eppendorf. The sample was mixed and left on ice for 30 minutes, then given a heat shock of 42 °C for 1 minute. The vials were then put on ice for 2 minutes. After that, 200 µL of Hanahan's Broth (Sigma) without kanamycin were

added to allow cells to recover. The mixture was then incubated in the shaker for 1 hour at 37 ºC. After this time, the mixture was transferred to a 15-mL falcon tube containing 3 mL of Hanahan's Broth with 50 μg/mL of kanamycin (Sigma). The sample was left overnight in the 37 ºC shaker. To prepare the glycerol stocks, 320 μL of the overnight culture were mixed with 80 μL of autoclaved 50% glycerol (Sigma), mixed well, flash frozen in liquid nitrogen and stored at -80 ºC.

SOB medium was prepared by mixing 28.1 g of Hanahan's Broth (Sigma) with DI H_2O for a total of 1L of medium. A 200 g/L dextrose (D-glucose, Sigma) stock solution was also prepared. In order to sterilise and avoid contamination by the presence of other proteins or DNA, both SOB and dextrose solutions were filtered by passing them through 0.22 μm PES filters (Stericup Millipore Express Plus, Millipore) and autoclaved during 50 minutes at 125 °C, alongside all the Erlenmeyers used for *E. coli* growth. Fresh stocks of 50 mg/mL of kanamycin (Sigma) and 50 mg/mL of IPTG (Sigma) were also prepared.

A starter culture was generated by mixing 40 mL of SOB medium, 1.6 mL of glucose stock, 40 μL of kanamycin stock and 200 μL of thawed BL21 *E. coli* glycerol stock. This starter culture was then incubated for 2 hours at 37 °C and 250 rpm on a Stuart S1500 incubator. Regarding the final culture, and to ensure correct aeration in the Erlenmeyer flasks, 2 L flasks were used for each 500-mL culture. For each autoclaved Erlenmeyer, cultures were prepared by mixing 469.5 mL of SOB medium, 20 mL of glucose stock, 500 μL of kanamycin stock and 10 mL of the previously generated started culture to complete 500 mL. The final cultures were then incubated at 37ºC and 250 rpm on a Thermo Scientific MaxQ 8000 incubator. We carried out the cultures at 37°C to increase the growth rate until the OD for induction was reached.

In order to get a proper protein expression using BL21 *E. coli*, it is necessary to grow the liquid cultures to mid-logarithmic (midlog) phase before adding a gene transcription-inducing agent, typically IPTG. Expression must be induced when the OD600 of the cells in the culture is between 0.5 and 1 [206], as this range of OD600 represents the midlog phase. Measurements of the OD600 of the culture were taken in the Pharmacia Biotech Ultrospec 2000 UV/Visible spectrophotometer. These measurements were taken by diluting 0.25 mL of the culture in 0.75 mL of SOC medium. SOC medium was also used as blank. To calculate the OD of the cells in the culture, the formula applied is shown in equation (4-1):

$$OD_{cells} = \left(OD_{sample} - OD_{blank}\right) \cdot 4 \quad (4\text{-}1)$$

Measurements were repeated every 5 minutes until the value of OD_{cells} was close to 0.6. Once that OD value was reached, the cultures were taken out of the incubator and cooled to the induction temperature (15 °C) in ice-water. The cultures were then induced by adding 2.38 mL of IPTG stock solution to reach a final concentration of IPTG of 1 mM. After the addition of IPTG, the cultures were incubated overnight at 15 ºC in a Thermo Scientific MaxQ 8000 incubator. The reasoning behind using 15 ºC and not 37 °C for the expression was to avoid G9a aggregation and precipitation in the culture. After culturing overnight, the cells were pelleted by using the centrifuge Thermo Scientific Heraeus Megafuge 40R at 4800g and 4ºC for 10 minutes, and the supernatant was discarded.

4.2.2.2. G9a purification

In order to purify the previously expressed G9a, a series of buffers were prepared:

- Lysis buffer with protease inhibitors: phosphate-buffered saline (PBS) pH 7.4 (Invitrogen), 0.25 M NaCl (Sigma), 5 mM imidazole (Sigma), 1 mM DTT (Sigma), 5% glycerol (Sigma), 0.1mM PMSF (Sigma).
- Wash buffer: 20 mM Tris-HCl buffer (Sigma/Amresco) at pH 8.0, containing 250 mM NaCl, 50 mM imidazole and 5% glycerol.
- Elution buffer: 20 mM Tris-HCl at pH 8.0, 250 mM NaCl, 250 mM imidazole, 5% glycerol.
- Dialysis buffer: 10 mM Tris-HCl at pH 7.5, 500 mM NaCl.

The cell pellets obtained in section 4.2.2.1. were thawed on ice and resuspended in lysis buffer with protease inhibitors. The cells were lysed on ice by sonication (Fisher Scientific Model 505 Sonic Dismembrator with probe CL-334) at a frequency of 20 kHz, with 10 cycles of 10 seconds followed by 30s for cooling for a total of 400s. The crude extract was then cleared by centrifugation at 18000xg for 20 min at 4ºC (centrifuge Thermo Scientific Heraeus Multifuge X3R). The supernatant from the centrifugation (containing recombinant G9a protein) was placed on ice and transferred to a cold room (4ºC) for purification.

The G9a recombinant proteins were purified using a two-column separation sequence. First, a DE52 anion exchange cellulose column (DEAE Sepharose Fast Flow, GE Healthcare) was prepared by mixing 10g of it in 50ml of lysis buffer, and washed with 30-40ml lysis buffer. The main reason for using this column was to eliminate DNA from the sample and therefore reduce its chemical complexity before adding it to the Ni-NTA column. The second column was a Ni-NTA affinity chromatography matrix column (QIAGEN); this column was prepared by mixing 3mL of the resin with 3 mL of lysis buffer, then washing with 20-30ml of lysis buffer. Both resins were packed in Econo-Column Chromatography Columns (BIO-RAD).

The G9a supernatant was loaded onto the DE52 column and the flow through was collected. Then 25mL of lysis buffer (with protease inhibitors) were also passed through and collected. These fractions were then loaded onto the Ni-NTA column to bind the poly-His-tagged G9a. The column was then washed with 10 column volumes of wash buffer, and the protein was eluted with elution buffer (3-5 mL) in volumes of 0.5 mL. To assess if the protein had been completely eluted, an analysis was carried out by adding 10 µl of each fraction of Ni-NTA elution to 100 µl of 1X Bradford reagent. The sample collection was stopped when the resulting blue colour of the assay became lighter.

The elutions containing G9a were then dialysed overnight at 4ºC in 3L dialysis buffer. The filter used for dialysis was Fisherbrand Dialysis Tubing with nominal MWCO 6,000-8,000 (Fisher Scientific). After overnight dialysis, the sample was purified using FPLC using dialysis buffer as mobile phase, in collaboration with Fabio Fisher (Baum Lab, Imperial College London). The FPLC clean-up was carried out in a GE Hiload superdex 16/600 column, loading 5ml of the G9a solution at a flow rate of 1ml/min, beginning fractionation after 0.2 column volumes. After FPLC purification, the concentration of the sample was determined by NanoDrop (Thermo Fisher). Purified recombinant proteins were aliquoted adding 20% glycerol (Sigma), flash frozen in liquid nitrogen and stored at −80 °C until used. All proteins were judged to be > 90% pure by SDS-PAGE.

4.2.2.3. Protein gel analysis

12% SDS-PAGE gels were prepared to analyse the amount and purity of G9a. The samples were mixed in equal volumes of DTT 54 mg/mL in Laemmli Buffer and denatured at 95 °C

for 5 minutes prior loading. The gel was run on BIO-RAD Mini-PROTEAN Tetra System at 175V for 60 min. Once the electrophoresis finished, the gel was cleaned with DI water and left dying overnight in coomasie blue stain (1% Brilliant Blue G-250 (Fisher Scientific), 10% AcOH (Sigma), 50% MeOH (Fisher Scientific) and 40% DI H_2O) and then destained in destaining buffer (20% AcOH (Sigma), 40% MeOH (Fisher Scientific) and 40% DI H_2O).

4.2.2.4. G9a in-solution digestion and MS identification

To assess the contents of the purified G9a sample, an in-solution digestion of the sample was performed. To do so, a 50 µL volume of the G9a-containing solution was dried in the SpeedVac Concentrator (Thermo Scientific Savant SC210A) at low drying rate down to ~5 µL. The pellet was then suspended in 50 µL of 100 mM ammonium bicarbonate (NH_4HCO_3) and sonicated for 5 minutes to ensure that all the contents were in the solution. After checking that the pH was ~8, DTT was added to a final concentration of 5 mM. The resulting mixture was then incubated at 51 °C in a heating block for 1 hour. Once it was back to room temperature (RT), iodoacetamide (IA) was added to a final concentration of 14 mM. After checking that the resulting pH was ~8, the sample was incubated for 45 minutes in the dark at RT. The reaction was then quenched by adding DTT to a final concentration of 0.5 mM. Trypsin was then added in a ratio of 100:1 protein : trypsin and the mixture was incubated at 37 °C for 6 hours. After that time, the reaction was quenched by adding glacial acetic acid to ensure the final pH was below 3. The resulting sample was then dried in the SpeedVac Concentrator (Thermo Scientific Savant SC210A). The sample underwent stage tipping to prepare it for MS analysis as described in section 4.2.2.6. .

A general MS protein identification method was performed to assess the quality and purity of the recombinant protein stocks. Real-time tandem mass spectra were acquired on an LTQ Velos Pro linear ion trap (Thermo Scientific) coupled to a nanoLC running a general 120-min gradient. Nanospray ionisation of the eluent was performed over a mass range of 300–2000 m/z followed by data dependent CID fragmentation using a normalised collision energy of 35 V.

Proteome Discoverer 1.3 (Thermo Scientific) was used for identification of the MS/MS spectra. Briefly, the SEQUEST algorithm was used to match tandem MS data against an *in silico* digest of the Uniprot database containing a modified version of G9a, only including

the residues of the expressed protein (N913-D1193). The search was performed with the following non-default specifications: mass limits 350-5000 Da; activation type CID; unrecognised charge replacements 2,3 and 4; precursor mass tolerance of 2.5 Da; fragment mass tolerance of 0.8 Da; dynamic side chain modifications of carbamidomethylation of cysteines and oxidation of methionines.

4.2.2.5. *In vitro* methyltransferase assay

Reactions were performed at a total volume of 200 µL in KMT buffer. KMT buffer was composed of 100 mM TRIS pH 8.8 (Sigma/Amresco), 10 mM $MgCl_2$ (Sigma), and 4 mM dithiothreitol (Sigma). Commercial HeLa native nucleosomes were obtained from BPS Bioscience (product number 52039) at a concentration of 0.85 mg/mL in a buffer containing 10 mM Tris-HCl pH 7.2, 5 mM NaCl, 1 mM $MgCl_2$, 0.5 mM $CaCl_2$, 125 mM EDTA, and 10 mM Sodium Butyrate, with purity ≥80% by SDS-Page.

The recombinant G9a used in these assays was expressed and purified as described in sections 4.2.2.1. and 4.2.2.2. . Recombinant SETD7 was obtained from Abcam (product number ab51285) at a concentration of 1 mg/mL in a buffer containing 20% glycerol, 0.2M sodium chloride, 50mM Tris HCl pH 7.5, and 5mM DTT, with purity > 95 % by SDS-PAGE. Both G9a and SETD7 were used at 200, 500 or 1000 nM depending on the experiment as specified in the results sections 4.3.2.4. and 4.3.2.5. .

The KTM buffer, enzyme and nucleosome solutions were mixed and preincubated during 5 min at 37 °C to get the solution at that temperature. CD_3-SAM or CD_3-SAM cofactor analogues (synthesised by Katie Addison, Chemistry Department, Imperial College London) were then added and mixed with the preincubated mixture at concentrations of 200, 500 or 1000 µM depending on the assay, as specified in the results sections 4.3.2.4. and 4.3.2.5. .

Reactions were incubated on a 37 °C heating block for 1 hour. After that time, the reaction was stopped by adding 400 µL of 0.6N H_2SO_4 (Sigma). This addition denatured the enzyme and avoided methylation after the allocated reaction time. The sample was then ready for histone extraction.

4.2.2.6. Mass spectrometry workflow for histones

Once the *in vitro* reaction had taken place, the samples were processed to the adequate conditions for LC-MS/MS analysis. This was accomplished in three steps: histone extraction, propionic anhydride derivatisation with trypsin digestion, and peptide desalting. This protocol is based on the method previously described by Garcia Lab [207] but adapted to the conditions of the assay.

Histone extraction

The goal of histone acid extraction is to get the histones out of the nucleosomes. As mentioned in section 4.2.2.5. , each reaction was stopped by addition of 400 µL of 0.6N H_2SO_4 (Sigma) for a total volume of 600 µL. After this addition, the sample was mixed with the pipette, then spun on a rotator mixer (Clifton) for 90 minutes at 4ºC. The samples were then centrifuged (4°C, 18000·g, 5 minutes) on the centrifuge Eppendorf 5424R. The same centrifuge was used for the rest of this section. The pellet formed (if any) would consist of non-histone proteins, which precipitate in strong acid as they are not as basic as histones. The supernatant of each sample, containing the histones, was then transferred to a protein LoBind Eppendorf microcentrifuge tube. A volume of 150 µL of 100% trichloroacetic acid (Sigma) was added to each tube and mixed. The mixture was then left to precipitate on ice for 3 hours. The samples were then centrifuged (4°C, 18000·g, 5 minutes) to pellet the proteins. After discarding the supernatant, 500 µL of acetone (Sigma) with 0.1% HCl (Sigma) was gently added to each sample with glass pipettes, avoiding disruption of the pellet. This was repeated another two times but substituting the acetone with HCl for pure, 100% acetone. After the last centrifugation, the supernatant was discarded, and the remaining sample was left air drying overnight.

Propionic anhydride derivatisation with trypsin digestion

The histone proteins were propionylated to label unmodified and monomethylated lysines with a propionyl group (+ 56 Da) so trypsin would cleave only at arginine residues and not the lysines, producing the same histone peptide lengths regardless of any post-translational modification present on lysine residues. This enables relative quantitation between different modifications on the same peptide based on signal intensity in the mass spectrometer. Besides, a subsequent propionylation of the N-termini of the corresponding histone peptides

derives in an increased chromatographic resolution of isobaric (i.e. same mass) peptides based on their hydrophobicity, reducing signal interference and increasing the confidence in the identification of PTMs, given that trypsin does not cleave as efficiently C-terminally to methylated lysines, therefore decreasing the peptide search space [193].

The propionic anhydride derivatisation protocol was performed as follows: once the samples were dry from histone extraction, 10 μL of 100 mM of ammonium bicarbonate (Sigma) were added to get the sample at a pH between 7 and 9. Propionic anhydride reagent was prepared by adding propionic anhydride reagent (Sigma) to isopropanol (Fisher Scientific) in a 1:3 ratio. A volume of 5 μL of this propionic anhydride reagent were then added to the sample. The pH was checked to reach 4-6 values, and 3.5 μL of ammonium hydroxide were immediately added to get the pH up again to ~8. After this, the samples were incubated in a heating block at 37ºC for 15 minutes, then dried using a SpeedVac Concentrator (Thermo Scientific Savant SC210A) until the volume was reduced to 10 μL or less. This procedure was repeated twice to ensure that the propionylation was efficient. After that, the samples were diluted into 100 μL of 100 mM ammonium bicarbonate (Sigma) at pH = 8.5. Trypsin (Promega) was then added in a 1:20 protease : histone mass ratio and the sample was left to react for 6 hours at 37ºC in a heating block. Then, another two rounds of propionic derivatisation took place to make sure that the propionylation was effective. The propionylated samples were then reconstituted in 200 μL of 0.1% AcOH in water, checking that the resulting pH was between 3 and 4.

Peptide Desalting

The salt existing in the sample hinders HPLC-MS analysis because it ionises during electrospray and suppress the signal from peptides. More importantly, salts can also form ionic adducts on peptides. The adducted peptide will have a different mass, therefore toughening its correct identification or quantification [157]. To remove buffer salts and exchange the solvent to one compatible with LC-MS/MS, stage tipping was performed on the histone samples. The stage tips were prepared by loading C-18 filters (Empore, 3M) onto P200 pipette tips to a thickness of 2 mm. The filter of the tips was then conditioned by passing 200 μL of MeOH (Fisher Scientific) through it, centrifuging (800·g, 5ºC, 5 min), then 200 μL of 0.1% AcOH (Sigma) in water solution, and centrifuging (800·g, 5ºC, 5 min). Once the filter was ready, the propionylated samples were loaded onto the filter-tip,

centrifuged (800·g, 5ºC, 5 min), then washed with 200 µL of 0.1% AcOH in water. The tips were then brought to new protein lo-bind microcentrifuge tubes where they were eluted with 200 µL of 75% MeCN (Sigma), 20% H_2O and 5% AcOH by centrifuging (800·g, 5ºC, 5 min). The eluted peptides were then dried in the previously mentioned SpeedVac Concentrator until ~2 µL. Once dried, MS-sample loading buffer (99.4% H_2O, 0.5% MeCN and 0.1% TFA (Sigma)) was added to the samples to obtain total sample volumes of 11 µL. These samples were then transferred to MS vials (Sun-Sri 500 212) for analysis.

4.2.2.7. Targeted LC-MS method to characterise histone PTMs

Nano-liquid chromatography tandem mass spectrometry (nano LC-MS/MS) was used to identify sites of G9a/SETD7 methylation on the lysine residues of histones, specifically for histone H3. The identification was based on the observation of a mass shift due to the incorporation of a CD_3 methyl group onto histone lysine side chains from CD_3-SAM and its synthetic analogues. The histone peptides were chromatographically resolved using an Ultimate 3000 RS-LC-nano System (Dionex), with an Acclaim PepMap100, C18 stationary phase, 2 µm particle size, 100 Å pore size, 75 µm internal diameter x 15 cm length column (Thermo Fisher). The nanoLC gradient started at 1% B (95% MeCN, and 0.1% formic acid) and 99% A (0.1% formic acid in water) for 3.5 minutes and increased to 80% B over 59.5 minutes followed by an increase to 98% B over 7 minutes. The solvents used were all LC-MS grade (Sigma Aldrich or Fisher Scientific).

The eluent from the nanoLC was ionised at 1.9 kV and real-time mass spectra were acquired on an LTQ Velos Pro linear ion trap (Thermo Scientific) with an 80-min acquisition time. The MS survey scan was performed over a mass range of 250–1650 m/z followed by CID fragmentation using a normalised collision energy of 35 V. Following the MS survey scan, several targeted SIM (Selected Ion Monitoring) zoom scans were executed for m/z values corresponding to the modified histone peptides. Lastly, tandem mass spectra corresponding to the methylated histone peptides were acquired using a priority queue. The Xcalibur software (Thermo Scientific) was used to generate and visualise extracted ion chromatograms corresponding to the m/z values for the histone proteoforms and quantitate relative methylation levels.

4.2.2.8. Estimation of the abundance of heavy methylation

The relative incorporation of heavy methylation ("abundance") within a specific histone proteoform was calculated as indicated in equation (4-2):

$$\frac{A_{j_{CD_3}}}{\sum_{j=1}^{n} A_{j_{proteoform}}} \tag{4-2}$$

where the numerator indicates the area under the chromatographic curve for the monoisotopic peak m/z ("area") of the heavy methylated species under study and the denominator indicates the summation of the areas of all the peptides of the proteoform that the species of the numerator belongs to.

The abundance of a proteoform was calculated according to equation (4-3):

$$\frac{\sum_{j=1}^{n} A_{j_{proteoform}}}{\sum_{i=1}^{p} A_{i_{peptide}}} \tag{4-3}$$

where the numerator indicates summation of the areas of all the peptides belonging to a specific proteoform and the denominator indicates the summation of the areas of all the proteoforms of a specific peptide.

4.3. Results and discussion

4.3.1. Computational validation: DOT1L "hole"

To validate the feasibility of our program to identify naturally occurring and selective KMT pockets, we decided to test if it could identify known structural differences between the SAM binding pockets of DOT1L and SET domain-containing KMTs, such as G9a. Specifically, the program searched in the direction of a new chemical bond to a specified design atom on SAM and checked for steric clashes with G9a and DOT1L residues. The SAM design point chosen was the N7 position as the addition of a single bromine atom at this position results in an 8-fold increased potency against DOT1L and reduced activities against other methyltransferases as proven by Yu et al. and their SAH-based inhibitor Bromo-deaza-SAH [86]. Furthermore, the SAM version of this inhibitor (CD_3BrSAM) was previously synthesised and tested for DOT1L selectivity and activity in our lab by Dr Nicole Trainor, and it was found that this bumped SAM cofactor resulted in selective histone H3K79 heavy methylation by DOT1L [208].

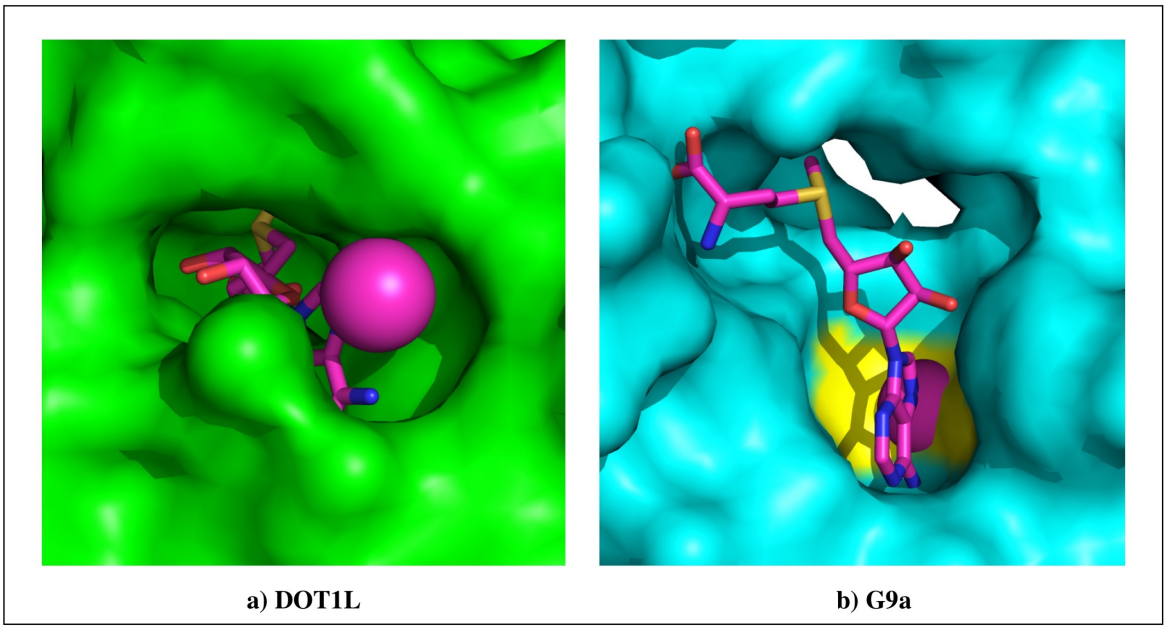


Figure 4-2. Structural differences between the SAM binding pockets of DOT1L (a) and G9a (b). These differences could enable the differential binding of SAM analogues when inserting groups at SAM N7 position. G9a residues H1113 and W1159 that clash with carbon atoms inserted at SAM's N7 position (represented as spheres) are shown in yellow.

As input to the program, PDB structures of DOT1L (PDB: 3QOW) and G9a (PDB: 2O8J) were aligned around their bound cofactor (SAM or SAH). This aligned PDB file was used as input into our algorithm for the identification of KMT pockets (described in Section 4.2.1.1.). The program output consisted of PDB files containing the G9a and DOT1L atoms within a specified radius of all SAH atoms for visual interpretation, and these structures are presented in Figure 4-2. This figure also shows the effects of appending a carbon atom to the N7 design atom of SAM fits within the pockets of the DOT1L and G9a structures, respectively. As can be seen from the figure, the insertion of a carbon atom does not generate any major clash for DOT1L, but clashes extensively with G9a.

In addition to the PDB structures, the program also outputs information regarding the number of "occurrences", being the number of clashes between the carbons appended to the SAH design atom with the KMT residues by counting the number of discretised volume units (e.g. 0.25 x 0.25 x 0.25 Å cubes) occupied by both. Any atom could in theory be "added" to the design point. In this case, carbon atoms were selected for simplicity. The results of our algorithm can be used only to ascertain the size of a naturally occurring cavity based on steric clashes. Computational methods for selecting the specific groups to add will be addressed in Chapter 5.

Using our program we identified that the addition of one carbon to the design atom resulted in strong clashes of 437 occurrences with G9a's H1113 and 331 occurrences with G9a's W1159 residues while only presenting a minor clash of 15 occurrences with P133 of DOT1L. Together with the information shown in the PDB structures seen in Figure 4-2, this provides an explanation of why inserting atoms at N7 position would make the SAM analogues specific towards DOT1L and validates the feasibility of our program to detect clashes.

4.3.2. SETD7 vs a post-SET domain-containing KMT (G9a)

4.3.2.1. Identification of structural differences unique to SETD7

After validating the program on previously discovered "pockets" in the structure of DOT1L, we used it to assess the structural differences between the SAM binding pockets of SETD7 and G9a, which was selected as a representative structure for the post-SET domain KMTs. Up to three carbon atoms were appended to the C2 design atom of SAH to search for steric clashes arising in SETD7 and DOT1L.

PDB crystal structures for SETD7 (PDB: 1XQH) and G9a (PDB: 2O8J) were aligned using their attached ligand as a basis for the alignment as described in section 4.2.1.1. and the alignment was inputted into our program. The output PDB structures showing how a linear three carbon chain from the C2 design atom of SAH interacts with these KMT structures are presented in Figure 4-3.

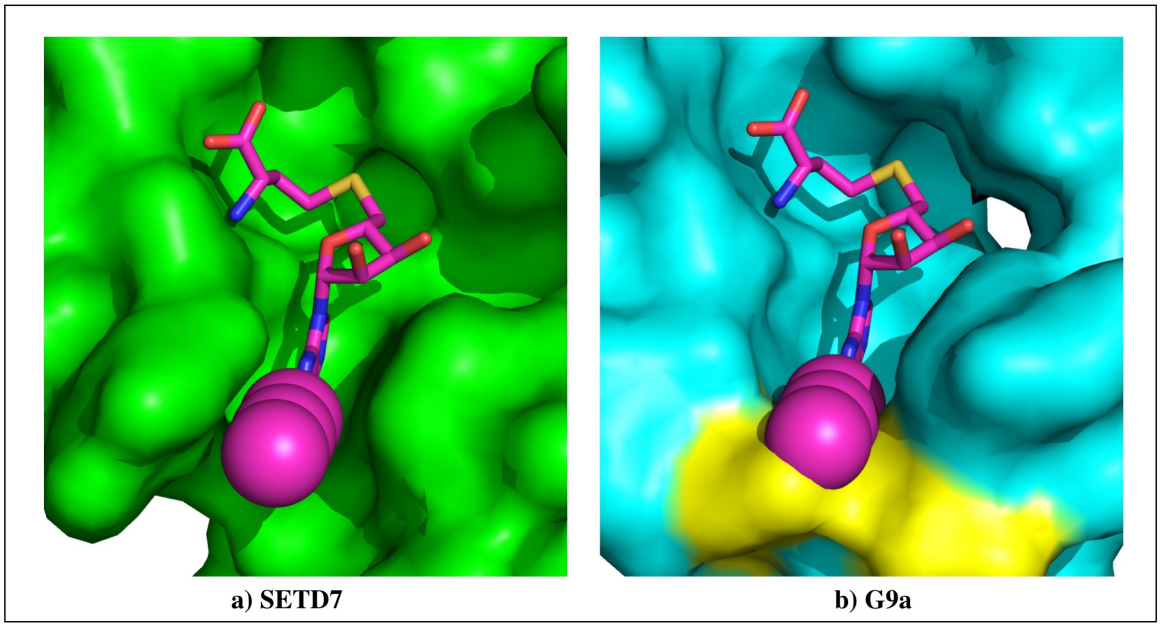


Figure 4-3. Structural differences between the SAM binding pockets of SETD7 (a) and G9a (b). These differences enable the differential binding of SAM analogues when inserting groups at SAM C2 position. G9a residues T1167 and Q1169 clashing with carbon atoms inserted at SAM's C2 position (represented as spheres) are shown in yellow.

As can be seen from Figure 4-3, the insertion of three carbon atoms does not generate any major clash for SETD7 as its structure is open around the C2 atom, but clashes with G9a's post-SET domain. As for DOT1L, the program also provided information about clashes of the inserted carbons at the C2 position of SAH with both SETD7 and G9a, counting the number of bins involved in that clash as "occurrences". No clash was identified for any of the three carbons inserted with SETD7. However, we identified that the addition of the first carbon to the design atom resulted in a strong clash of 117 occurrences with G9a's Q1169 and a minor clash of 2 occurrences with G9a's T1167; a second carbon also strongly clashed (276 occurrences) with G9a's Q1169 and presented another minor clash of 20 occurrences with G9a's T1167, and finally a third atom presented no clash with G9a's T1167 but collided strongly with G9a's Q1169, presenting 396 occurrences. Therefore, the insertion of moieties at SAM's C2 position has the potential to increase its selectivity towards SETD7 vs G9a.

Together with the graphical information shown in Figure 4-3, the program enabled the identification of a natural pocket in the SET domains of KMTs lacking a post-SET domain (i.e. SETD7) which does not exist in post-SET domain-containing KMTs. The output of the program also indicated that the insertion of a "bump" onto the C2 position of SAM should make the SAM analogues specific towards SETD7 vs post-SET domain-containing KMTs, such as G9a.

4.3.2.2. G9a purification and characterization

G9a was expressed and purified as explained in the sections 4.2.2.1. and 4.2.2.2. . The spectrum obtained after FPLC purification is shown in Figure 4-4. Even fractions obtained after FPLC purification were further analysed by SDS-PAGE electrophoresis as specified in section 4.2.2.3. . The result of this analysis is shown in Figure 4-5, which shows that the protein obtained in the FPLC fractions shows a similar MW to the one expected for G9a with the His-tag, 34.5 kDa by ExPASy [209]. To further assess if the protein present was indeed G9a, an in-solution digestion, mass-spectrometry analysis and proteome search was performed on the pooled purified G9a samples as described in section 4.2.2.4. . The result of this analysis showed that G9a had a coverage of 92.53% and a score (sum of the score of the individual peptides constituting the protein) of 3629.87, being the highest of all the

proteins found. The purified samples of G9a were flash frozen and stored in a 20 % glycerol solution at a concentration of 0.464 mg/mL of protein as measured by Nanodrop (Thermo Scientific).

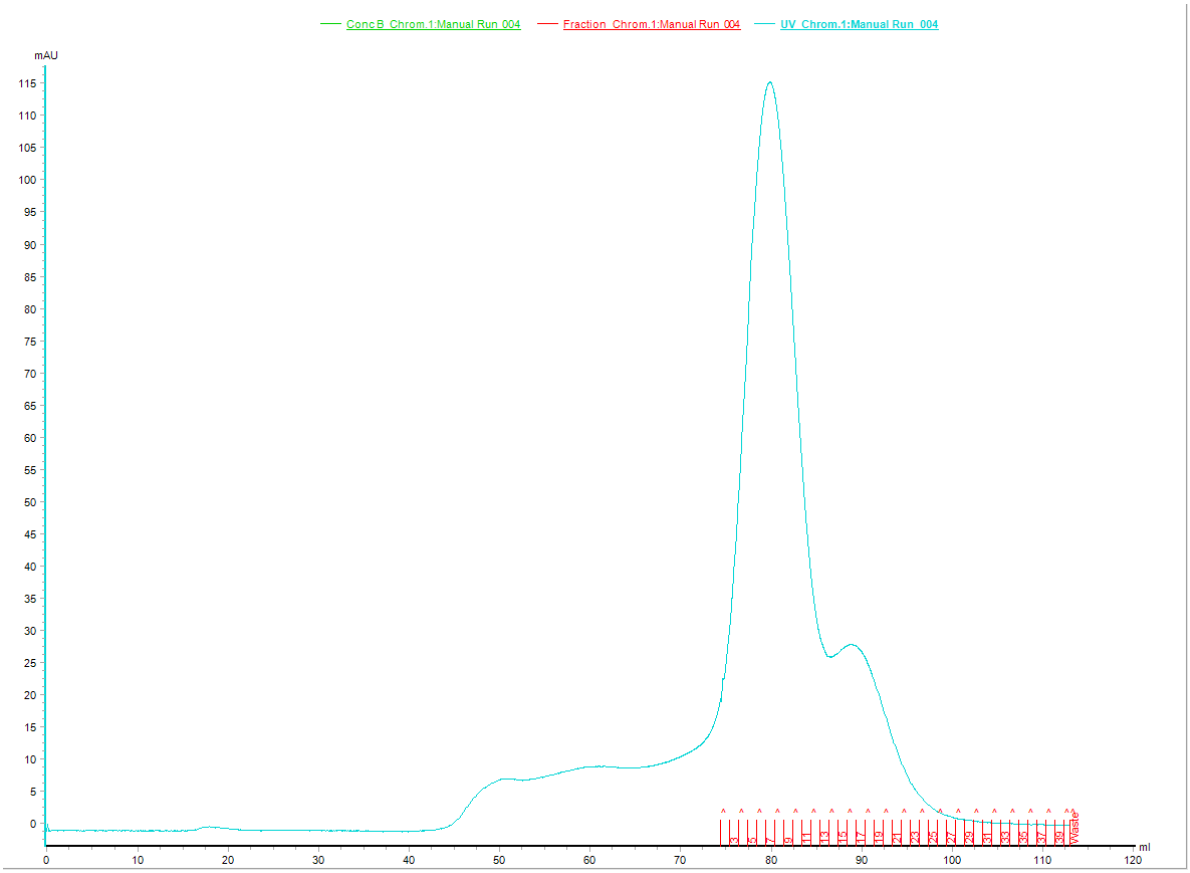


Figure 4-4. Spectrum obtained during the FPLC purification of G9a. The fractions collected appear in red at the bottom of the figure.

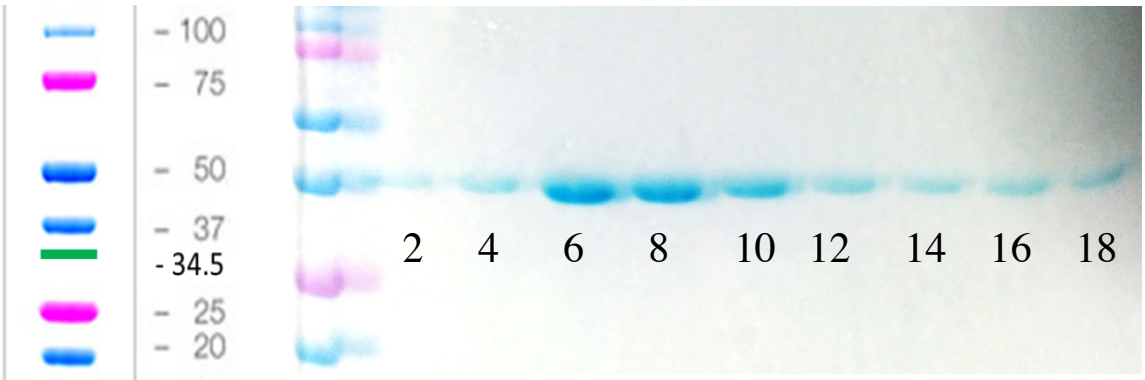


Figure 4-5. SDS-PAGE gel electrophoresis analysis of even fractions of the expressed G9a after FPLC. The added band (in green) indicates the MW expected for G9a (34.5 kDa).

4.3.2.3. SAM cofactor analogues design and synthesis

The position chosen to insert a group to act as a "bump" is SAM's C2 position due to a naturally occurring structural "hole" in the SAM binding pocket of SETD7 identified in section 4.3.2.1. . The insertions introduced at C2 position should clash with G9a post-SET domain but not with SETD7, as it lacks a post-SET domain. SAM cofactor analogues were proposed based on steric considerations and the chemical feasibility of adding specific chemical groups to ease the chemical synthesis. The cofactors were then synthesised by Katie Addison from the Department of Chemistry, Imperial College London. The proposed cofactors are shown in Figure 4-6.

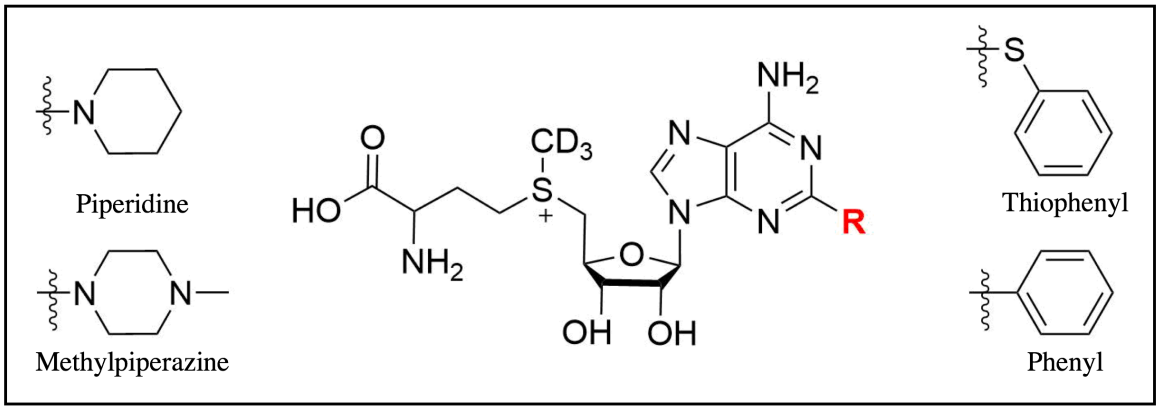


Figure 4-6. Proposed C2-substituted SAM cofactor analogues to be specific towards SETD7 vs G9a.

It was found that a nucleophilic substitution reaction could be utilised at the 2-position using various nucleophiles. Nitrogen nucleophiles were initially tested due to their common usage as nucleophiles with the reactions providing acceptable yields. Piperidine and methylpiperazine groups were chosen due to their similar size to a phenyl group, which is bulky enough to disrupt the correct binding of the SAM analogue with G9a but should be able to fit in the existing hole of SETD7. In addition, a thiophenyl group was installed with the sulphur still acting as a good nucleophile, however less electronegative than the nitrogen atom installed in the first two analogues.

A C2-modified SAM with a phenyl group insertion was also synthesised to test if this approach had a lower interference in the binding of the SAM analogues towards SETD7. Phenyl groups have been successfully utilised in previous implementations of the bump-and-hole approach [167,177]. The phenyl group was installed using a Suzuki-Miyaura palladium

cross-coupling reaction, where the coupling partners are a boronic acid and an organohalide catalyzed by a palladium(0) complex [210].

Overall, four C2-substituted SAM cofactor analogues were synthesised: piperidine-SAM (Pip-SAM), methylpiperazine-SAM (MPip-SAM), thiophenyl-SAM (TPhe-SAM) and phenyl-SAM (Phe-SAM).

4.3.2.4. *In vitro* assays with SETD7 and SAM cofactor analogues

In vitro experiments were performed to assess the activity of recombinant SETD7 with our engineered SAM analogues and chromatin substrates. As a first approximation to choosing the conditions for the *in vitro* assay, experiments at different concentration of heavy SAM and SETD7 were performed. The concentrations chosen were those that provided more heavy methylation in the same amount of time. The results of this first study are shown in Figure 4-7.

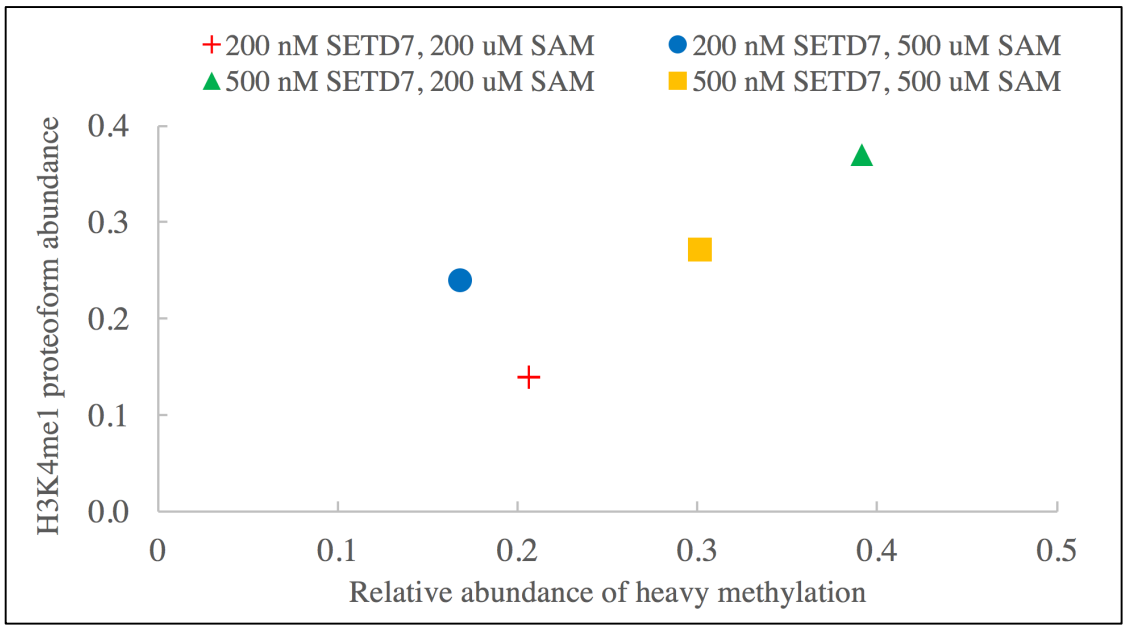


Figure 4-7. SETD7 H3K4me1 proteoform abundance vs relative abundance of heavy methylation for different SETD7 and SAM concentrations.

As observed in Figure 4-7, the amount of heavy methylation increased with the concentration of SETD7. However, the abundance of heavy methylation was reduced when the SAM concentration was increased from 200 to 500 μM. Therefore, the concentrations

of SETD7 and heavy SAM to obtain the maximum amount of heavy methylation under the conditions tested were 500 nM of SETD7 and 200 µM of SAM.

In vitro assays with the optimised concentrations were performed both with native heavy SAM and with the heavy SAM cofactor analogues. Two replicates of each experiment were performed to assess the reproducibility of the assay. The result of this analysis is shown in Figure 4-8. This figure shows that piperidine-SAM and methylpiperazine-SAM present slightly higher relative abundances of heavy monomethylation of H3K4 compared to native SAM (6% and 5% higher, respectively), whereas thiophenyl-SAM and phenyl-SAM present substantially lower abundances (40% and 37% less, respectively). In summary, these results indicate that the SAM cofactor analogues that are suitable to achieve high activity levels with SETD7 are piperidine-SAM and methylpiperazine-SAM.

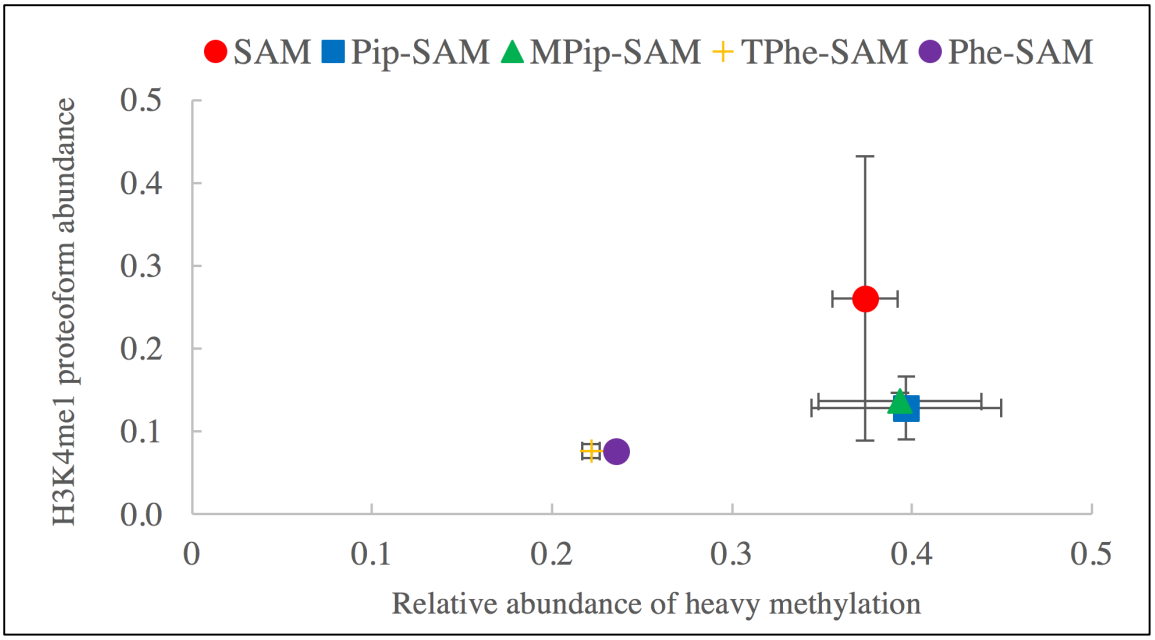


Figure 4-8. SETD7 H3K4me1 proteoform abundance vs relative abundance of heavy methylation for different SAM analogues. The error bars indicate the standard deviation of the measurements.

4.3.2.5. *In vitro* assays with G9a and SAM cofactor analogues

In vitro assays at the saturating concentrations of 1 mM of SAM cofactor and 1 µM of G9a were performed with both native heavy SAM and the heavy SAM cofactor analogues. The result of this analysis is shown in Figure 4-9.

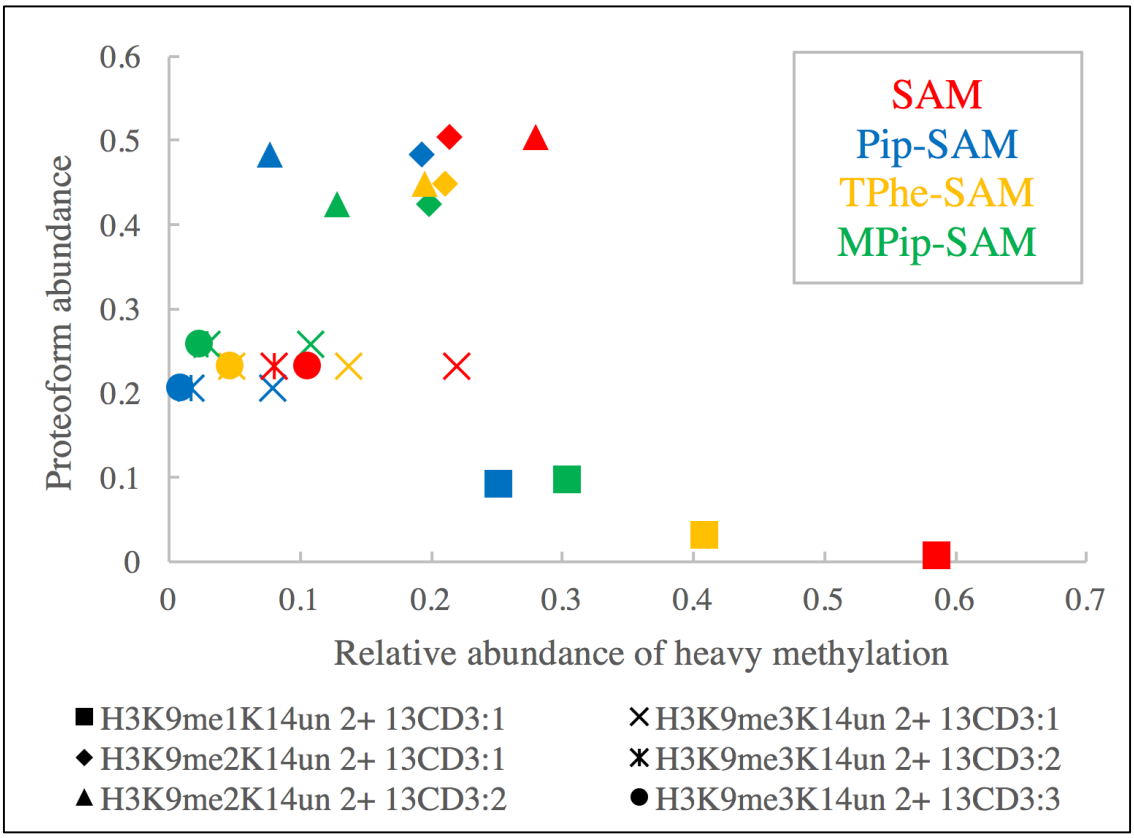


Figure 4-9. G9a peptide proteoform abundances vs relative abundance of heavy methylation with different SAM analogues. Un = unmethylated.

As shown in Figure 4-9, there appears to be a global decrease in heavy methylation when using our designed SAM cofactor analogues compared to the result obtained with native SAM. Overall, the SAM cofactor analogue achieving the highest changes in G9a's activity when compared with native SAM is piperidine-SAM. More specifically, G9a heavy methylation presents a decrease of 57% for H3K9me1K14un $^{13}CD_3$:1, 72% for H3K9me2K14un $^{13}CD_3$:2 and 92% for H3K9me3K14un $^{13}CD_3$:3 with piperidine-SAM when compared to SAM. It is necessary to note that this experiment was carried out in high saturating conditions (1 mM of SAM cofactor and 1 µM of G9a), therefore it is expected that the decrease in methylation activity could be higher when the conditions of the assay are optimised.

Overall, the most promising SAM cofactor analogue to evaluate the differential activity of SETD7 and G9a appears to be piperidine-SAM, as it preserves the activity of native SAM with SETD7 (6% higher methylation activity) but it dramatically worsens the results when used with G9a (up to 92% less methylation activity).

4.4. Conclusions and future work

We introduced a novel KMT "methylome" profiling assay based on the identification of unique "holes" in KMT structures via an in-house program based on the ESBTL library, followed by the engineering of synthetic SAM cofactor analogues with a complementary "bump" so that they bind selectively to the KMT with the "hole". The design is followed by *in vitro* assays with the SAM cofactor analogues proposed and analysis using mass spectrometry (MS) proteomics.

This study represents the first successful attempt in designing cofactors specific towards different KMTs based solely on existing natural pockets to guide the cofactor design. In addition, one of the main novelties of this approach is the introduction of a program to identify "holes" exclusively present in a subset of KMTs. Another advantage is that it can be easily extended towards the design of cofactors towards other enzymes such as, for example, kinases.

The program takes aligned KMT-cofactor structures as input and identifies pockets that are unique to one of the KMTs under analysis. The program served to identify a natural pocket in the SET domains of KMTs lacking a post-SET domain (e.g. SETD7) which does not exist in post-SET containing KMTs such as G9a. The output of the program was also used to elucidate design points for incorporation of synthetic "bumps" on the SAM analogue (C2 position) to match the identified "hole" in SETD7.

Several SAM analogues were proposed based on the steric properties of the SETD7 binding pocket and synthesised for experimental validation (in collaboration with Dr Matthew J. Fuchter and Katie Addison, Imperial College London). The SAM analogues also contained a heavy methyl donor group (e.g. SAM-CD_3) in substitution of the native methyl to ensure that the methylation event is both biologically consistent and detectable by MS proteomics.

A panel of SAM analogues were tested through *in vitro* assays with chromatin substrates and SETD7 or G9a to demonstrate selective utilisation by the KMT of interest, where mass spectrometry (MS) proteomics was the primary readout for validating catalytic activity. These assays revealed cofactor selectivity of the SAM cofactor analogue piperidine-SAM towards SETD7. More specifically, piperidine-SAM preserves the activity of native SAM with SETD7, with a 6% higher methylation activity than with SAM, but it dramatically

worsens the results when used with G9a, showing up to 92% less methylation activity when compared to SAM. These results support the feasibility of our proposed design methodology to distinguish between the methylation activity of different KMTs.

Future work about this project could be performed in some areas of improvement. First, G9a *in vitro* assays were only successfully performed on very saturating conditions and with only one replicate due to technical difficulties and time constraints. Further experiments were run at varied concentrations of SAM and G9a but none of them seemed to show activity. We attribute this to the loss of activity of the G9a in use. We propose to perform a similar analysis to the one performed with SETD7 but for G9a, where the conditions achieving higher heavy methylation are obtained and then replicates of the *in vitro* assays of G9a with different SAM analogues are performed. Besides, more replicates of SETD7 at the identified optimal conditions would provide further reliability to the results.

Furthermore, we propose to extend our approach to cellular assays. One of the main advantages of this extension would be the identification of non-histone targets. However, it would be necessary to assess the permeability of SAM cofactors, as they are known to present low permeability through the cell membranes [211]. One possible solution would be the production of these analogues inside the cells, adapting the human methionine adenosyltransferase (MAT) to accommodate bulky adenine analogues, in a similar way to the accommodation of bulky methionine analogues introduced by Luo's lab [212].

The design of SAM cofactor analogues in this chapter was purely based on the lack of steric clashes and the feasibility of the cofactor chemical synthesis. In the next chapter, we will go deeper into the design of SAM cofactor analogues and we will discuss computational methods to prioritise the synthesis of specific cofactors.

Chapter 5

In silico screening of engineered KMT-SAM systems

Biomolecular simulations are increasingly used in the pharmaceutical industry as screening tools for new drugs, to identify novel binding sites or to compute small-molecule binding energies to different receptors [213]. The reasons for this trend are the development of novel, more accurate, sampling algorithms [214], the improvements in computational power and a reduction in the cost of computational equipment.

This project involves the design of SAM cofactor analogues to detect the activity of KMTs exploiting differences between binding pockets. The implementation of a computational screening tool prior to experimentation is necessary to prioritise the analogues for chemical synthesis. This is because the chemical synthesis of naturally occurring cofactors is generally challenging and costly in time, reagents and risk [215]. A computational screening approach would reduce the number of experiments to perform and provide theoretical support for the generated experimental data.

Therefore, biomolecular simulations were included in this project to assess the binding and suitability of the different enzyme-cofactor pairs for protein methylation. To test if the

cofactors could be utilised by the enzymes, the simulations and their analysis should provide insight into the following:

- The binding orientation of SAM cofactor analogues with respect to the substrate channel (as a necessary condition for catalytic efficiency); more specifically the distances and angles between the sulphur-methyl group of the SAM analogues and an atom near the nitrogen acceptor of the histone lysine.
- The relative binding affinities to prioritise which SAM cofactor analogues should be used in the experiments, based on them having similar binding energies with respect to native SAM.

The approach proposed could be easily extended to any other project involving the design of cofactor analogues.

The analysis of crystallographic structures using PyMOL [216] or even the generation of homology models using Rosetta [217] do not consider detailed modelling of cofactor-KMT interactions [218]. An option taken into consideration was the use of scoring functions of docking programs such as Autodock Vina [219] due to the speed and capacity to screen complete libraries of compounds in relatively short times. However, the level of accuracy is empirically much lower than estimates based on molecular dynamics simulations [220].

To model the intermolecular interactions accurately, it is necessary to allow for flexibility in both the protein and ligand atoms, not only for the side chains but also for the backbone atoms of the protein. This can be achieved by using the detailed atomic force fields included in molecular dynamics (MD) packages, which are also required to infer any level of binding. Therefore, MD was utilised in this project and it is the focus of the next section.

5.1. Introduction to molecular dynamics

5.1.1. The basis of molecular dynamics simulations

Molecular dynamics (MD) is a widely-used tool to assess the structure and dynamics of biomolecular systems and obtain detailed information about features such as ligand binding and protein folding. To represent the physical behavior of a system, MD relies on the representation of the energy of the system as a function of its atomic coordinates, with the function expected to achieve a minimum value at equilibrium. As the forces on individual atoms are computed from the gradient of this function, the function is commonly referred as a "force field". The most common potential energy function incorporated by force fields is the called "Class I" [221]:

$$U(x) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi[\cos(n\phi + \delta) + 1] + \sum_{\substack{nonbonded \\ pairs}} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{{r_{ij}}^{12}} + \frac{C_{ij}}{{r_{ij}}^{6}}\right] \quad (5\text{-}1)$$

In equation (5-1), the first term represents the bonds, the second the angles and the third the torsions or dihedrals (1-2, 1-3 or 1-4 interactions) contributions, as shown in Figure 5-1.

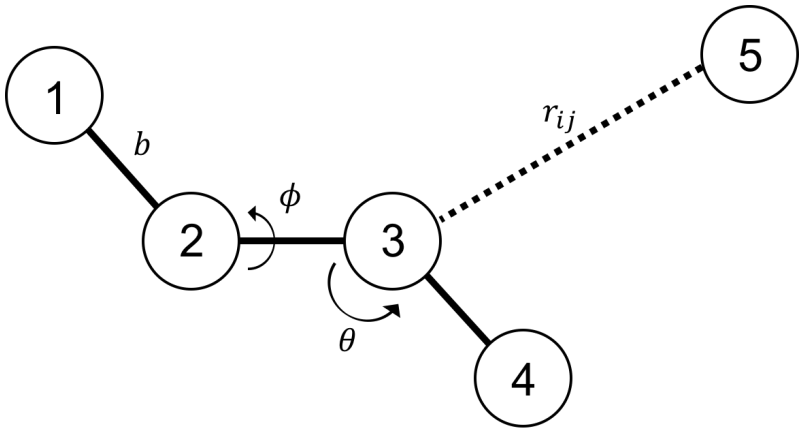


Figure 5-1. Scheme of force field interactions. b represents the distance, θ the angle and ϕ the dihedral. Covalent bonds are shown with solid lines, whereas non-bonded interactions are shown with a dashed line. Adapted from Ponder et al. [221].

The final summation term over pairs of atoms i and j in equation (5-1) represents the electrostatics and the van der Waals energies. In particular, the electrostatic forces are estimated using the partial charges q_i on each atom interacting though Coulomb's law in the first term of the summation. The remaining terms estimate the 6-12 Lennard-Jones potential representing the dispersion and exchange repulsion forces in the van der Waals equation. Equation (5-1) is the simplest that can reproduce the protein energy interactions at an atomic detail. The combination of the potential energy function described above with specific parameters for each atom type (e.g. k_b, k_θ, etc) composes the "force field" (FF) [222]. The parameters could be obtained either experimentally or through ab initio quantum mechanical simulations.

The energy function enables the calculation of the force experienced by an atom with respect to the positions of the rest of the atoms. Then, the position and velocity of each atom is updated by using Newton's laws of motion for every time step. This time step needs to be at least on the order of femtoseconds (10^{-15} s) to achieve an accurate representation [223]. According to Newton's second law:

$$F = ma \tag{5-2}$$

where F is the force on an atom, m is the mass of the atom and a is the acceleration of the atom. The force acting on each atom can be obtained from the potential energy function as follows:

$$F(x) = -\nabla U(x) \tag{5-3}$$

where x represents the atom coordinates and U is the potential energy function. Given that acceleration is the derivative of velocity, and velocity is the derivative of the position:

$$\frac{dv}{dt} = \frac{F(x)}{m} = \frac{-\nabla U(x)}{m} \tag{5-4}$$

$$\frac{dx}{dt} = v \tag{5-5}$$

The basic solution to update the velocity and position for each atom is:

$$v_{i+1} = v_i + \delta_t F(x_i)/m \tag{5-6}$$

$$x_{i+1} = x_i + \delta_t v_i \tag{5-7}$$

where δ_t is the time step. Current MD packages such as Amber rely on more accurate numerical integrating methods of Newton's equations of motion such as the Verlet integrator [224,225], which is based on additions of the Taylor expansions for +Δt and -Δt to update positions and velocities [222], obtaining equation (5-8) and equation (5-9):

$$r_i(t_0 + \Delta t) = -r_i(t_0 - \Delta t) + 2r_i(t_0) + a_i(t_0)\Delta t^2 + O(\Delta t^4) \tag{5-8}$$

$$v_i(t_0) = \frac{1}{2\Delta t}\left[r_i(t_0 + \Delta t) - r_i(t_0 - \Delta t)\right] \tag{5-9}$$

Variants of equations (5-8) and (5-9), such as the velocity-Verlet algorithm and the similar leap-frog algorithm [222] are used at every time step to update the position and velocity of each atom, resulting in the MD trajectory.

5.1.2. Methods to calculate the binding affinity

When an estimation of the binding affinity between a protein and a small molecule is required, there are two main categories of methods that can be used: exact methods where no empirically fitted parameters are used, and approximate methods using both physical laws and empirical parameters.

Regarding the exact methods, the most popular ones are thermodynamic integration (TI), based on the energy change that one system undergoes when it changes steadily into another, and free energy perturbation, based on exponential averages of changes in potential energy [218]. Both methods estimate the free energy differences from system simulations that are not physics-based (alchemical). The application of thermodynamic cycles on the simulation

results can provide an estimation of the free energy differences of interest. An example of thermodynamic cycle is shown in Figure 5-2.

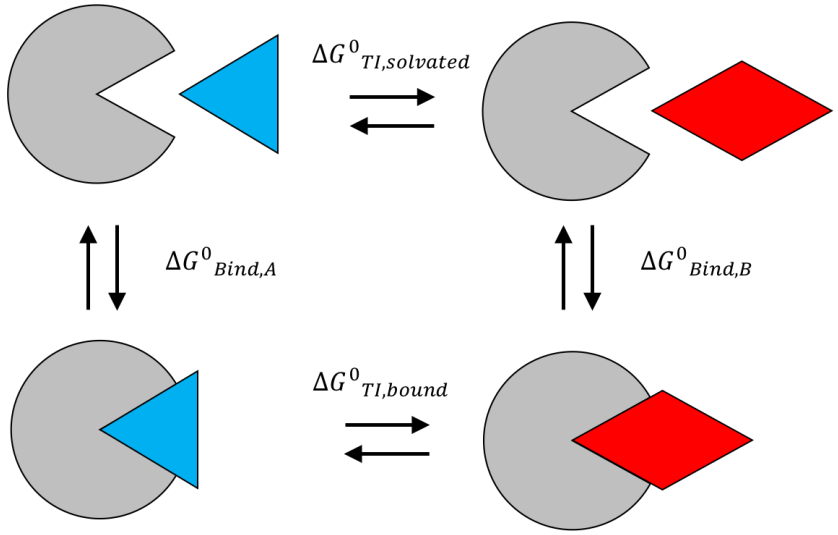


Figure 5-2. Thermodynamic cycle used to calculate the relative binding free energy difference, $\Delta\Delta G^0{}_{Bind}$, between two different ligands A and B, shown in blue and red respectively. $\Delta G^0{}_{Bind,A}$ and $\Delta G^0{}_{Bind,B}$ correspond to the binding free energies of each of the ligands, while $\Delta G^0{}_{TI,solvated}$ and $\Delta G^0{}_{TI,bound}$ are the variations in free energy occurring when transforming the ligands in the bound and solvated (or unbound) state, respectively. Adapted with permission from Steinbrecher, T., Case, D. A. & Labahn, A. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. J. Med. Chem. 49, 1837–1844 (2006). Copyright 2006 American Chemical Society[226].

The relative binding free energy can thus be computed according to Equation (5-10):

$$\Delta\Delta G^0{}_{Bind} = \Delta G^0{}_{Bind,A} - \Delta G^0{}_{Bind,B} = \Delta G^0{}_{TI,solvated} - \Delta G^0{}_{TI,bound} \qquad (5\text{-}10)$$

Equation (5-10) indicates that the value obtained is independent from the path taken to get from the protein bound to ligand A to the protein bound to the ligand B. This path is represented by the alchemical Hamiltonian operator representing to the total energy of a system (kinetic and potential). The Hamiltonian is usually provided as a function of the coupling parameter λ, where $0 < \lambda < 1$, 0 represents the protein bound to ligand A and 1 the protein bound to the ligand B [227]. The free energy perturbation and thermodynamic integration methods to estimate binding energies are described below.

5.1.2.1. Free energy perturbation (FEP)

The free energy perturbation is a method that calculates the differences in free energy from one system to another after undergoing alchemical transformations, characterised by

changes in the chemical identity of the atomic models during the simulation. It is based on exponential averages of the change in the potential energy of the system, as it is assumed that the kinetic component does not change between the systems [228]. The term "perturbation" refers in this case to the alchemical change of one of the molecules of the system [229]. The free energy change is calculated by the application of the Zwanzig relationship [230], according to Equation (5-11):

$$\Delta G = \beta^{-1} \ln\langle e^{-\beta[H_1(x)-H_0(x)]}\rangle_0 = \beta^{-1} \ln\langle e^{-\beta\Delta H(x)}\rangle_0 \qquad (5\text{-}11)$$

where $\beta = (k_B*T)^{-1}$. This exponential average converges only when small changes occur between one system and another. The transformation is therefore broken into a series of free-energy calculations with nonphysical intermediates. For example, these intermediates can be generated by turning off the atomic charges or by changing the element of one atom into another. The total free energy change is estimated as the sum of the individual free energy changes between the intermediate states, which can be entirely nonphysical.

5.1.2.2. Thermodynamic integration (TI)

The thermodynamic integration method is based on integrating the change in energy that one system undergoes when it is steadily changed into another (e.g. alchemically). It requires a series of simulations that are run at different values of the coupling parameter λ, that describes a series of intermediate alchemical states [229]. The derivative of the potential energy of the system with respect to λ is numerically integrated [226] to obtain the difference in free energy as shown in Equation (5-12):

$$\Delta G^0 = \int_{\lambda=0}^{1} \left\langle \frac{dU}{d\lambda} \right\rangle_\lambda d\lambda \qquad (5\text{-}12)$$

Thermodynamic integration requires fewer uncorrelated samples to get an acceptable value of relative error than FEP, but the mandatory use of a numerical method to compute the integration introduces bias in the final result [229].

In summary, the calculation of binding free energy between two molecules with FEP and TI methods require alchemical transformations in both bound and unbound states. The increments in the coupling parameter λ need to be small to ensure phase space overlap

between the states for FEP or to provide enough points to perform an accurate numerical integration in TI. These demands translate into computationally expensive calculations that do not enable the high-throughput of ligands [231]. More approximate methods have been developed to save both time and computational resources. These methods make use of less accurate physical models and empirically fitted parameters to simplify the modelling. Among those, the most important ones in increasing order of accuracy are: docking, linear interaction energy (LIE) and Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA).

5.1.2.3. Docking

Docking methods aim to identify the most stable conformations of a ligand in a receptor using a simplified energy model, which is usually an empirical force field with an implicit solvent model. The conformations found are then used to compute a binding energy or score a specific ligand conformation. This score can be computed with the same energy model used for the docking or with a more complex and computationally expensive model [232]. The three main criteria used to compare the performance of docking programs are the ability to: discern the correct bound conformation of a ligand in a receptor, assign accurate binding scores to known, high-affinity ligands, and generate scores that correlate to measured binding affinities with known ligands.

A variety of scoring functions have been developed to score different ligand-protein pairs. Different scoring functions are available depending on the software. The most popular docking programs are: AutoDock [233], GOLD [234], Swissdock [235], Glide [236], Surflex [237], LigandFit [238], FlexX [239], eHiTS [240], Drugscore [241], and Autodock Vina, the latter being a faster and more accurate version of Autodock, with extended ease of use and the possibility of setting flexible receptor side chains during docking [219]. For these reasons Autodock Vina was the docking program used for this project.

Although these methods are fast and do not require extensive computational resources, they present some critical disadvantages. The main disadvantage is the simplicity of the scoring functions, which are not designed for physical accuracy but for computational throughput. The scores are therefore often weakly predictive of the experimental binding affinities. Docking methods also rely heavily on details such as the protonation state of the ligands [242].

Due to the aforementioned disadvantages, docking methods are mainly used for the initial screening of drug libraries, when speed is critical. For example, to screen the ZINC database of chemical compounds, which contains over 120 million commercially available molecules [243,244] with a receptor. Since docking programs utilise empirically based algorithms and simple equations, with little or no protein motion incorporated into the evaluation [245], they are not accurate enough to make rigorous distinctions between the binding affinity of a limited set of ligands to the same receptor, as this project requires.

In order to overcome the limitations of docking, the "end point free energy methods" were developed. These methods use conformational sampling to generate thermodynamic averages, as opposed to docking programs that calculate the binding energies based on a single bound conformation of the ligand-protein complex. End point methods eliminate the problems associated with the calculation of a single conformation associated with docking, but require more computational time and power to converge to meaningful results. The basis of these two methods is similar: the program inputs frames of different conformations of the free and bound states and computes the binding affinity as the energy difference between the states. The most common end-point approaches, LIE and MM-PBSA, are described in sections 5.1.2.4. and 5.1.2.5. .

5.1.2.4. Linear interaction energy (LIE)

The Linear Interaction Energy (LIE) method involves an approximate binding affinity calculation that considers the end points of the binding process. This method involves running two MD simulations, one for the ligand in solution and another for the ligand in the binding site. Snapshots of the MD trajectories are extracted and used to calculate Boltzmann electrostatic (U_{elec}) and van der Waals (U_{vdw}) interaction energies of the ligand with the rest of the molecules in its bound and free state [246]. The binding free energy can be approximated using the following formula:

$$\Delta G^0 \approx \beta\left(\langle U_{elec}\rangle_{bound} - \langle U_{elec}\rangle_{free}\right) + \alpha\left(\langle U_{vdw}\rangle_{bound} - \langle U_{vdw}\rangle_{free}\right) \quad (5\text{-}13)$$

where the < > symbol indicates that that parameter is calculated as the ensemble average over the MD trajectory, and α and β are empirical parameters accounting for changes in the internal energies of both solvent and protein.

Some of the benefits of LIE is that it is over five times faster than "exact" methods such as FEP, as LIE directly simulates the end states. Furthermore, compared to docking LIE uses an explicit solvent model, so the desolvation free energy can be accurately computed. Finally, and as opposed to docking/scoring function approaches, LIE considers the protein flexibility similarly to FEP but with a much lower computational cost [247]. The disadvantages of LIE are that it does not account explicitly for concentration, changes in the configurational entropy nor the internal energy of the ligand [232]. The method is therefore limited to comparing very similar ligands that do not strongly alter these terms.

5.1.2.5. Molecular mechanics Poisson-Boltzmann surface area (MM-PBSA)

The Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) method and its Generalised Born variant (MM-GBSA) are end-point methods. These methods input MD trajectories of the free ligand, free protein, and their complex to compute the average potential and solvation energies [248]. The calculation is done according to Equation (5-14), which represents the multiple trajectory protocol (MTP):

$$\Delta G_{bind} = \langle G_{PL} \rangle_{PL} - \langle G_P \rangle_P - \langle G_L \rangle_L \quad (5\text{-}14)$$

where PL: complex, P: protein, and L: ligand. However, it is more common to start from an MD simulation of the complex only, then obtain trajectories of the free ligand and receptor by removing the pertinent atoms. This approach is called single trajectory protocol (STP). Equation (5-15) summarises this approach.

$$\Delta G_{bind} = \langle G_{PL} - G_P - G_L \rangle_{PL} \quad (5\text{-}15)$$

The STP requires less simulation time and is appropriate if the receptor and ligand ensembles are comparable in the bound and unbound states. For both MTP and STP, the free energy is computed as a function of the terms shown in Equation (5-16) [249].

$$G = E_{bnd} + E_{el} + E_{vdW} + G_{pol} + G_{np} - TS \quad (5\text{-}16)$$

The first three items are based on standard molecular mechanics (MM) energies: bonded (E_{bnd}, bond, angle and dihedral), electrostatic (E_{el}) and van der Waals (E_{vdW}) interactions. G_{pol}

and G_{np} are the polar and non-polar contributions to the solvation free energies, respectively. G_{pol} is usually calculated by applying the Poisson-Boltzmann (PB) equation (or by using the Generalised Born (GB) model in MM-GBSA approach), while G_{np} is estimated from a linear relation to the solvent accessible surface area (SASA). Finally, the last term in Equation (5-16) is the absolute temperature of the system, T, multiplied by the entropy, S, which can be optionally computed by a normal-mode analysis of the vibrational frequencies [250].

Compared to other end-point methods such as LIE, MM-PBSA performance is comparable and their relative accuracy depends on the tested system. For example, Genheden et al. demonstrated that the binding estimate of avidin to a set of seven biotin analogues were more accurate for LIE, but MM/PB(GB)SA outperformed LIE for the binding of nine inhibitors to fXa [251]. Besides, it was proposed that inaccuracies in the MM/PB(GB)SA method might be caused by perturbations in the results as a consequence of introducing the entropy term [252]. This is because an additional energy function is used to estimate the entropy; therefore, the entropy term typically gives the largest statistical uncertainty [250]. In order to overcome this limitation, the entropy term is commonly omitted when running the MM/PB(GB)SA calculations [253], given that it is only necessary for the computation of absolute affinities and not to compute the relative affinities between different ligand-receptor pairs.

In conclusion, the MM-PBSA method is based on MD trajectories of the ligand-receptor complex, and it is an intermediate method both in terms of accuracy and computational cost compared to the fast but simple and purely empirical docking/scoring methods [220] and the strict but slow and complex alchemical methods such as TI or FEP [254]. Some studies show that the accuracy of MM-PBSA is even comparable to the one of alchemical methods depending on the system [250]. Another advantage of the MM-PBSA method is its integration within the AmberTools package in a program called MMPBSA.py [248]. For all these reasons, the MM-PBSA method was chosen for this project.

5.1.3. Molecular dynamics simulations of KMTs

From early studies of methyltransferases through molecular dynamics simulations involving the study of Catechol O-methyltransferase [255], these enzymes have become a recurrent target

of biomolecular simulations due to their importance in the regulation of many gene expression programs and in the development and progression of many diseases such as cancer.

5.1.3.1. MD simulations of SETD7 and G9a-like protein

Regarding SETD7, there are various efforts in the literature where it has been studied through MD. In 2007, quantum mechanical/molecular mechanical (QM/MM) MD and free-energy simulations were used to understand the deprotonation of the positively charged methyl lysine process occurring during the SETD7 histone H3K4 methylation catalysis and to explain its specificity towards specific products. These studies suggested that Y335 play the role of the base for the deprotonation [62]. This was tested with QM/MM free-energy (potential of mean force) simulations using CHARMM [256] in simulations of 1.5 ns.

Soon after, further (QM/MM) MD simulations were performed on SETD7 to confirm that the methylation state specificity was controlled by the catalytic residues involved in the methyl transfer and further proof to classify SETD7 as a mono-methyltransferase. These studies also found that once the lysine substrate is methylated, it is able to open up the SAM binding pocket so water molecules can reach the active site [257]. In this case, the QM/MM MD simulations were carried out in Gaussian03 [258], Q-Chem [259] and TINKER [221] achieving 1 ns simulation time.

A more recent simulation of SET7/9 was performed by Yao [260], focusing on understanding the differences in the methyl-transfer processes, the specificity and the efficiency, between SET7/9 and its Y245A mutant, concluding that the Y245A mutation increased the ability of the enzyme to transfer more methyl groups to the lysine target. Hybrid QM/MM free energy simulations were performed using CHARMM [256] in 1 ns simulations for each of the methyl transfer events.

QM/MM MD simulations were also performed on endogenous G9a-like protein (GLP) and its F1209Y and Y1124F mutants in order to understand the influence that the Phe/Tyr switch has on the product specificity of GLP [261]. The results supported the role of these residues on the product specificity of GLP. Regarding the simulations, SAM and the target lysine

residue were simulated using QM, whereas the rest of the system was modelled using MM. MD simulations of 1 ns were performed using CHARMM [256].

5.1.3.2. MD of PMTs to generate binding affinity estimates

The simulations for SET7/9 and GLP mentioned in section 5.1.3.1. were focused on gaining a fundamental understanding about the methyl transfer process and product specificity. This required the more computationally expensive QM/MM methods. This approach is generally not scalable to discern which SAM cofactor analogues bind with stronger affinity and better conformation to enable the methyl transfer, as the level of detail is unnecessarily high and the required simulation time would make use of extensive computational resources and time. The next paragraphs will review on previous MD simulations of protein methyltransferases (PMTs) to generate binding affinity estimates.

The binding mode of small molecule inhibitors to the SET domain lysine methyltransferase Enhancer of Zeste Homolog 2 (EZH2) was analysed through molecular dynamics simulations [262]. In this work, the system was energy minimised, heated and equilibrated for 5 ns prior a 4-ns production run. The production run was used to generate 500 snapshots (also called frames) of the complex as input for the MMPBSA.py program. The results produced binding affinity estimates of the inhibitors to EZH2, showing correlation with available experimental results. All calculations were performed with Ambertools12 [263].

A similar strategy was followed to test the binding affinities of a range of inhibitors of human protein arginine methyltransferase 1 (PRMT1), performing the MM-PBSA calculations with the MMPBSA.py program [264]. To generate the MD simulation required, Amber12 was used [263]. The protocol involved energy minimization, heating (20 ps), equilibration (320 ps) and production (28 ns) runs. Out of the 28 ns production run, the last 10 ns were taken to generate 1000 snapshots for the MM-PBSA calculation.

Recently, further MD simulations were performed for PRMT1 (including SAM and the arginine substrate) to understand the impact of M48F, H293A, H293S, and H293S-M48F mutations [265]. It was concluded that two residues, M48 and H293, play a key role in the regulation of PRMT1 product specificity. These MD simulations were performed in Amber12 [263], involving minimization, heating (50 ps), equilibration (1 ns) and a 300 ns

production run. The MM-PBSA calculation was performed based on all the snapshots from the 300 ns production trajectories.

Phenoxyacetamide derivatives were recently identified as Dot1-like histone methyltransferase (DOT1L) inhibitors [124] via docking with Glide [236] and MD simulations using GROMACS [266]. The protocol involved energy-minimization, heating (200 ps) and equilibration (200 ps) followed by a production run of 10 ns. The binding free energy of DOT1L towards the phenoxyacetamide compounds was computed from the production trajectory using the g_mmpbsa tool [267], a GROMACS tool enabling MM-PBSA calculations. It was concluded that 4 of the compounds showed much higher binding affinity than others, with one in particular having the lowest Glide score and therefore supporting the results of the g_mmpbsa tool.

5.1.4. Technical challenges and opportunities

MD simulations are increasingly being used as the basis to compute binding free energy estimates of small molecules towards methyltransferases with more accuracy than docking programs. The preferential tool to perform the actual free energy calculations from MD simulation trajectories is the MM-PBSA method, due to its versatility, accuracy and affordable computational cost. For these reasons, MD and MM-PBSA were the chosen tools to evaluate which SAM cofactor analogues bind with stronger affinity and better conformation to enable the methyl transfer towards the enzymes under study: G9a and SETD7.

This project is not exempt of challenges as it is commonly the case with method development research. First, there are no crystallographic structures of the enzymes bound to the SAM cofactor analogues but only with native SAH. Therefore, it is necessary to perform docking of the SAM cofactor analogues prior to the MD simulation and select which binding mode is likely to resemble the real scenario by comparing the obtained conformation with the one of native SAM in the equivalent crystallographic structure. This project also demands a strategy for the MD simulation (minimization, heating, equilibration and production) that is adapted both to the size and physical properties of G9a and SETD7 and to the objective of these simulations, an equilibrated trajectory long enough to derive

the binding affinities from it. In addition, G9a contains a ZnCys motif [268]. Metal sites demand a separate parametrization prior MD simulations that is not included in standard protocols for the application of the MM-PBSA approach.

To the best of our knowledge, molecular dynamics simulations have not been previously applied to the estimation of binding affinities for a diverse set of SAM cofactor analogues towards SETD7 and G9a. Only QM/MM MD studies exist for SETD7 [62,257,260] that aim to characterise the methylation event of the enzymes, but not with the purpose of using it as screening tools to assess selective binding. The work presented in this chapter aims to fill this gap in knowledge and presents an opportunity for a more accurate computational determination of both the binding orientation and relative binding affinities of a set of SAM cofactor analogues towards G9a and SETD7.

5.2. Methodologies

The objective of this chapter is to develop a computational method to rank the SAM cofactor analogues with regards to their capacity to be utilised by G9a and SETD7, assuming that the more the similarities to the endogenous SAM binding mode, the higher the compatibility. To assess this, it is necessary to gain further insight into the binding mode of the proposed SAM cofactor analogues and evaluate if they bind with distances, angles and affinities similar to the ones of native SAM.

In order to choose the software to accomplish the goals of this project, it is necessary to first consider the project requirements, namely: parametrization of novel organic molecules, parametrization of metal centers, generation of topology and coordinate files of solvated enzyme-cofactor complexes, adequate post-analysis tools to assess when methylation might be occurring, and a relatively fast and accurate estimation of relative binding affinity between some enzyme-SAM analogue pairs and the same enzyme with native SAM.

Amber (Assisted Model Building with Energy Refinement) is a package of computer programs used to apply molecular mechanics, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules [225]. It can be used for the preparation, performance and analysis of molecular dynamics simulations [269]. Amber also includes a series of molecular mechanics force fields such as Amberff14SB for amino acids and nucleic acids [270], and GAFF (General Amber Force Field) for general organic molecules [271].

Amber includes three main MD simulation engines, each of them adapted to different purposes [272]. Sander is the less computationally efficient, but it is the most flexible and therefore mainly used for development purposes or less computationally demanding tasks. Pmemd is a version built on sander but with improved parallel scalability and higher performance, therefore used for more expensive tasks. Finally, pmemd.cuda is a version of pmemd adapted to its use on GPUs, including calls to specific CUDA kernels for GPU acceleration. This version is thus ideal for large solvated systems to run for long periods of time. For this project, sander was used for the minimization and heating of the systems, whereas pmemd.cuda was used for the production runs.

In order to achieve the objectives of this project, Amber16 was selected among other popular Molecular Dynamics (MD) packages such as CHARMM [256] or GROMACS [273]. The reason why Amber was chosen was the fact that it includes programs to accomplish all the requirements of this project:

- **Antechamber:** a set of tools to automatically generate missing parameters of organic molecules [274]. This program served for the parametrization of the SAM analogues.
- **Metal Center Parameter Builder (MCPB):** used to build models of metal proteins. It generates input files to use on QM packages such as Gaussian [275] and uses the output to generate inpcrd and prmtop files of the complexes. This functionality was used for the parametrization of the Zn metal centers in G9a.
- **LEaP:** generates force fields compatible with the Amber MD packages [272]. It was used to merge together the information contained in mol2, frcmod, pdb and other files to generate the topology and coordinate files of the protein-ligand complexes needed to run MD simulations in Amber.
- **CPPTRAJ:** generates multiple types of analysis from previously generated MD trajectories, such as RMSD vs time, distances, angles, cluster analysis etc. [276]. It was used as a post-analysis tool to assess which SAM analogues are more likely to facilitate the methyl transfer reaction.
- **MMPBSA:** estimates the binding-free energies between different molecules based on MM/Poisson-Boltzmann Surface Area calculations (MM/PBSA). It was used to generate a fast and accurate estimation of relative binding affinity between different enzyme-cofactor pairs [248].

A description about the methodology followed for this project is shown in Figure 5-3.

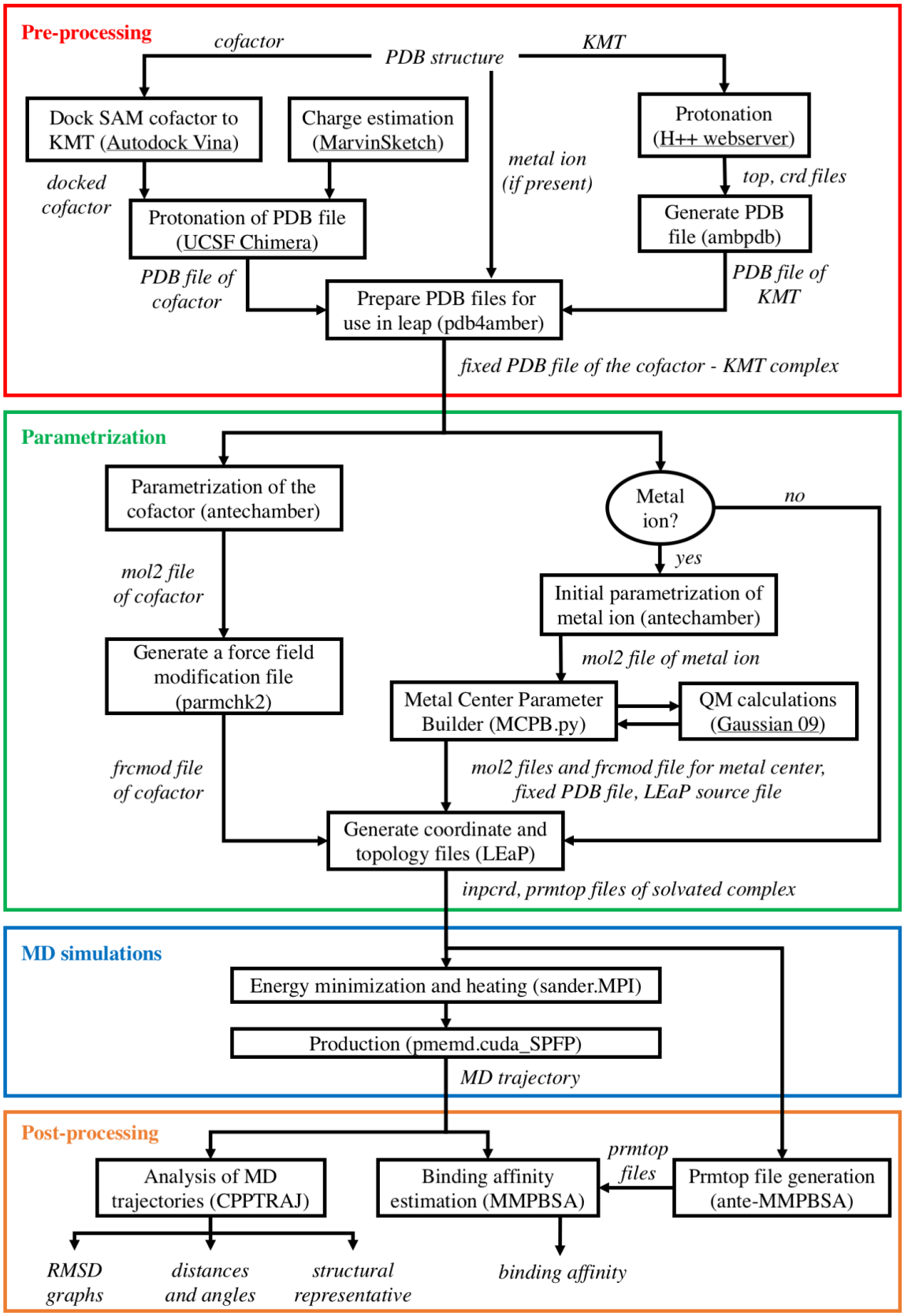


Figure 5-3. Methodology followed for the pre-processing (red), parametrization (green), MD simulations (blue) and post-processing (orange) of the cofactor-KMT pairs under study. All the programs that are external to the Amber16 suite appear underlined.

5.2.1. Pre-processing

5.2.1.1. Pre-processing the cofactor structures

The protonation and charge state of each of the SAM cofactor analogues at pH = 7.4 was calculated with the pKa estimation functionality of the program MarvinSketch [277]. PDB files of SAM cofactor analogues were generated from an available crystallographic structure of SAH with G9a (PDB: 2O8J). To generate the SAM cofactor analogues in the PDB format, methyl groups were added with PyMOL to convert SAH into SAM, and different groups were inserted at C2 position depending on the cofactor using the "build" functionality of PyMOL.

To avoid clashes between the SAM cofactor analogues and their receptors that could lead to high energy perturbations in the MD simulation, the SAM cofactor analogues were docked to each receptor using Autodock Vina [219]. Prior to docking, the receptor structures were prepared using Dock Prep in UCSF Chimera [278]. The specifications used were: replace incomplete side chains using the Dunbrack rotamer library, add hydrogens, and assign charges according to Amberff14SB for standard residues. Autodock Vina was then run for each of the KMT (G9a and SETD7) and cofactor (SAM, piperidine-SAM, methylpiperazine-SAM, thiophenyl-SAM, phenyl-SAM) pairs, using a 25 $Å^3$ box in the SAM binding site as the receptor search volume. The selected docked structures of SAM were those presenting the lowest energy among the structures that kept a similar conformation to the experimental binding mode of native SAM. The selected structures for the SAM analogues were then saved as separate PDB files.

Finally, UCSF Chimera [278] was used to change the total charge and protonation state of the PDB of the docked SAM analogues according to the results of the pKa calculation in MarvinSketch.

5.2.1.2. Pre-processing the KMT structures

Structures of G9a and SETD7 were isolated from crystal structures in the PDB. The criteria to choose the structures were: best resolution, highest number of residues around the SET

domain and presence of SAH/SAM. For those reasons, the B chain of structure 2O8J [65] was chosen for G9a (resolution: 1.8 Å, residues 922-1192), whereas chain A of structure 1O9S [40] was chosen for SETD7 (resolution: 1.75 Å, residues 117-366). Regarding the standard residues of the protein, protonated and charged topology (top) and coordinate (crd) files were generated using the webserver H++ [279], specifying pH=7.4. The topology and coordinate files were then inputted into the program ambpdb (Ambertools) to generate a PDB file of the standard residues with the Amber naming scheme.

5.2.1.3. Preparing the PDB file of the cofactor-protein complex

The PDB file of each docked SAM cofactor analogue was combined with the one of each KMT (and the metal ion, if present) into a single PDB using the cat command in Unix. Then, the PDB file was renamed for its use in LEaP using pdb4amber (Ambertools), giving as a result a fixed PDB file of the SAM cofactor analogue - KMT complex.

5.2.2. Parametrization of the complexes

5.2.2.1. Parametrization of the SAM cofactor analogues

The PDB files of the SAM cofactor analogues were used as input files in Antechamber to generate their corresponding mol2 files. The AM1-BCC charge method was specified and the GAFF force field and atom types were used, specifying the total charge as the one predicted by MarvinSketch. Then, the program parmchk2 read in the generated mol2 files and used its default atomtype similarity index file and force field file (*.dat) to generate a force field modification (frcmod) file containing any force field parameters that are needed for the molecule but not supplied by the force field (*.dat) file.

5.2.2.2. Parametrization of the G9a complex

In order to run MD simulations, it is necessary to first obtain the topology (prmtop) and coordinates (inpcrd) files of the solvated complexes. The steps to obtain these files differ

for G9a and SETD7, as G9a contains a Zn metal site. Whenever a metal is present, it is necessary to specifically parametrise the protein atoms involved in it as the partial charges of the residues surrounding the metal are modified. In this project, the ZnCys motif present in G9a was parametrised by using the Metal Center Parameter Builder program included in Ambertools [275]. This program acts as a bridge between the quantum mechanical (QM) software Gaussian [280] and the programs of the Amber suite [269].

Antechamber was used to generate the initial mol2 file of the metal ion. Once the mol2 file is generated, the charge was modified to have a value of +2.0, as it is the charge value Zn adopts in biomolecules [281]. The PDB and MOL2 files of the cofactors were obtained as described in section 5.2.2.1. . The parametrization of the metal site with the program Metal Center Parameter Builder (MCPB.py) will then alter the partial charges of both the Zn ion and the cysteine residues involved in the ZnCys motif due to their proximity to the metal.

The PDB file of the complex, the mol2 files of metal and cofactors and the frcmod files of the cofactors were all inputted into MCPB.py to generate the PDB and fingerprint models for the small, standard and large models of the system, as well as the Gaussian input files for the small and large models. To generate these files, the input file for MCPB.py specified a bond cut-off of 2.8 Å, with a geometry optimization of the hydrogen positions of the large model for the RESP calculation and selecting Gaussian09 [280] as the program to perform the quantum mechanics (QM) calculations.

A bonded model of the metal ion with the surrounding cysteines was chosen as it is less computationally expensive than the non-bonded model [275]. The bonded model considers the interaction between the metal ion and the involved residues through bond, angle, dihedral, electrostatic and van der Waals (VDW) terms [275]. Therefore, the metal center is simulated as bound to the surrounding residues. The bonded method has also been shown to present the best performance when compared to the non-bonded due to the complexity of the second [282].

Gaussian09 calculations included geometry optimization, force constant calculation and generation of the fchk file for the small model, performing the Merz-Kollman RESP charge calculation for the large model to generate its chk file. Once all Gaussian09 output files were generated, MCPB.py was used to generate the frcmod file for the metal center to use later for LEaP. For this step, the force field parameters were generated using the Seminario method. In this method, the bond and angle force constants are determined from the sub

matrices of the Cartesian Hessian matrix. The RESP charge fitting was performed to generate the mol2 files for the metal site residues with the option ChgModB. ChgModB restrains the charges of the backbone heavy atoms (CA, N, C, O) to the values in AMBER FF94. The combination of Seminario to generate the frcmod files + ChgModB for the mol2 files was chosen to achieve higher simulation accuracy as recommended by Peters et al. [283].

The last step prior the use of LEaP was to generate the LEaP source file with MCPB.py. This file contains information about the PDB file of the system, the force fields to use, the size of the solvation box, which ions to add to equilibrate the charges of the system, and the mol2 and frcmod files for the molecules that require them for a specific parametrisation (e.g. non-standard residues). The source file generated by MCPB.py was modified in two points: a line was added so the mol2 file of SAM was identified, and another line was modified from "solvatebox mol TIP3PBOX 10.0" to "solvateoct mol TIP3PBOX 15.0" to obtain an octahedral solvation shell instead of the cubic one generated by the solvatebox command, with the edge of the octahedron at least 15Å away of any atom originally present in solute. The octahedral shape for the solvation shell was chosen because it resembles more a sphere than the cube, so the distance between the complex and the solute shell boundary is more homogeneous given that G9a is a globular protein. This selection also represents a gain in efficiency, as less solvent molecules need to be accounted for the simulation. For the same reasons the octahedral shape was extended to SETD7. In addition, Na^+ ions were added as counter ions to compensate the existing negative charge of the system.

LEaP was then run calling for the generated input file and all the generated PDB, mol2 and frcmod files to produce the prmtop and inpcrd files of the solvated complex. This protocol was followed for G9a and five SAM cofactor analogues, for a total of 5 SAM cofactor analogue – G9a systems.

5.2.2.3. Parametrization of the complexes of SETD7 with LEaP

The protocol to generate the prmtop and inpcrd files for SETD7 differs from the one for G9a as there is no metal center in the SET domain of SETD7. The fixed PDB file of the SAM cofactor analogue – SETD7 complex generated with pdb4amber as described in section 5.2.1.3. were inputted alongside the mol2 and frcmod files of the SAM analogues generated as described in section 5.2.2.1. to produce the prmtop and inpcrd files of the

solvated complex. Other specifications were the use of the ff14SB force field and the TIP3P water model. A solvated octahedron was generated with a minimum distance of 12Å between any atom of the complex and the edge of the periodic box. Na^+ ions were added as counter ions to compensate the existing negative charge of the system.

5.2.3. Molecular dynamics simulations

Once the prmtop and inpcrd files of the solvated complexes were generated as specified in section 5.2. , MD simulations were performed. The minimization and heating steps were computed using the sander.MPI simulation engine on 20 CPUs. The CPUs used were Dual socket Intel Xeon E5-2660 v2 with 128GB of memory running Centos 7.3. The production runs were computed using the CUDA-accelerated version of Amber16 that includes the SPFP precision model, named pmemd.cuda_SPFP [270]. This precision model boosts performance on GPUs without sacrificing numerical accuracy. The GPUs used for the production simulations were Tesla K80. Using these resources, each of the production runs took 6 days to be completed. All simulations were performed using the AMBER ff14SB force field [270]. The same protocol was applied to all the systems included in the study.

The system was initially minimised with a 2 kcal mol^{-1} $Å^{-2}$ restraint on protein and cofactors and unrestrained movement for water and counter ions. This was followed by another minimization, this time with no energetic restrictions. For both minimizations, a maximum of 10000 steps of minimization were set, switching from the steepest descent algorithm to conjugate gradient after 3000 steps. The non-bonded interaction cut-off was set to 12 Å.

After minimization, two heating steps were carried out. The first heating step involved an NVE ensemble (conserved amount of substance, volume and energy) with a linear increase in temperature from 0 to 100K during 500 ps with 10 kcal mol^{-1} $Å^{-2}$ energy restraints on the protein and cofactor. The second heating step was performed at constant pressure and without any energy restraints, increasing the temperature linearly from 100 to 310K during 500 ps. The heating was followed by production runs in the NPT ensemble (conserved amount of substance, pressure and temperature) at 310K including a Monte Carlo barostat to control the pressure at 1 bar.

Both heating and production steps shared some common attributes. For temperature scaling, Langevin dynamics with collision frequency of 1 ps^{-1} were set. The random seed for this Langevin thermostat was given by the date and time of the start of the simulation. Bonds involving hydrogen were constrained using SHAKE bond length constraints to remove the hydrogen motion. This enabled to increase the time step from 1 fs to 2 fs without introducing instabilities caused by hydrogen vibrations [284]. The non-bonded interactions cut-off was set to 10.0 Å. The specific commands used are shown in the Appendices section (Figure A-1, Figure A-2, Figure A-3, Figure A-4 and Figure A-5).

5.2.4. CPPTRAJ post-analysis of the MD trajectories

The generated MD trajectories were processed with the program CPPTRAJ [276] and other programs included in the AmberTools16 package [270]. Further analyses of the results were performed using the Python programming language, plotting the data with the libraries matplotlib [285] and seaborn [286]. Xmgrace [287] was used to generate the RMSD plots.

5.2.4.1. RMSD Graphs

The RMSD for backbone atoms of each protein (CA, C, N) against the initial frame of the production trajectory was computed with CPPTRAJ and plotted with Xmgrace for the whole trajectory (100 ns).

5.2.4.2. Distances and angles

The hydroxyl group of the tyrosine Y1067 of G9a or Y245 of SETD7 hydrogen bonds to the methyl-accepting nitrogen of each respective histone lysine (H3K9 for G9a and H3K4 for SETD7). This interaction inhibits trimethylation for G9a [65] and di- or trimethylation for SETD7 [40]. The tyrosine hydroxyl group (OH) is located 3 Å away for G9a's closely related protein GLP and 2.9 Å away for SETD7 to the methyl-accepting nitrogen (N) of the histone lysine for the endogenous systems with native SAM. Therefore, we hypothesise that the distances formed between the oxygen of the hydroxyl group of the tyrosines with the carbon of the methyl transfer group, and the angles formed between these two atoms and the sulphur

atom of SAM must be comparable to the ones obtained with endogenous SAM to enable the catalytic activity. A visual explanation of this rationale is shown in Figure 5-4.

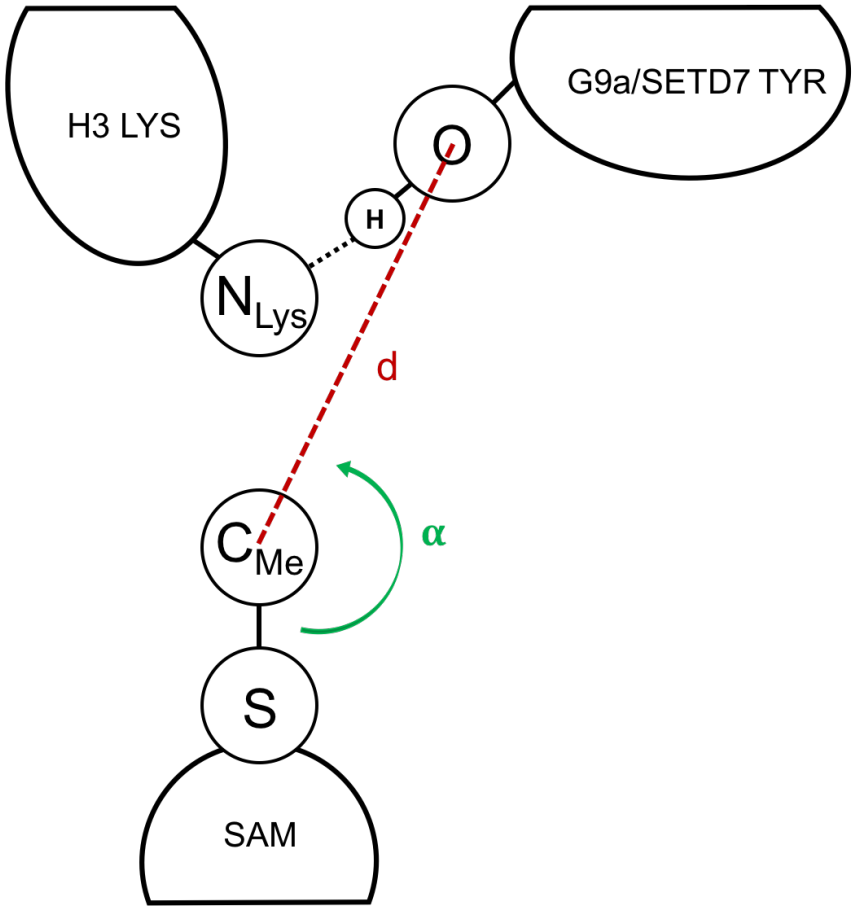


Figure 5-4. Distance (d) between the oxygen of the hydroxyl group of a G9a/SETD7 catalytic tyrosine with the carbon of the methyl transfer group and angle (α) formed between these two atoms and the sulphur atom of SAM.

To obtain the distance d and angle α shown in Figure 5-4 from the production trajectory, the functions "distance" and "angle" were used in CPPTRAJ. Then, both distances and angles for the last 80 ns of the production trajectory were used to generate density distribution graphs. This enabled the comparison of the angle and distance values obtained with SAM cofactor analogues with the ones obtained for native SAM.

5.2.4.3. Clustering of MD trajectories

Cluster analysis was performed on the production run with CPPTRAJ. This analysis was performed to determine the populations of the structures from the production trajectories and obtain a structural representative of the major cluster. To perform the cluster analysis,

water molecules and counter ions were removed from the generated MD trajectories to decrease their size and therefore the time of the computation. Then, the clustering algorithm DBSCAN was used with a minimum of 25 points to form a cluster. The decision to use DBSCAN was based on reducing the user bias, as the number of clusters to be obtained are not pre-specified. The distance cut-off for forming a cluster was set to 1.5Å for G9a and 1.2Å for SETD7, the reason being obtaining a similar number of clusters for both proteins. The distance metric was the RMSD of the backbone atoms (CA, C, O, N, H), using 10 initial frames for clustering. The specific commands used can be found in the Appendices, Figure A-6.

5.2.5. MMPBSA calculation of binding affinities

5.2.5.1. Generating prmtop files for the binding affinity estimation

The MM-PBSA method requires the following files to estimate the binding affinities of protein-cofactor complexes:

- An input file specifying the frames of the trajectory to analyse, the implicit solvation method to calculate the binding affinity (e.g. Poisson-Bolzmann) and parameters for that method.
- The production trajectory (nc) files previously generated.
- Prmtop files for the solvated complex (already generated) and for the non-solvated complex, non-solvated receptor and non-solvated cofactor.

Ante-MMPBSA.py was used to generate the prmtop files for the non-solvated complex, non-solvated receptor and non-solvated cofactor. The prmtop file of the solvated complex was inputted. Waters and Na^+ ions were stripped from the solvated complex (waters, Na^+), also specifying which atoms correspond to the cofactor. The Poisson-Bolzmann (PB) Radius set in the generated topology files was mbondi2. Following these guidelines, prmtop files for the non-solvated complex, non-solvated receptor and non-solvated cofactor were generated.

5.2.5.2. Running MMPBSA.py

The MMPBSA calculations were performed for each KMT-SAM cofactor analogue pair. The last 80 ns of the MD production trajectories were inputted into the program MMPBSA.py (Ambertools). As the total production run (100ns) represents 10000 frames, the last 80 ns of the simulation are represented by the frames 2001 to 10000. An interval of 80 frames was set to get more uncorrelated snapshots for a total of 100 frames evaluated. The entropy contributions were not computed as the purpose of this evaluation is to compare the binding affinity of cofactors binding to the same protein, which are expected to have a similar entropy contributions. Besides, entropy calculations are computationally expensive and tend to introduce a large margin of error in the final result [267]. Regarding the PB calculation, the non-polar optimization method used the solvent-accessible-surface area to correlate the nonpolar solvation free energy (method 1), the atomic radii was set according to the prmtop file (method 0), and an ionic strength of 0.1M was used.

5.3. Results and discussion

The goal of this chapter is to get molecular insights on the interaction of SAM cofactor analogues with G9a and SETD7, in order to select which enzyme-cofactor pairs could be more efficient regarding methylation activity. To assess this, the binding mode of the proposed SAM cofactor analogues was studied regarding distances, angles and binding affinities and comparing those values to the ones obtained with native SAM.

5.3.1. Pre-processing results

5.3.1.1. Protonation state of the SAM analogues major microspecies

The protonation states and structures for the major microspecies of SAM and SAM analogues were calculated at 310K and pH = 7.4 using the pKa functionality of Marvinsketch. The results obtained are shown in Figure 5-5.

The structure with more relative abundance at the specified conditions was the one selected for docking. The main characteristic of the major microspecies is the fact that the nitrogen of the methionine group appears protonated, leading to a total charge of +1. For all the cofactors except MPip-SAM, the first species has a distribution above 99%. However, for MPip-SAM, there exists an almost equal abundance between the conformation that is non-protonated at the nitrogen attached to the methyl group in the methylpiperazine (S3.1, RA: 50.67%) and the one that is protonated at the same nitrogen (S3.2, RA:48.37%). In order to reduce the simulation workload (6 days on Tesla K80 GPUs), only the non-protonated form (S3.1) was considered. The protonated form should be studied in future simulations.

5.3.1.2. Atom types and partial charges of the SAM analogues

Once the protonation state was estimated, the mol2 files were generated with Antechamber (see section 5.2.2.1. for reference). These files contain information about the charge and Amber atom types of each of the atoms of SAM or the SAM analogues. This information is shown on the structures of the cofactors in Figure 5-6.

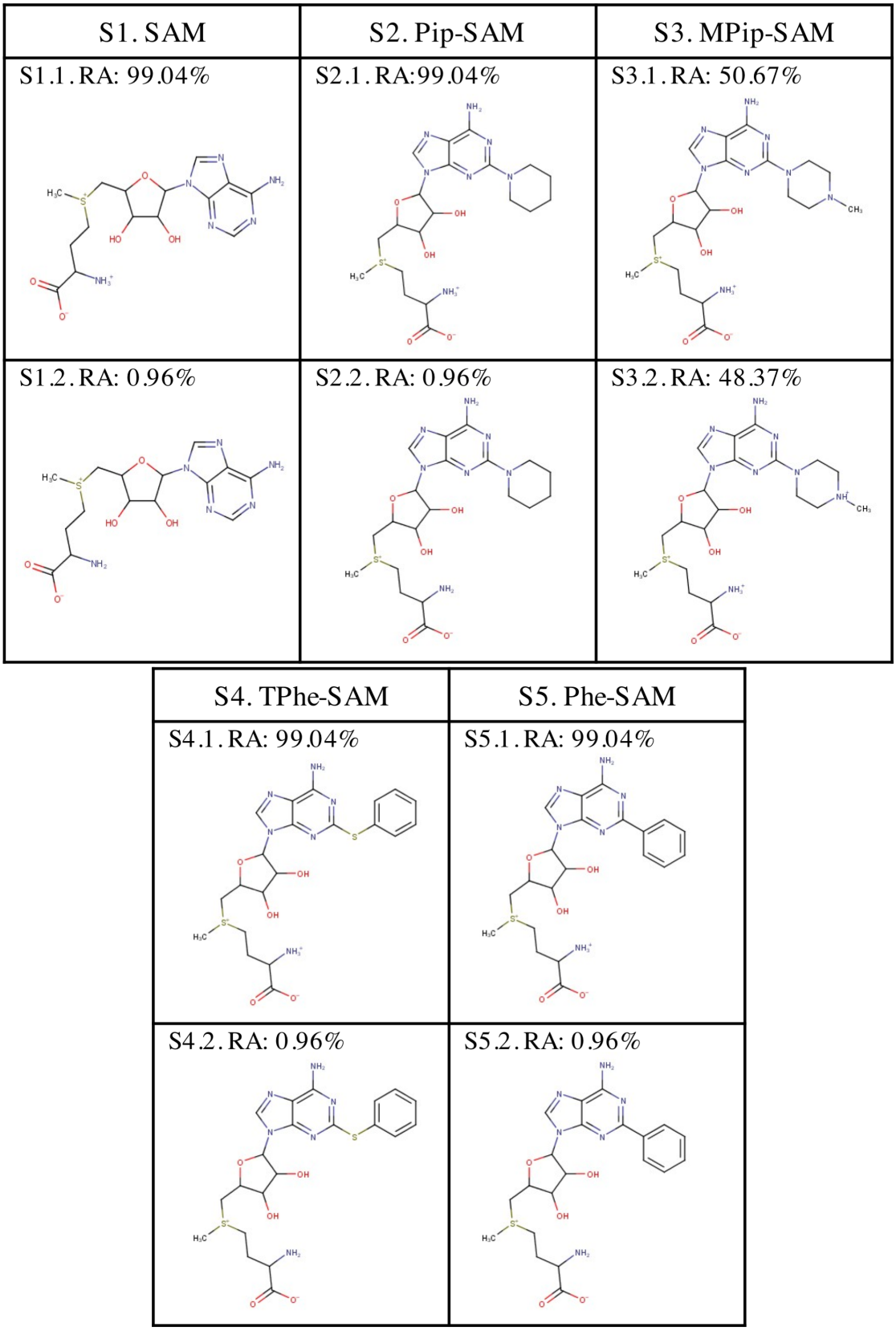


Figure 5-5. Relative abundance (RA) and protonation state prediction of the major microspecies of native SAM, Pip-SAM, MPip-SAM, TPhe-SAM and Phe-SAM at pH = 7.4 and T = 310K.

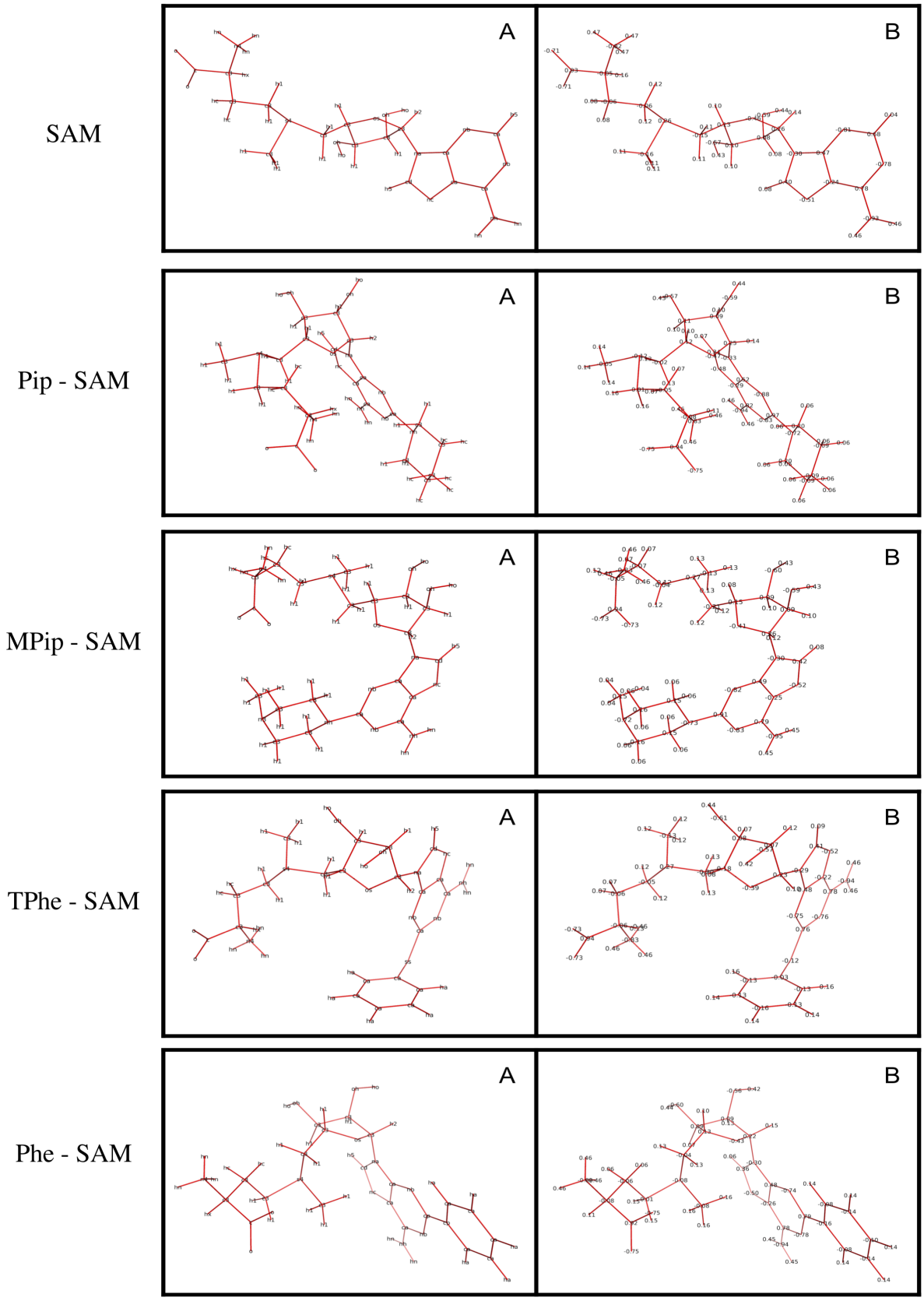


Figure 5-6. Amber atom types (A) and partial charges (B) for each of the atoms of SAM, Pip-SAM, MPip-SAM, TPhe-SAM and Phe-SAM.

The mol2 files of the SAM-based cofactors with the charges and atom types shown in Figure 5-6 were later inputted into MCPB.py for G9a as mentioned in the methods section 5.2.2.2. . These files were also inputted into LEaP to generate the prmtop and inpcrd files for the solvated complex of all the experiments.

5.3.1.3. Parametrization of the ZnCys motif

For G9a there exists a ZnCys motif that pins the post-SET domain to the main body of the SET domain. This ZnCys motif was parametrised with MCPB.py as specified in section 5.2.2.2. . More specifically, we used the bonded model for structural zinc by explicitly bonding the Zn^{2+} ion with its surrounding CYS residues: C1115 in G9a's SET domain and C1168, C1170 and C1175 in G9a's post-SET domain. The bonding between these residues and the metal is shown for native SAM with endogenous G9a in Figure 5-7. This parametrization was extended to all the experiments involving G9a.

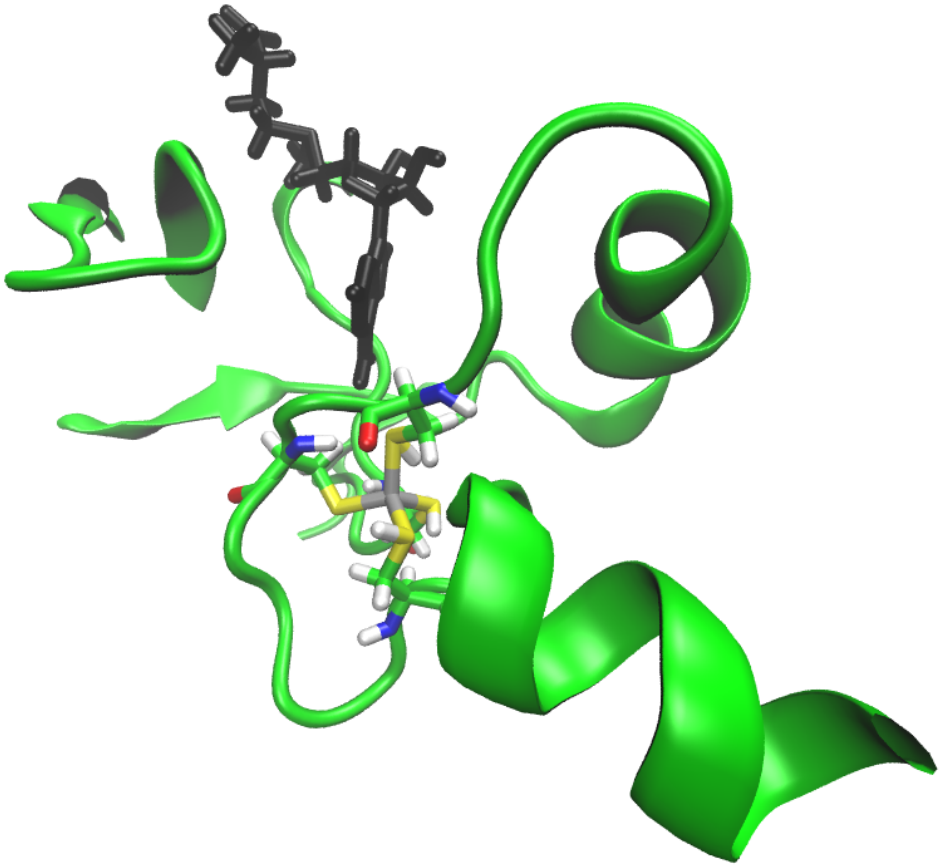


Figure 5-7. G9a's Zn motif parametrised by MCPB.py using the bonded model. The Zn atom is shown in grey, the S atoms of the CYS surrounding the Zn in yellow and G9a in green. Native SAM is shown in black for reference.

The parametrization of the metal site with MCPB.py alters the partial charges of both the Zn ion and the cysteine residues involved in the motif due to their proximity to the metal. From the original Zn charge of +2, the positive charge was distributed into the cysteines so the actual partial charge of the Zn atom resulted in 0.123489 for G9a.

5.3.1.4. Prmtop and inpcrd files for the solvated enzyme-cofactor complexes

All the PDB, mol2 and frcmod files generated were inputted into LEaP to generate the prmtop and inpcrd files of each solvated complex. This protocol was followed for G9a and SETD7, each with native SAM plus the four SAM analogues, for a total of 10 systems. The outputs of this calculation were prmtop and inpcrd files of the different enzyme-cofactor complexes, including the metal center for G9a, and equilibrated using Na^+ ions and a water solvation shell.

Periodic boundary conditions were used in Amber, meaning that the model is composed of an infinite number of copies of the primary cell. Amber considers one copy of the atoms, but it also calculates the forces between that system and other atoms in the infinite system that are relevant to the calculation. For example, the atoms within the cut-off distance for the van der Waals interaction, or all atoms for electrostatics. As mentioned in section 5.2.2.2. , the shape chosen for the solvation shell of the cell was octahedral due to its resemblance to a sphere. As an example, the primary cell for the system of endogenous G9a with native SAM is shown in Figure 5-8.

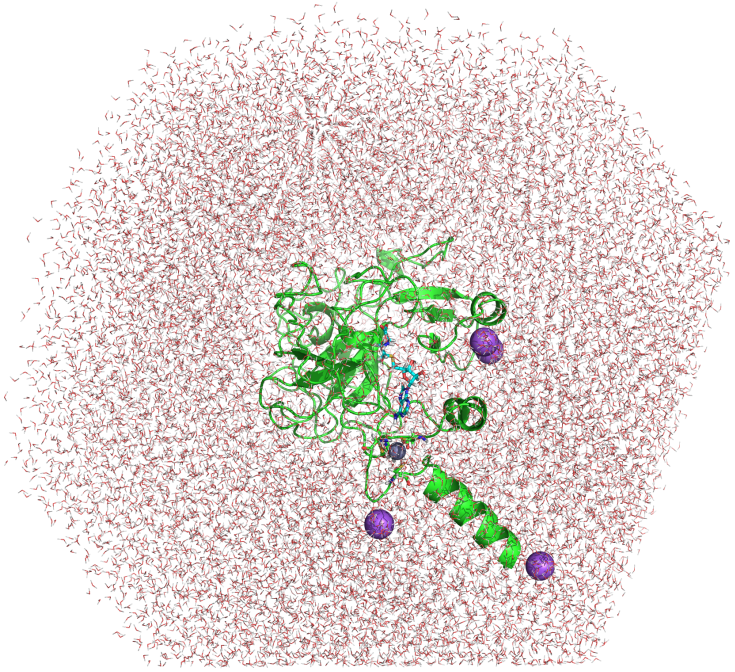


Figure 5-8. Primary cell of the simulation of endogenous G9a with native SAM. G9a is shown in green, SAM in blue, Na^+ counter ions in purple, Zn atom in grey and water molecules in red.

5.3.2. RMSD evolution of the protein backbone

To assess the overall stability of the complex and to obtain an estimate of when the systems become equilibrated, the root-mean-square-deviation (RMSD) of the protein backbone atoms (CA, C and N) was calculated with regards to the initial production frame for the different enzyme-cofactor pair production runs as shown for G9a and SETD7 in Figure 5-9.

Figure 5-9 shows that the RMSD of the runs increased during the first 20 ns, but it remained quasi-stable for most of the cases from there on with the exception of Pip-SAM and G9a, where longer simulation times could be performed to ensure convergence. The average RMSD values oscillate between 2.5 and 5 Å for G9a whereas the RMSD changed between 1.5 and 2.5 Å for SETD7. This is a first indicator of the higher adaptation needed for G9a to accommodate the modified SAM analogues from the initial docked structures when compared to SETD7. Because of the existence of a systematic increase in the RMSD in the first 20 ns of each run and to compare equal ranges between the different KMTs, the first 20 ns were discarded for the calculation of probability density distributions of distances and angles as shown in the next section 5.3.3. .

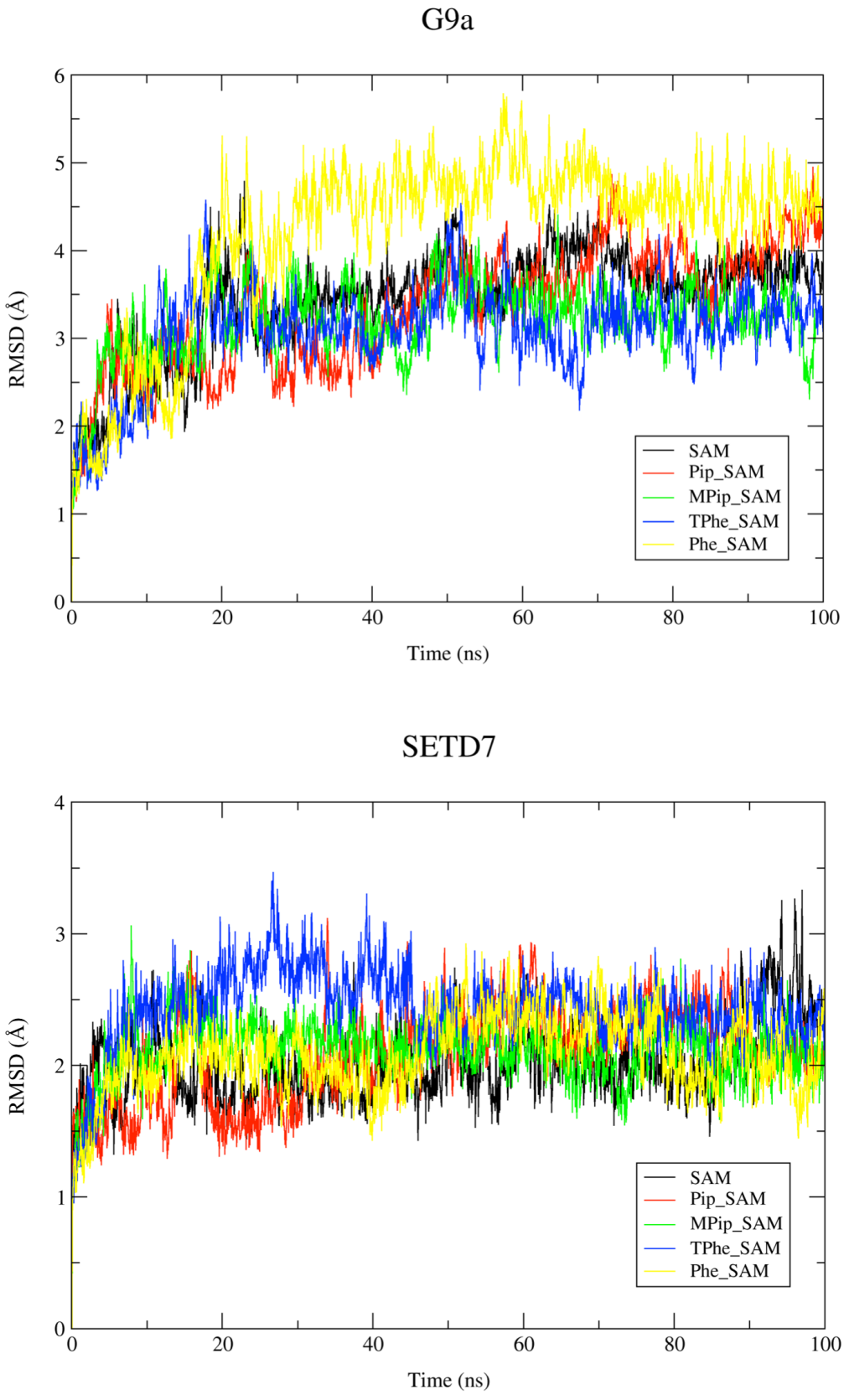


Figure 5-9. Root-mean-square-deviation (RMSD, Å) for the backbone atoms of G9a and SETD7 with regards to the initial frame for the production trajectory.

5.3.3. Distances and angles to enable methyl transfer

For endogenous G9a and SETD7, the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 for G9a or Y245 for SETD7 ("the distance") was measured on their crystallographic structures (PDB:2O8J for G9a and PDB: 1O9S for SETD7), as shown in Figure 5-10. The methyl group was added with PyMOL as the crystallographic structures contain SAH and not SAM.

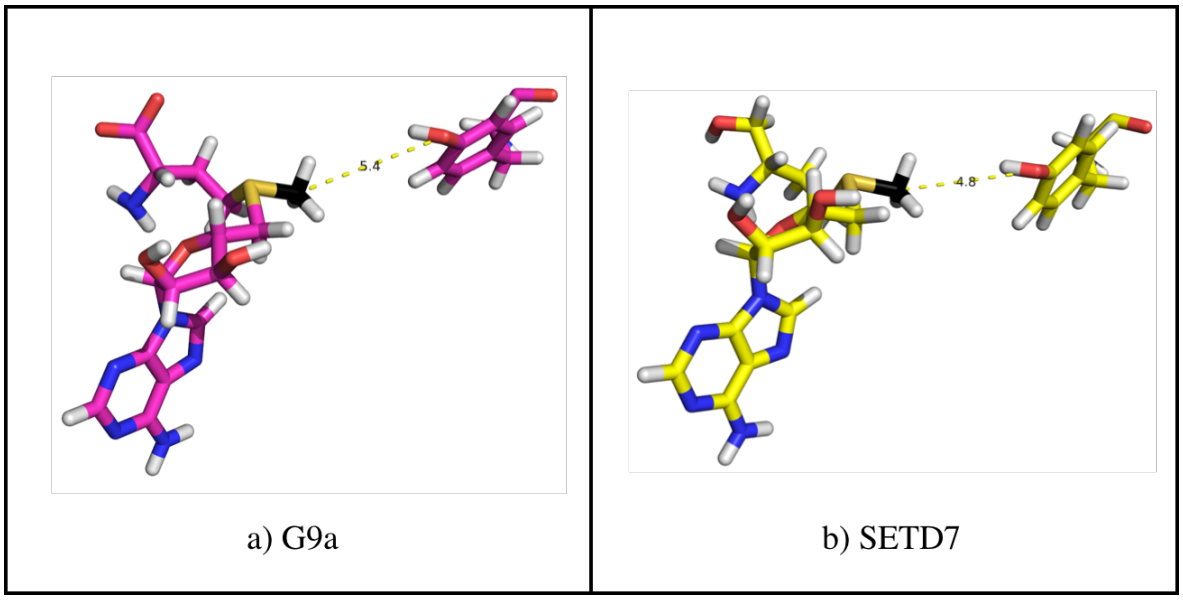


Figure 5-10. Experimental distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 in G9a (a) or Y254 in SETD7 (b).

The distance measured between the oxygen atom of the hydroxyl group of the catalytic tyrosine and the carbon atom of the methyl transfer group of SAM was 5.4 Å for G9a and 4.8 Å for SETD7. The angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 for G9a or Y245 for SETD7 ("the angle") was measured using the get_angle functionality of PyMOL. These angles resulted 147.413º for G9a and 158.446º for SETD7. The distances and angles measured in the crystallographic structures served as benchmark of the distances and angles obtained through the MD simulations with native SAM.

5.3.3.1. G9a distances and angles analysis

As shown in Figure 5-11, the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 in G9a is predominantly around 5.5 Å for native SAM. This agrees with the measured distance on the crystallographic structure (5.4 Å), and it is therefore considered to be the distance enabling the methyl transfer. However, when this distance is analysed with the SAM cofactor analogues proposed, it becomes much higher: between 7 and 12 Å for Pip-SAM, around 12 Å for MPip-SAM, between 5 and 12 Å for TPhe-SAM (unclear, wide distribution in this case) and around 8 Å for Phe-SAM. Therefore, assuming that the ideal distance is the one of SAM, the methyl transfer should become much worse with any of the SAM cofactor analogues.

A similar approach was performed for the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 in G9a, and the results are shown in Figure 5-12. The angle obtained between the aforementioned atoms with native SAM is predominantly around 150º, comparable to the experimental 147.413º. However, angles obtained with SAM analogues were substantially different: angles around 30 and around 120º for Pip-SAM, 120º for MPip-SAM, 150º for TPhe-SAM and 85º for Phe-SAM. Therefore, the only SAM cofactor analogue that seems to provide comparable angles for the methyl transfer compared to native SAM is TPhe-SAM.

In conclusion, analyzing both distance and angles together, it seems that none of our SAM cofactor analogues seem to resemble both the distance and angle that allow an efficient methyl transfer with native SAM. In the case of TPhe-SAM, the angle is similar to the one of native SAM, but as the distance distribution is much wider, it is expected that the methyl transfer is also less effective. TPhe-SAM also appeared to be the SAM analogue showing higher experimental heavy methylation with G9a (see section 4.3.2.5. , which further validates the use of MD simulations to predict biological outcomes. In conclusion, and as expected from the design, none of the SAM cofactor analogues provide distances and angles for methyl transfer comparable to native SAM.

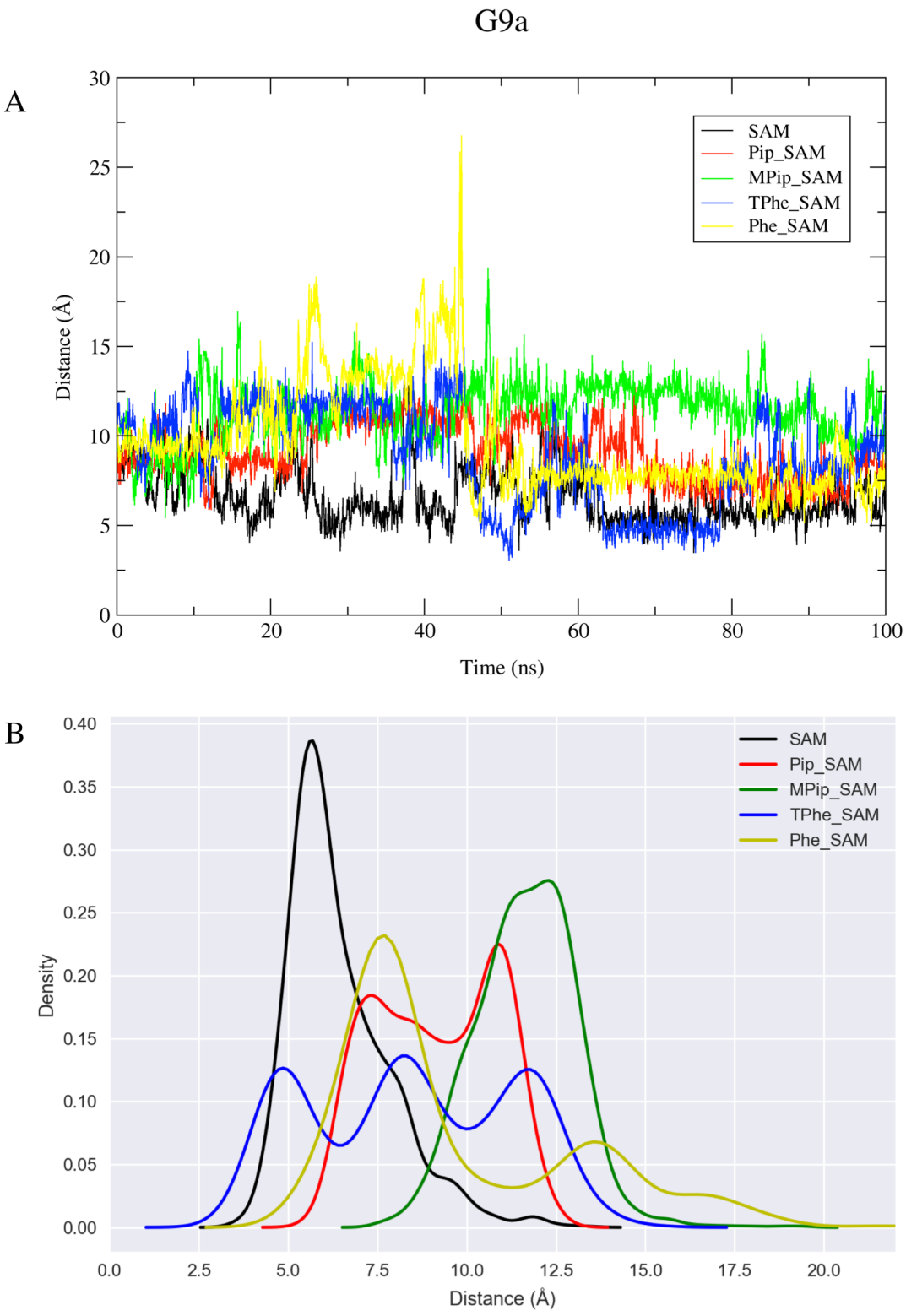


Figure 5-11. A) Evolution of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 during 100 ns of simulation. B) Probability distribution function of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 for the last 80 ns of the simulation.

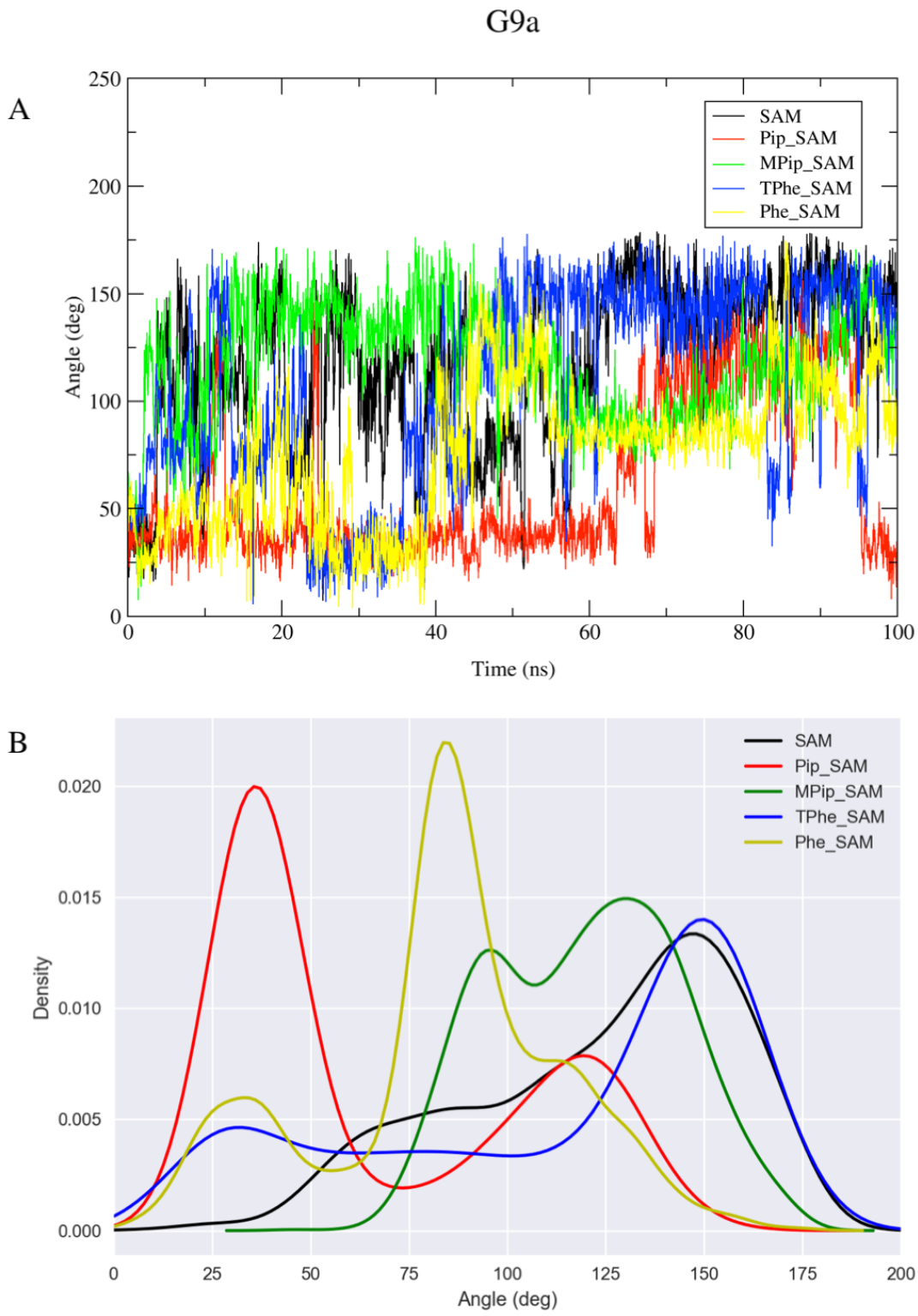


Figure 5-12. A) Evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 during 100 ns of simulation. B) Probability distribution function of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a's Y1067 for the last 80 ns of the simulation.

5.3.3.2. SETD7 distances and angles analysis

As Figure 5-13 shows, the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y245 in SETD7 is predominantly around 4.5 Å for native SAM, which resembles the experimental value of 4.8 Å. This is therefore considered to be the distance enabling the methyl transfer. The distance calculated for each of the SAM cofactor analogues is around 5 Å for Pip-SAM, 7.5 Å for MPip-SAM, 10 Å for TPhe-SAM and two clusters around 6 Å and 9.5 Å for Phe-SAM. Therefore, Pip-SAM presents similar distances to the ones of SAM, increasing in distance for Phe-SAM, MPip-SAM and TPhe-SAM, with the last two apparently too far to provide an adequate methyl transfer.

A similar approach was followed to analyse the evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y245 in SETD7. The results are shown in Figure 5-14. The angle formed between the previously mentioned atoms with native SAM is predominantly around 150º, comparable to the experimental 158.446º. This is assumed to be the angle that enables methyl transfer. The angles obtained with SAM analogues seem more favorable for SETD7 than for G9a in general. The angles varied depending on the cofactor: angles around 155º for Pip-SAM, around 110º (or 150º in a smaller cluster) for MPip-SAM, 90º for TPhe-SAM and between 80º and 140º for Phe-SAM. Therefore, the only SAM cofactor analogue that seems to provide comparable angles for the methyl transfer compared to native SAM is Pip-SAM.

In conclusion, analyzing both distance and angles together, it seems that the best SAM cofactor analogue to allow an efficient methyl transfer of SETD7 with native SAM is Pip-SAM, with a distance of 5 Å vs 4.5 Å in the native SAM and an angle of 155º vs 150º in the native SAM. This is also a good cofactor to differentiate between the activity of G9a and SETD7, because the distances and angles of G9a differed extensively between Pip-SAM and native SAM. In particular, section 5.3.3.1. showed a distance of 7-12 Å and an angle of 30 or 120º for Pip-SAM vs a distance of 6 Å and an angle of 150º for native SAM. This suggests that, when using Pip-SAM as a cofactor, a much more effective methyl transfer occurs with SETD7 than with G9a.

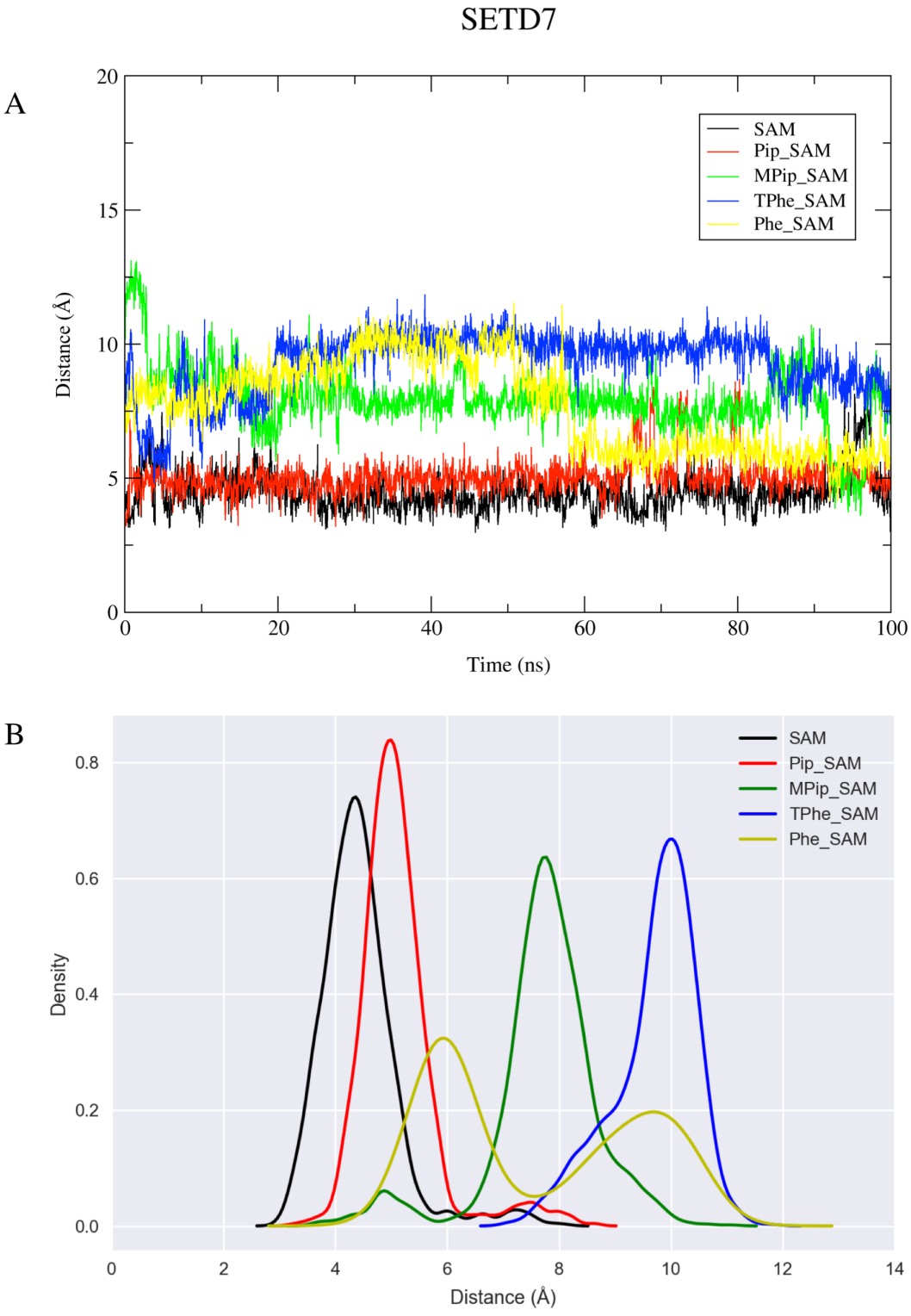


Figure 5-13. A) Evolution of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 during 100 ns of simulation. B) Probability distribution function of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 for the last 80 ns of the simulation.

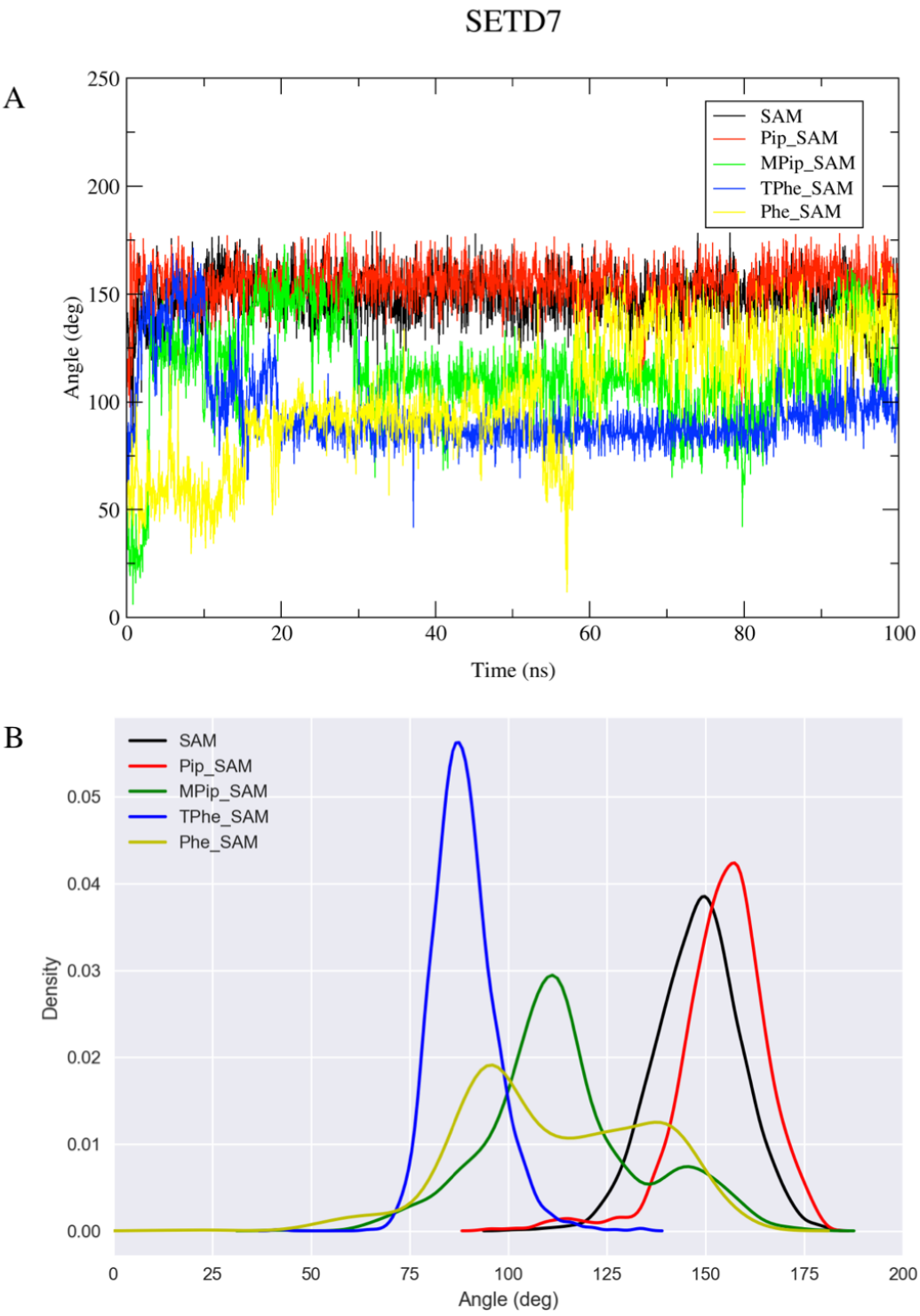


Figure 5-14. A) Evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 during 100 ns of simulation. B) Probability distribution function of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of SETD7's Y245 for the last 80 ns of the simulation.

Compared to the experimental results obtained in Chapter 4, the results obtained in the MD simulations agree for Pip-SAM, as this cofactor presented similar relative abundances of heavy monomethylation of H3K4 compared to native SAM. However, there is a lack of agreement for MPip-SAM, as experimentally it presented similar relative abundances of heavy monomethylation compared to native SAM or Pip-SAM, but different distances and angles in the MD simulation. We hypothesise this could be an artifact caused by the limited simulation time. Longer simulation times (e.g. 1 μs) could be run to investigate if there exists a conformational change later in the simulation that could indeed explain the experimental results.

5.3.4. Representative structures of the major clusters

The representative structures of the major cluster of each MD simulation were analysed with a focus on the enzyme-cofactor pair that was previously shortlisted for their orthogonal application, Pip-SAM and SETD7. Regarding SETD7, Figure 5-15 shows the differences in the binding of Pip-SAM to G9a and SETD7.

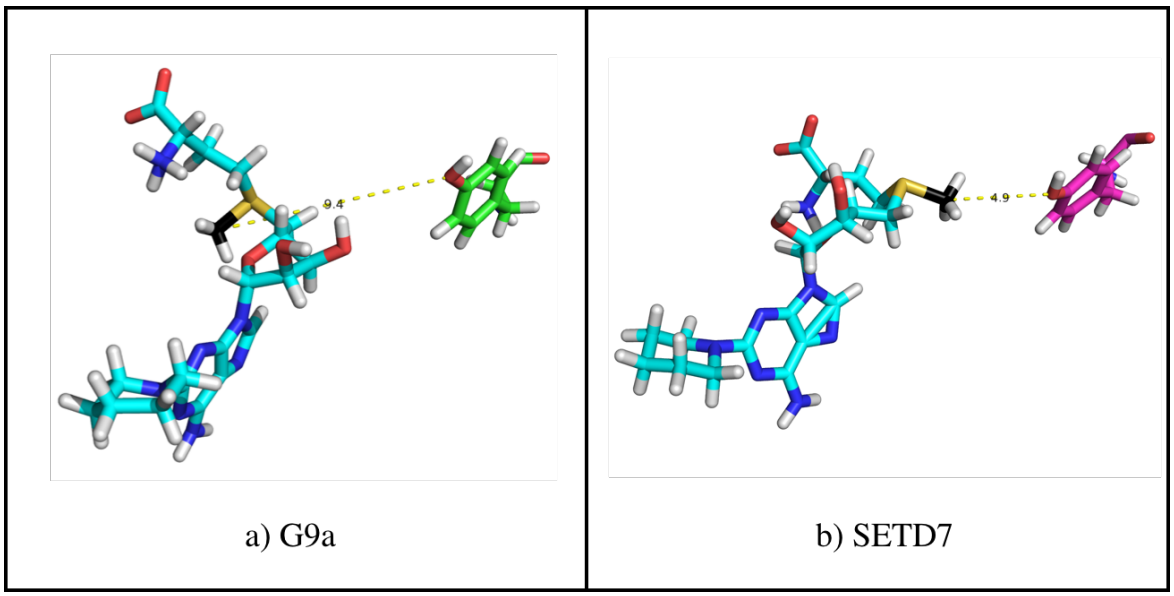


Figure 5-15. Differences in the relative positioning of Pip-SAM (cyan) to a) G9a's Y1067 (green) and b) SETD7's Y245 (magenta) based on their respective representative structures obtained from the MD production trajectories. Both the angle and the distance between Pip-SAM's methyl group (depicted in black) and the catalytic tyrosine are not adequate for G9a, but for SETD7 they resemble the values obtained with native SAM.

When observing the representative structures of the major clusters of the MD trajectories of G9a and SETD7 with piperidine-SAM, it could be observed that both the angle and the

distance between the methyl group (depicted in black) and the catalytic tyrosine are not adequate in G9a, but in SETD7 they resemble the values obtained with native SAM, possibly facilitating the transfer of the methyl group.

Furthermore, an adequate binding and interactions of the SAM analogue to SETD7 are required for the methyl transfer to occur. Therefore, the hydrogen bonding interactions between SAM and piperidine-SAM towards SETD7 were analysed using Ligplot [288] to assess if the interactions are preserved when using piperidine-SAM, using the representative structure of the most populated cluster for the analysis. The result of this analysis is shown in Figure 5-16.

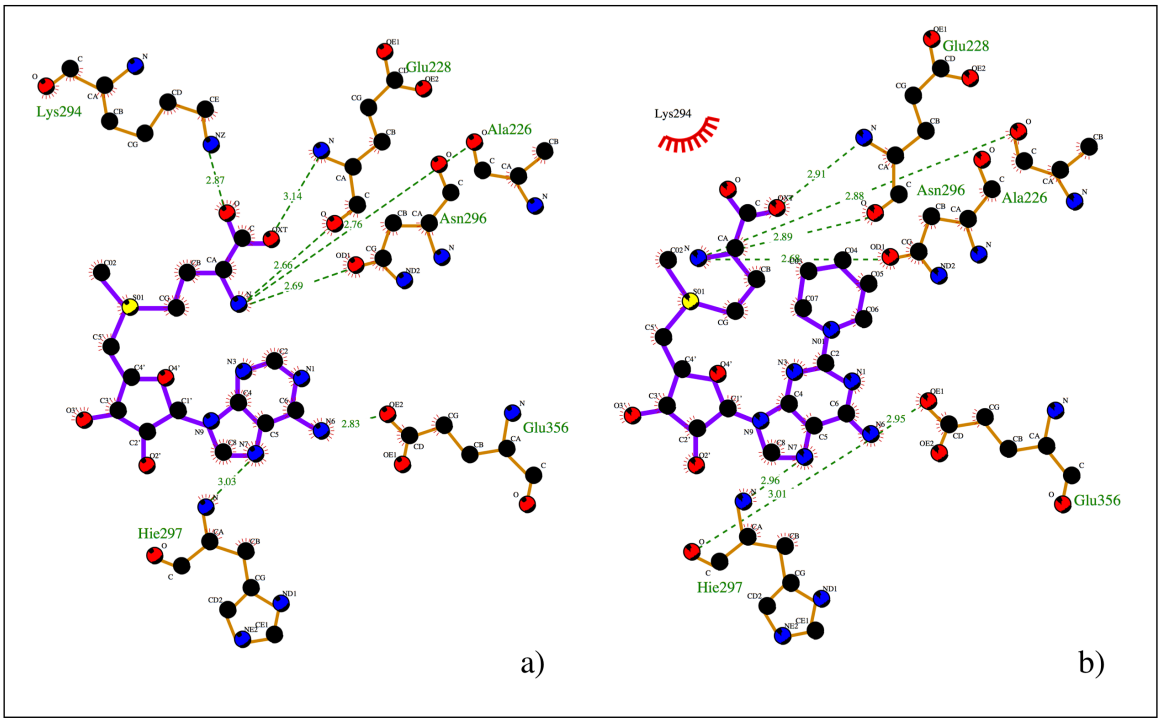


Figure 5-16. Comparative of the interactions of SAM (a) or piperidine-SAM (b) with SETD7.

As can be observed in Figure 5-16, the main interactions between SAM and SETD7 are preserved when exchanging SAM for piperidine-SAM. SETD7's K294 appears not to form a hydrogen bond when using piperidine-SAM, but it appears to be close to the methionine group. Therefore, it is possible that the interaction is preserved experimentally.

Together with the information generated about the preferential distances and angles of each SAM cofactor analogue shown in sections 5.3.3.1. and 5.3.3.2. , these results set piperidine-SAM as the preferential SAM cofactor analogue to differentiate between the

activity of G9a and SETD7. This result agrees with the one described in Chapter 4, sections 4.3.2.4. and 4.3.2.5. , where the *in vitro* assay results showed that piperidine-SAM preserved the activity of native SAM with SETD7 (in fact, showing a 6% higher methylation), but it dramatically worsens the results when used with G9a (up to 92% less methylation activity).

5.3.5. MM-PBSA calculation of the binding affinities

Besides the correct binding mode (distances, angles) of the SAM cofactor analogues with respect to their target proteins, it is also important to assess if the binding affinity of the selected cofactors is comparable to the one of native SAM. This is because a lower interaction could indicate weaker activity, or a much stronger interaction could suggest that the molecule might act as an inhibitor instead of a cofactor. The binding affinities of SAM cofactor analogues towards SETD7 were calculated with the program mmpbsa.py to elucidate if they are similar to the binding affinity of SAM. The calculation was done using the last 80 ns of the MD production trajectories, as specified in section 5.2.5. . A special focus was on Pip-SAM, as it appeared to be the best SAM cofactor analogue to resemble the characteristic distances and angles that allow an efficient methyl transfer in native SAM for SETD7 (see section 5.3.3.2. for reference). Table 5-1 shows a summary of the energetic analysis of SETD7 with different SAM cofactor analogues.

Table 5-1. Energetic analysis for the SETD7-SAM analogue complexes as obtained by MM–PBSA calculations.

Energy (kcal/mol)	S1	S2	S3	S4	S5
	SAM	Pip-SAM	MPip-SAM	TPhe-SAM	Phe-SAM
ΔE_{vdw}	-40.85	-48.72	-46.87	-45.09	-48.05
ΔE_{el}	-413.08	-407.14	-355.22	-366.43	-351.48
ΔE_{pol}	407.57	409.91	375.51	362.96	365.98
ΔE_{ap}	-4.05	-4.79	-4.90	-4.85	-4.90
ΔG_{gas}	-453.92	-455.86	-402.09	-413.42	-399.53
ΔG_{solv}	403.52	405.11	370.60	358.11	361.08
ΔG_{bind}	-50.40	-50.75	-31.48	-55.30	-38.45
$\Delta\Delta G_{MM\text{-}PBSA}$ (relative to SAM)	-	-0.35	18.92	-4.90	11.95

where ΔE_{vdw} is the van der Waals energy, ΔE_{el} is the electrostatic energy, ΔE_{pol} is the polar solvation energy, ΔE_{np} is the non-polar solvation energy, ΔG_{gas} is the total gas phase free energy, ΔG_{solv} is the total solvation free energy, ΔG_{bind} is the total binding energy and $\Delta\Delta G_{MM-PBSA}$ is calculated as the difference in binding energy between SETD7 and SAM and SETD7 and the corresponding SAM analogue. All the energies are shown in kcal/mol.

In section 5.3.3.2. and section 5.3.4. it was observed that the best SAM cofactor analogue for SETD7 was Pip-SAM, based on Pip-SAM having similar distances and angles that allow an efficient methyl transfer in native SAM. The binding affinity estimate seems to support this theory, as the calculated binding energy difference between SAM and Pip-SAM is extremely similar, only differing in 0.35 kcal/mol tighter binding for Pip-SAM. According to the MMPBSA calculation, MPip-SAM and Phe-SAM would bind with lower affinities, whereas TPhe-SAM would bind tighter with a difference of 4.90 kcal/mol. However, the binding mode for TPhe-SAM was not as similar to the one of SAM than with Pip-SAM. In conclusion, MD simulations indicate that Pip-SAM might be the ideal the ideal SAM analogue to resemble the activity of native SAM with SETD7.

5.4. Conclusions and future work

The conformational properties, binding modes and affinities of four SAM cofactor analogues (piperidine-SAM, methylpiperazine-SAM, thiophenyl-SAM and phenyl-SAM) have been investigated by means of MD simulations and binding free-energy (MM–PBSA) calculations. The SAM cofactor analogues were ranked according to their capacity to be utilised specifically by G9a or SETD7. The assumption adopted for this classification was the more the similarities to the SAM binding mode, the higher the compatibility. To assess this, distances, angles and binding affinities critical for the methyl transfer were computed for the different SAM analogue-enzyme pairs, and compared to the values obtained for SAM.

Distance and angle calculations presented Pip-SAM as the best SAM cofactor analogue to differentiate between the activity of G9a and SETD7 among the cofactors analysed. This conclusion was reached because the characteristic distances and angles that allow an efficient methyl transfer differed extensively between Pip-SAM and native SAM for G9a, but they were similar for SETD7. Besides, when observing the representative structures of the major clusters of the MD trajectories of G9a and SETD7 with Pip-SAM, both the angle and the distance between the methyl group and the catalytic tyrosine were not adequate for G9a. However, for SETD7 they resembled the binding mode of native SAM, with similar values of distance and angle to the ones observed in the experimental crystal structure of SETD7. Finally, a MM-PBSA calculation was performed, showing that Pip-SAM was the SAM cofactor analogue achieving a more similar binding energy with regards to the one obtained for native SAM with SETD7, only differing in a 0.35 kcal/mol tighter binding for Pip-SAM. Together, these arguments suggest that Pip-SAM enables selective methylation by SETD7 with regards to G9a. This result has been verified experimentally as mentioned in Chapter 4.

Future work should involve the design of further SAM cofactor analogues based on the current results of the simulations, the assessment of those SAM cofactor analogues through further MD simulations and application of the approach described. Another interesting point would be to perform further independent MD simulation runs of the same systems described, each of them with different initial velocities, to assess the reproducibility of the

results. This work shows that the study of distances, angles and binding affinities can help discerning the compatibility of different SAM analogues towards KMTs.

Our method has demonstrated that it is possible to design selective SAM cofactor analogues towards native KMTs when their structures are distinct enough between them. The next chapter will focus on those cases where the structures of different KMTs are so similar that it may not be possible to design selective cofactors, therefore requiring mutations in the KMT structure to generate a “hole” to accommodate the “bumped” cofactors.

Chapter 6

A bump-and-hole approach to identify G9a's targets

The focus of this work thus far has been on engineering SAM cofactor analogues to be selective for a subclass of wild type methyltransferases. However, some KMT structures are so similar that it may not be possible to design selective cofactors towards only one target KMT, as it is the case for G9a and GLP (see Chapter 2, Figure 2-3 for a phylogenetic tree of KMTs). In such cases, it may be desirable to pursue a classical bump-and-hole strategy, where the KMT of interest is mutated to accommodate a bumped SAM cofactor. Since the bump-and-hole problem involves simultaneous engineering of enzyme and cofactor, it is considerably more difficult than the problems addressed in Chapters 4 and 5. To simplify the analysis and provide a proof-of-principle, this chapter will utilise the proposed SAM cofactors with the bump on the C2 position to design mutations on the KMT of interest, using G9a as a case study.

The proposed methodology for identifying mutations to G9a to accommodate the C2-bumped SAM analogues involved a bioinformatic analysis of the binding pocket of G9a to identify mutant candidate pairs followed by comparative modelling with energy refinement to assess the structural changes of the proposed mutations and MD simulations to assess the

suitability of the SAM cofactor analogues. The analysis and computational assessment is then substantiated via *in vitro* assays of the mutant enzyme-cofactor pairs and their consequent analysis based on LC-MS/MS to decipher the KMT-substrate targets.

In summary, the objective of the work in this chapter is to develop a bump-and-hole-based assay for the identification of G9a-specific methylation. The mG9a should catalyse the transfer of a traceable donor group from the SAM analogue, preserving the biological function of the original mark with similar kinetics. Furthermore, the modified SAM should ideally bind specifically to mG9a and have lower affinity to other endogenous KMTs. Finally, mG9a should retain the same SET domain fold as native G9a so that the correct binding and orientation of the SAM cofactor relative to the histone substrate is achieved, which is important in facilitating the transfer of the methyl group.

6.1. Introduction

6.1.1. Introduction to G9a

G9a is a KMT responsible for histone H3 lysine 9 (H3K9) mono- and dimethylation. These histone methylation marks are then identified by the heterochromatin protein 1 (HP1), leading to transcriptional silencing [289]. One reason for selecting G9a as the KMT for the case study in this chapter is its crucial role in cancer biology. This enzyme is upregulated in many different types of cancer, showing implications in cancer initiation and progression. Specifically, higher expression levels of G9a are correlated with higher histone methylation, leading to the lack of expression of critical tumor suppressor genes and therefore to cancer. For example, G9a has been found to be overexpressed in aggressive lung cancer, brain cancer, multiple myeloma, aggressive ovarian carcinoma, esophageal squamous cell carcinoma, and hepatocellular carcinoma [290]. The apparent importance of G9a in cancer have led to the conclusion that this enzyme represents an exceptional target for forthcoming therapies, with some inhibitors already developed to block the epigenetic activity of G9a, such as BIX [291], UNC0638 [74], UNC0642 [292].

G9a also appears involved in protein complexes. In fact, a G9a-GLP heterodimer has been shown to be the predominant form of G9a *in vivo*, with the dimer appearing to be a functional H3K9 KMT [76]. It is important to highlight the fact that only the overexpression of G9a, and not of GLP, has been associated with cancer phenotypes, and G9a's activity has been proven to be more important than GLP's for the *in vivo* function of the complex [293]. In addition to G9a's heterodimer with GLP, a subset of each of the main H3K9 KMTs (G9a, GLP, SETDB1 and Suv39h1) form a multimeric complex [78]. Furthermore, when depleting G9a from cells, the rest of KMTs are destabilised at the protein level [78], indicating the dependence of the enzymes involved in the complex with respect to G9a and its fundamental importance in cancer regulation.

Understanding the structure of G9a and the binding mode of SAM in G9a's binding pocket is a necessary condition to apply a rational bump-and-hole approach as it is critical to assess which groups could be inserted into SAM to match G9a's mutant candidates. The

availability of crystallographic structures for G9a and its homologue GLP bound to SAH and to the histone H3 tail (e.g. G9a PDB: 2O8J, GLP PDB: 2RFI) was therefore another important reason why G9a was selected for this case study.

6.1.2. Non-histone targets of G9a and methods for non-histone target identification

As mentioned in Chapter 2, section 2.3.1. , G9a has been observed to methylate both histone and non-histone targets. Although the mechanisms for its non-histone methylation are not as characterised as the ones for histones, G9a controls important cellular functions through these modifications.

There are various existing methods aiming to identify methylation targets of KMTs. Among the substrate methylation detection methods, radiometric quantification with radiolabeled methyl groups, antibody-based detection pairing the KMT substrate and an anti-methyl-lysine antibody with donor and acceptor coloring agents and mass spectrometry are the most commonly used. These methods were analysed in detail in sections 3.1.1. , 3.1.2. and 3.1.3. respectively. Some of the identified G9a's non-histone targets, their function and the techniques used to identify them are:

- Tumor suppressor p53 at K373 residue: inactivation of p53. SiRNA knockdown of G9a followed by western blot analysis. Not MS confirmation [294].
- Automethylation of G9a at K239 position: facilitates the binding of the transcriptional repressor HP1 at the methylated site. This is probably critical for the recruitment of other factors. Studied via SDS gels, fluorography, LC-MS/MS identification [104].
- The chromatin remodeling factor Reptin at K67: represses gene expression. Quantitative Real-Time RT-PCR and ChIP assays. Confirmed by LC-MS/MS. [295].
- The chromatin remodeling factor Pontin: activates gene expression. Quantitative Real-Time RT-PCR and ChIP assays. Not confirmed by LC-MS/MS [296].
- The myogenic regulatory factor MyoD at K104: inactivates its regulation of gene expression. Western Blot analysis, ChIP assay, no MS confirmation [297].

- The transcription corepressor CDYL1 at K135: alters its chromodomain binding to H3K9me3 and regulates the activity of chromatin factors. Studied via SDS gels, MS confirmed [80].

Further confirmed G9a non-histone targets were shown in Chapter 2, Figure 2-9. As can be deduced from the previous lines, even among the most "confirmed" non-histone targets of G9a, the findings are sometimes ambiguous and not accurately quantified as they lack MS confirmation. There have been previous efforts trying to overcome the limitations of the previously mentioned studies. For instance, the strategy of Luo and his collaborators Islam and Wang focused on using clickable SAM analogue cofactors built by modifying the methyl donor group of SAM by various alkyl clickable groups; this clickable alkyne could be transferred by G9a to its targets. The outcome of the transfer could then be analysed using MALDI-MS [140,166,169,212]. Further details about this strategy are described in chapter 3, section 3.3.2. . Besides, Minkui Luo's lab extended this approach to identify >500 targets of G9a or GLP, generating SAM analogues by replacing SAM's methyl group with bulky, chemically active functionalities, and matching mG9a/GLP enzymes, were MALDI-MS was used as the primary assay [298].

6.1.3. MD simulations for the bump-and-hole approach

The bump-and-hole technique consists on introducing a mutation to the protein active site that makes an enzyme able to selectively bind to a small molecule [167]. To our knowledge, molecular dynamics simulations have not been implemented in the application of a bump-and-hole approach. The closest was the application of docking programs such as Glide that were used to design BET bromodomain chemical probes [299]. However, as mentioned previously in this chapter, the docking results are merely approximate and highly dependent on the system. This project merged the use of biomolecular simulations within the bump-and-hole design, making the decisions on the design more rational, saving experimental resources and time.

6.1.4. Technical challenges and opportunities

The click chemistry-based approach for the identification of G9a methylation targets that was developed in Luo's lab [140,166,169,212] presents some limitations when extended to a cellular setting [181]. Specifically, the SAM analogues used by this method show basal activity for other endogenous KMTs. In addition, the modified alkyl group is not structurally comparable to the methyl group, and thus the propagation of methylation function (e.g. formation of heterochromatin via HP1-binding) is not conserved.

Thus, there exists a need for a cell-based, biologically consistent assay. To overcome the limitations that are present in Luo's work, we propose the substitution of the alkyl group in Luo's work by a heavy methyl donor (CD_3) that is biologically compatible. To ensure specificity towards G9a, we designed SAM cofactor analogues with a "bump" to match a mutation in G9a's binding pocket. This strategy was combined with *in vitro* assays and LC-MS/MS analysis to guide the identification of novel methylation sites.

As mentioned in section 6.1.3. , MD simulations have not been applied to the bump-and-hole approach. Currently, the closest to model the enzyme-cofactor pairs prior to experimentation are docking/scoring methods that do not reflect accurately the real binding event [299]. This paper introduces the first merging of biomolecular simulations and bump-and-hole design.

The research performed to achieve the goals of this project presented some challenges, for example, in the mutagenesis of G9a to generate specific mG9a-SAM cofactor analogue pairs. For this strategy to work, it is necessary to identify mutations in G9a that do not affect enzymatic activity, the substrate specificity or the folding of the enzyme and make the enzyme bind and utilise a bumped SAM cofactor analogue showing reduced activity for other endogenous KMTs. Complementarily, there are some challenges associated with the cofactor design: it is necessary to synthesise a stable cofactor with a bump that is selective to mG9a compared to other native KMTs, such as native G9a. This challenge becomes even tougher since not all KMTs present the exact same arrangement of residues in the SET domain. Section 6.2. focuses in describing the methodology followed to overcome these challenges.

6.2. Methodologies

6.2.1. Rosetta structure prediction

Rosetta is a software suite created by the Baker Lab that can perform both protein structure prediction and functional design [217]. It has been successfully applied to predict the structure of proteins, with or without the help of experimental data, protein-protein and protein-ligand docking, protein design of original proteins and protein redesign to obtain novel functions [300]. The outputs of this suite allow the fast verification of hypotheses that would otherwise be very expensive or time-consuming to test experimentally via X-ray crystallography, Nuclear Magnetic Resonance (NMR), Electron Paramagnetic Resonance (EPR) or electron microscopy.

Rosetta has consistently been one of the best programs for protein structure prediction according to the Critical Assessment of protein Structure Prediction (CASP) [301–304]. A summary of some of the most important applications within Rosetta is shown in Table 6-1.

Table 6-1. Main applications within Rosetta [300,305].

	Rosetta application	**Brief description**
Structure prediction	AbinitioRelax	Predicts the tertiary structure of a protein from its amino acid sequence
	Comparative_modeling	Predicts the structure of a novel protein using one or more structures as templates
	loop_modeling	Predicts the conformation of a set of loops within a certain protein
Docking	protein_docking	Given 2 proteins with known structure, predicts their docking
	ligand_docking	Given a protein and a small molecule, predicts their docking
Protein design	enzdes	Designs the protein active site to perform a certain catalysis
	FixedBBProteinDesign	Given a fixed protein backbone, redesigns the amino acids

The application that has been used for this project is Rosetta comparative modeling. This application builds structural models of proteins from their sequence by using one or more known structures of homologous proteins as templates. The election of comparative modelling and not *ab initio* approaches was based on the hypothesis that the new protein should keep a similar fold to the endogenous one, as the number of changing residues is reduced, and the higher accuracy of comparative modelling when templates are available when templates are available with a sequence identity above 30%[306,307]. The approach followed by Rosetta comparative modelling is shown in the upcoming points.

6.2.1.1. Sequence alignment

Alignments of the query sequence are generated with one or more sequences of homologous proteins with available structures in the PDB. Currently, this alignment relies on the use of external programs such as BLAST or HHpred [308].

6.2.1.2. Model generation

In the first place, an incomplete model is generated based on the template structure by copying the coordinates of the query sequence over the aligned regions. The quality of the alignment determines the way Rosetta samples. If there exists a high degree of sequence homology between the query and the template (>50% identity), the protein backbone is only rebuilt in regions involving insertions and deletions in the sequence alignment [301]. More specifically, Rosetta starts with the template structure and builds in missing loops using fragment insertion [309] or randomization of ϕ and ψ angles followed by a the loop closure algorithms such as kinematic closure or cyclic coordinate descent [310].

6.2.1.3. Refinement of the structure using the Rosetta full-atom energy function.

Rosetta contains two types of energy functions: a low-resolution function called centroid energy function, used to search the conformational space of the query protein in the *ab initio* approach, and a high-resolution one called all atom energy function [311] used in both *ab initio*

and comparative modelling approaches to optimise the structure, where the side chains are represented with atomic detail. The all atom refinement slightly "relaxes" the ϕ and ψ angles of the backbone, but maintains the global conformation of the protein. Regarding the side chains, they are adjusted using a simulated annealing Metropolis Monte Carlo search of the rotamer space. Ultimately, gradient minimization is applied to all the torsional degrees of freedom: ϕ, ψ, ω and χ [312].

The Rosetta all atom energy function contains a solvation approximation, a knowledge-based electrostatics term, an orientation-dependent hydrogen bonding potential, a knowledge-based conformation-dependent internal free energy term and a 6-12 Lennard-Jones potential for van der Waals forces [312], which allows a more detailed prediction when compared to the centroid function. The hydrogen bond potential is of special interest, as it is able to reproduce the angular dependence of high-level *ab initio* quantum mechanical energies for hydrogen-bonded side chain analogs better than traditional physics-based force fields that lack explicit hydrogen bond terms [313,314].

6.2.1.4. Clustering and model selection

Once several models have been generated according to the previous steps, it is necessary to assess which is the most accurate model among all generated. A possibility to discern this is the use of a clustering program such as the Rosetta Cluster application, as described by Gray et al. [315]. Using this program, the decoys are clustered based on their pair-wise root-mean-squared deviation (RMSD) using a hierarchical clustering algorithm. Structures that fall within a clustering threshold are added to the same cluster. The clusters with more members are selected as the ones containing final predictions, ranked according to the cluster sizes. The lowest-scoring decoy within the most populated cluster is selected as the representative model of the simulation.

6.2.2. Proposed KMT design strategy

The objective of this chapter is to design a mG9a that conserves its function and selectively uses a bumped SAM analogue as a methylation cofactor. The designed mG9a-SAM analogue complex should keep the folding of G9a; the mG9a should be active and able to

utilise the SAM analogue to methylate its target substrates; besides, the SAM analogue should be specific towards G9a and not towards other KMTs. Each of these goals and their consequent research guidelines are shown in Table 6-2.

Table 6-2. Design objectives of the mG9a-SAM analogue complex and their influence on the analysis

Objective	Reasoning of this objective	Hypothesis on research
Preserve folding	The mG9a should keep the same SET domain fold than the endogenous G9a, as this structure enables the recognition of both SAM analogue and the substrates.	To avoid the unfolding of the binding site, mutations should be performed on residues that are not evolutionarily conserved.
Specificity	The mG9a should preferentially bind the SAM analogue and vice versa to ensure the results of the MS analysis correspond to G9a's activity.	Analyse which residues in the SAM binding pocket of G9a can be modified to increase the specificity towards "bumped" SAM analogues, creating a "hole" to accommodate the "bump".
Activity	The mG9a should be able to transfer the chemical mark from the SAM analogue to G9a's substrates (such as H3K9) with the appropriate kinetics and allowing the preservation of its biological function.	Unlike other approaches using an alkyl group as the transfer group [137], the current approach utilises a heavy methyl donor (CD_3) as a transfer group. This group has been successful in the preservation of the biological function due to its resemblance to the methyl group, while allowing its detection via MS. To implement this approach, it is necessary to modify SAM at points other than the leaving group.

To achieve the design objectives of the mG9a-SAM analogue complex, the following design strategy is proposed: (1) the identification of G9a residues for mutation based on the analysis of the structures and sequences of G9a and other related KMTs; (2) predicting the fold of mG9a candidates with comparative modelling and MD simulation software to assess their match with proposed SAM analogues; and (3) a MS-based *in vitro* validation of the ability of the proposed mG9a to methylate nucleosomes using the SAM cofactor analogue pairs.

6.2.2.1. Identification of candidate G9a residues for mutation

The design hypotheses for our application of the bump-and-hole approach to G9a are: (1) conserving the G9a-SAM interactions (e.g. hydrogen bonds) that are critical to achieve the orientation of SAM towards the histone substrate; and (2) preserving the overall fold of G9a

to conserve appropriate binding and orientation of SAM for methylation to occur, and in theory preserving other important protein-protein interactions relating to G9a function.

PyMOL was the biomolecular visualization program selected to analyse the hydrogen bonds between SAH and G9a in its binding pocket, the sub-domains of G9a, discern possible SAM modification points and their corresponding G9a residues to mutate, visualise clashes of proposed SAM analogues with G9a and compare the crystallographic structure of G9a with those of other KMTs. PyMOL presents useful features such as the insertion and deletion of atoms or moieties in the structure that make it ideal for its use in this project. Information in the form of hydrogen bonding and atomistic distances between SAH and G9a was gathered from an available crystallographic structure of G9a with the PDB identifier 2O8J [65]. The hydrogen bonds were identified following the criteria of Xu et al. [316], where the maximum accepted distance between donor and acceptor is 3.9 Å and the minimum angle donor–hydrogen–acceptor is 90.0°. A representation of the different types of atoms involved in a hydrogen bond is shown in Figure 6-1.

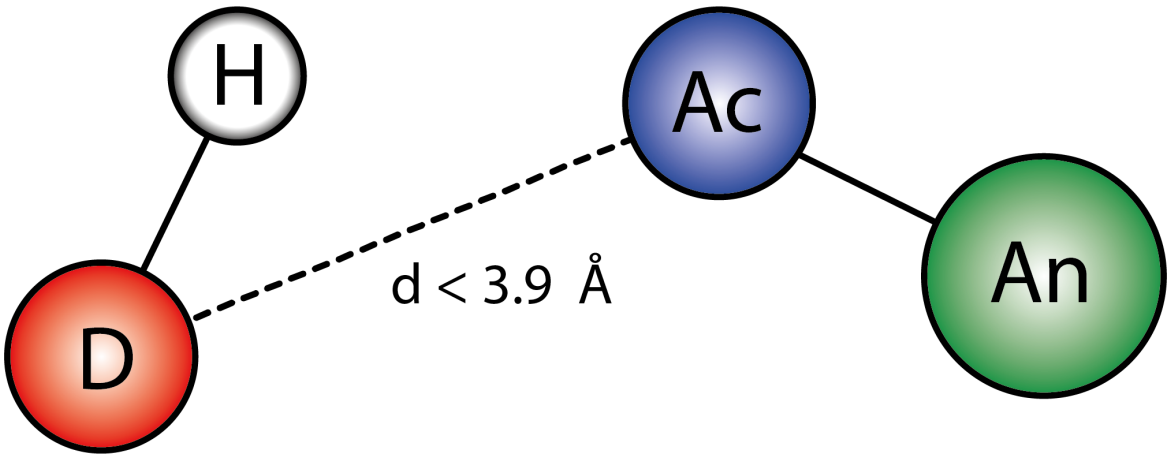


Figure 6-1. Description of a hydrogen bond. D: donor, H: hydrogen, Ac: acceptor, An: antecedent of the acceptor, d: distance donor-acceptor, maximum 3.9 Å.

A multiple sequence alignment of G9a with its closest related KMTs, named GLP, SETDB1, SETDB2, SETMAR, SUV39H1 and SUV39H2 (as described in the phylogenetic tree shown in Chapter 2, Figure 2-3) was performed using Clustal Omega, a multiple sequence alignment program. The alignment in Clustal Omega is driven by sites of evolutionary conservation, where conserved residues imply a conservation of function [317]. The resulting alignment was inputted into pLogo, a program that provides a representation

of a PWM (Position Weight Matrix) where each possible residue has a value at each position based on the statistical significance of the frequency of that residue in the input alignment [318]. In the graphs generated by pLogo, the residues that are more evolutionary conserved appear bigger and closer to the X axis. These graphs enabled to visualise the conservation of G9a's possible mutation residues in the SAM binding pocket and therefore assess if their mutation would potentially interfere with G9a's function or fold, and enabled the assessment of the conservation of the residues involved in G9a'. hydrophobic channel.

The multiple sequence alignment generated with Clustal Omega was inputted into a MATLAB algorithm that compares each residue of the KMTs considered in the multiple sequence alignment to the one of G9a at the aligned position and assigns a conservation value to this residue according to BLOSUM62 (BLOcks Substitution Matrix) (see Appendices Figure A-7 for information on the code used). BLOSUM62 is a scoring matrix used for the sequence alignment of proteins [319] based on local alignments having a certain value for each pair of residues in the alignment, making the BLOSUM62 matrix accurate even for evolutionary divergent protein sequences (see Appendices Figure A-8 for a representation of the BLOSUM62 matrix). The values obtained were made relative to the highest value (to obtain figures between 0 and 1) using Microsoft Excel and used to visualise the conservation of different residues in the structure of G9a using PyMOL by changing the B-value of each atom (temperature factor) for its conservation value in the PDB file of G9a's structure. Overall, potential G9a sites of mutation were proposed combining the results obtained from the conservation analysis and the structural analysis in PyMOL. This methodology is summarised in Figure 6-2.

6.2.2.2. Computational screening of mG9a candidates

Once possible mutation points were identified, the effect of these mutations on G9a's structure was evaluated using the comparative modeling functionality of Rosetta 3.5 [301]. This program requires some input files: the FASTA file of the query sequence, the PDB file of the template structure, a grishin alignment file between the query and template sequences and fragment files to model the loops. To generate the sequence alignment between the query sequence and the template, the FASTA file of the query sequence was inputted into HHpred [308]. The generated alignment file was then modified so only the alignment with the

most closely related protein found (GLP, PDB: 3hna) was kept. The alignment was then converted into the grishin format using the convert_aln script included in Rosetta 3.5. Then, the alignment was fixed to match the PDB file template using the program fix_alignment_to_match_pdb also included in Rosetta 3.5. Besides, the FASTA file of the query sequence was inputted into the Robetta Fragment Server [320] to generate the 3mer and 9mer fragment libraries for the modeling of the loops.

The models of mG9a were generated using the minirosetta program included in Rosetta 3.5 using the previously generated fragment files, fixed grishin alignment file, PDB file of the template and FASTA file of the query sequence as inputs. For each mutant candidate, 10000 models (or decoys) were generated. The input file is shown in Appendices, Figure A-9. The output of the calculation was a silent file containing information about the best models. The computations were run on the Linux Cluster at the Department of Chemical Engineering of Imperial College London.

The generated decoys were clustered using the cluster program included in Rosetta 3.5. The previously generated silent file was used as input with the structures read in full atom mode. The calculation was performed specifying a cluster radius of 1.5 Å. The energies of representative decoys from each cluster were estimated, and the models were ranked in terms of their energy within each cluster. The input file is shown in Appendices, Figure A-10. The model having the lowest energy within the most populated cluster was chosen as the solution. Finally, the PDB file of the solution structure was extracted from the silent file using the extract_pdbs program within Rosetta 3.5, ready to be further examined in PyMOL. The methodology followed is summarised in Figure 6-2.

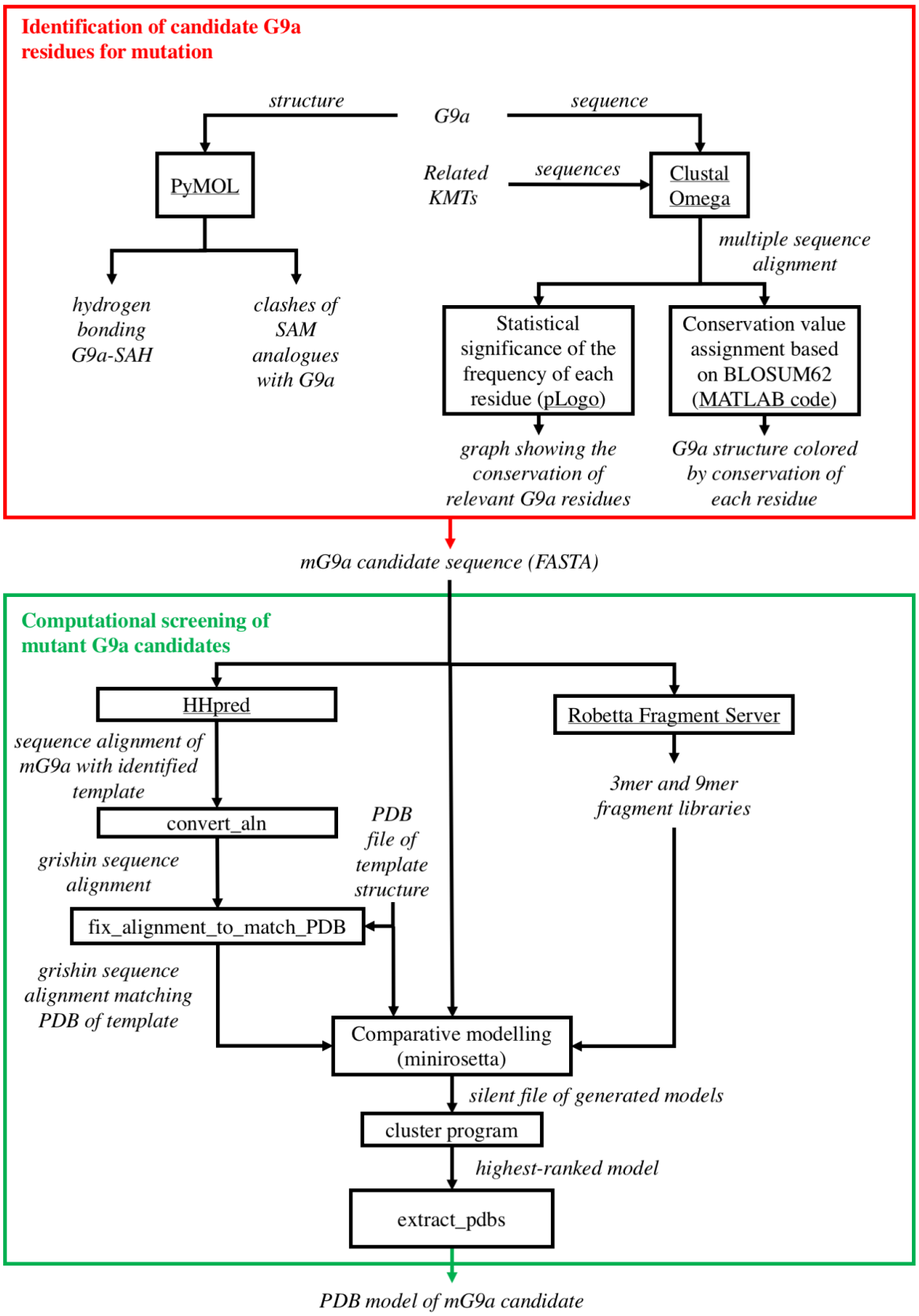


Figure 6-2. Methodology followed for the identification of candidate G9a residues for mutation (red) and computational screening of mG9a candidates (green). All the programs that are external to the Rosetta 3.5 suite appear underlined.

6.2.3. Experimental materials and methods

To obtain the mG9a, site-directed mutagenesis was performed by Till Bartke's group at Imperial College London on the plasmid number #25503 on Addgene of endogenous G9a provided by the Structural Genomics Consortium (SGC). The site-directed mutagenesis was performed using a QuikChange® Site-Directed Mutagenesis Kit (Startagene) according to the instructions of the kit, consisting of denaturing the original plasmids and annealing the oligonucleotide primers containing the desired mutation. Then, the PCR was carried out extending and incorporating the mutagenic forward and reverse primers in nicked circular strands with Pfu Turbo DNA polymerase. Dpn I restriction enzyme was then added to digest the methylated, nonmutated parental DNA template. The nicked circular dsDNA was then transformed into XL1-Blue supercompetent cells to repair the nicks in the mutated plasmid and colonies were checked for mutagenesis. Plasmids were then extracted from the XL1s using Plasmid Plus Maxi kit (QIAGEN) according to the instructions and the sequence mutations were validated via DNA sequencing.

The mG9a-containing plasmids were then transfected by electroporation into BL21 competent *E. coli* and selected in the presence of 50 μg/mL of kanamycin in collaboration with Fabio Fisher from Jake Baum's lab, Imperial College London. The mG9a was then expressed and purified as specified for G9a in Chapter 4, sections 4.2.2.1. and 4.2.2.2. except for substituting the Hanahan's Broth by Select APS Super Broth (BD). Protein gels were run as specified in Chapter 4, section 4.2.2.3. The purity and validation of mG9a was also assessed through in-solution digestion and MS identification as described in section 4.2.2.4. *In vitro* methyltransferase assays were then performed as mentioned in Chapter 4, section 4.2.2.5. The MS workflow for histones was the same as the one mentioned in section 4.2.2.6. , and the MS operation the same as described over section 4.2.2.7.

6.3. Results and discussion

6.3.1. G9a structural characterization for a bump-and-hole design

A detailed structural analysis of G9a's SAM binding pocket and its structural features was performed to guide the bump-and-hole design. The bump-and-hole approach requires simultaneous engineering of both the SAM cofactor and mutations to the KMT, which is considerably harder than the problems considered in Chapters 4 and 5. However, this work will show that appending a bump to the C2 position of SAM is structurally ideal for engineering the complementary mutations to the G9a post-SET domain, and so the SAM analogues designed in Chapter 4 can also be used to test mG9a.

6.3.1.1. Bonding interactions of G9a with SAM

G9a catalyses the transfer of a methyl group from the cofactor SAM to a substrate lysine, such as the histone target H3K9, using its highly-conserved SET domain. Due to the lack of crystallographic structures of G9a bound to the H3K9 substrate, and based on the high degree of sequence similarity between G9a and GLP, a G9a-SAH-H3K9 structure was created using PyMOL by aligning (RMSD = 0.508 Å) the crystallographic structures of G9a (PDB: 2O8J) and GLP (PDB: 2RFI). The proximity of histone H3K9 and SAH in G9a's binding pocket is shown in Figure 6-3.

Hydrogen bonding is a key contributor in the recognition and activation of specific substrates of enzymatic systems [321]. Therefore, it is critical to assess and maintain the native interactions between G9a and SAH in order to preserve correct cofactor orientation relative to the histone substrate. A summary of the existing hydrogen bonding between SAH and G9a is shown in Table 6-3. As the objective of this bump-and-hole approach is to create an engineered mG9a-SAM analogue pair with high specificity, it is important to keep as many highly-conserved and functionally important interactions as possible. The results are presented in Table 6-3 and the 3D visualization of these interactions is shown in Figure 6-4.

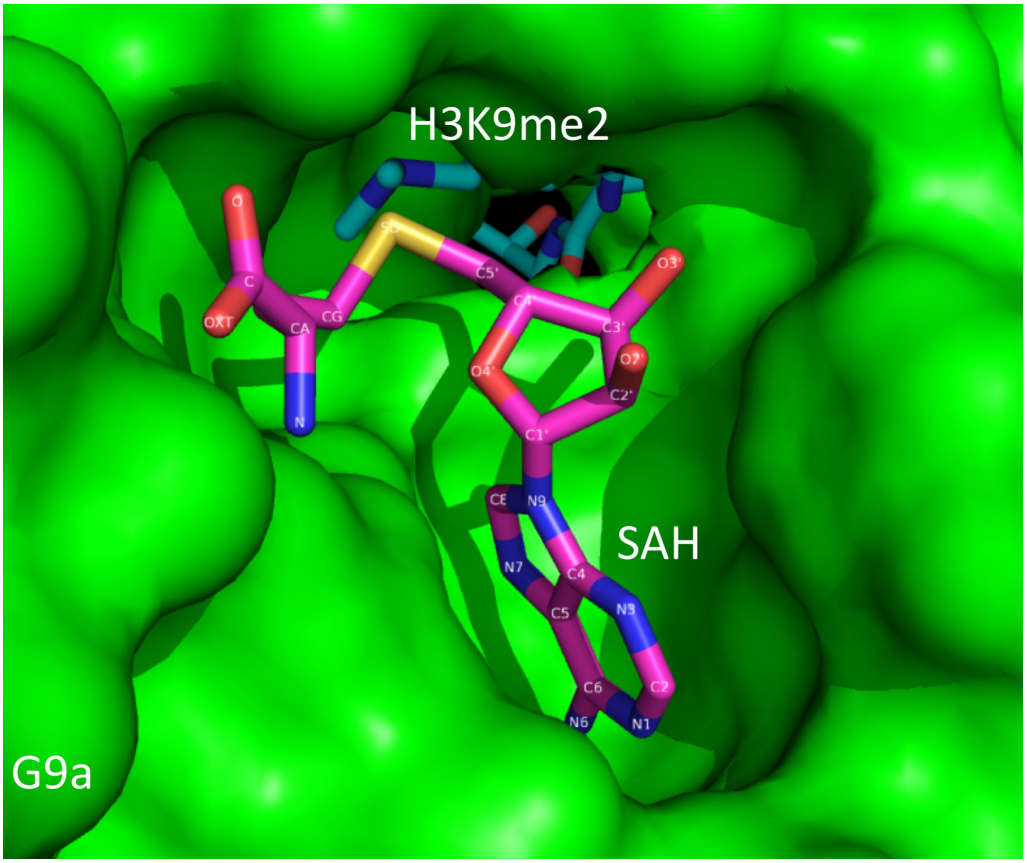


Figure 6-3. Detail of G9a's SET domain (green surface) bound to SAH (magenta sticks), both from PDB: 2O8J, and H3K9me2 (cyan sticks) from PDB: 2RFI.

Table 6-3. Identified hydrogen bonds between S-Adenosyl-homocysteine and G9a. BB: backbone. SC: side chain.

	SAH atom name	G9a Residue			Domain name	Distance Donor-Acceptor
		Name	Number	Interacting atom		
Homocysteine	N	MET	1048	O-BB	SET	2.7
		TRP	1050	N-BB	SET	2.8
		ASN	1112	N-SC	SET	2.8
	OXT	TRP	1050	N-BB	SET	2.8
	O	TYR	1085	O-SC	I-SET	2.8
Ribose	O3'	SER	1084	O-SC	I-SET	3.8
	O4'	ASN	1112	N-SC	SET	3.7
Adenine	N1	GLN	1169	N-BB	Post-SET	3.0
	N6	HIS	1113	O-BB	SET	2.9
	N7	HIS	1113	N-BB	SET	3.0

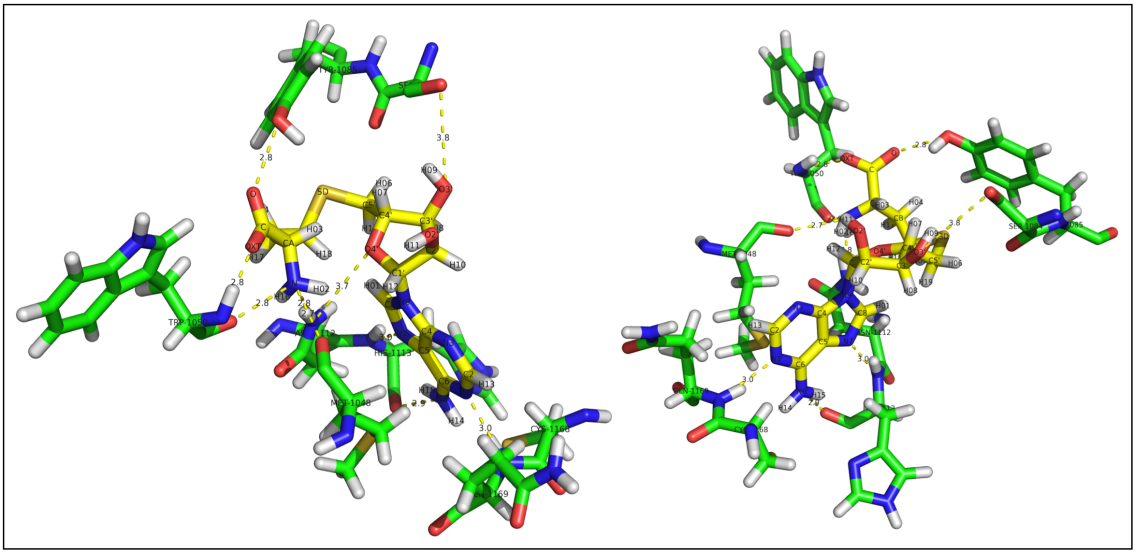


Figure 6-4. Hydrogen bonding interactions between SAH (yellow sticks) and G9a (green sticks) from two viewpoints. Generated with PyMOL using PDB: 2O8J.

It is also important to highlight that G9a is a multi-domain protein composed of several domains. Together, these domains contact and bring together the SAM cofactor and protein substrate to facilitate methylation of histone lysines and other targets. The G9a domains surrounding SAH and the possible atoms to incorporate groups are shown in Figure 6-5.

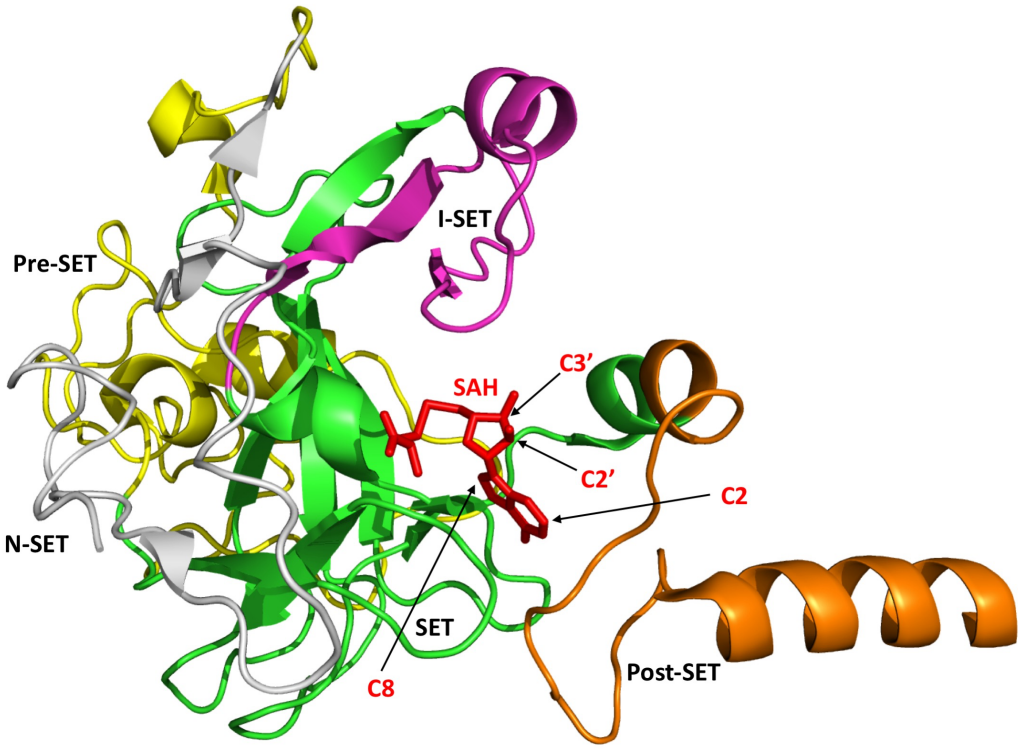


Figure 6-5. G9a's sub-domains within the SET domain shown with SAH (red sticks) and the possible atoms to insert groups at.

6.3.1.2. Suitability of SAM design atoms and complementary G9a mutations

In the process of identifying candidate G9a residues for mutation, it is necessary to avoid modifying the catalytic residues of G9a. The key G9a residues for histone recognition and catalysis are: (1) Y1067, which creates H-bonds with the methyl accepting nitrogen of the histone lysine, preventing the trimethylation of H3K9; (2) F1152, which allows the dimethylation of H3K9; and (3) Y1154, whose hydroxyl group participates in the catalysis [65]. Combining these important residues with those identified in Table 6-3 provides the basis for determining which SAM atom would be ideally suited to contain a synthetic bump. From Table 6-3, the N6 position is an unsuitable substitution point as it holds two key hydrogen bonding interactions with G9a, which contrasts with previous work (Lin et al. 2001)[167]. Recent work by the Zhou laboratory [177] inserted alkyl groups in substitution of the hydrogen atom bound to the atoms O2' and O3', with the latter being identified to be important in the binding of SAM to G9a (Table 6-3).

Thus, the remaining SAM atoms we identified as candidates for appending a synthetic bump to are C2', C3', C2 and C8. Therefore, the rest of the analysis focused on how modifications at the C2, C8, C2' and C3' positions of SAH would interact with the structure and potentially impact the function of G9a. The results, classified according to the corresponding SAH atom, are shown in Table 6-4.

Table 6-4. Possible SAH atoms to modify and analysis of the closest G9a residues to mutate.

	C2		**C8**		**C2'**	**C3'**		
Closest residues on G9a relative to SAH position	THR-1167	GLN-1169	ASN-1112	TYR-1154	PHE-1166	SER-1084	PHE-1158	PHE-1166
Distance to atom (Å)	3.5	3.8	2.9	3.5	3.7	4.9	4.9	3.8
Folding	-	-	β	-	-	-	α	-
SAH closer to G9a backbone (BB) or side chain (SC)?	BB	SC	SC	SC	SC	SC	SC	SC

According to Table 6-4, the C3' position seems to be a less favorable design candidate than C2', due to its proximity to residues located at the edge of the H3K9 hydrophobic channel of G9a. In principle, C2' would be a better candidate than C3' but is still close to the channel entrance. Additionally, C8 modifications would require G9a modification at N1112, which is part of the pseudoknot motif. This residue is involved in a β-sheet structure, and presents a highly-conserved hydrogen bond with both SAH homocysteine nitrogen and ribose O4'. Furthermore, the location of N1112 in the inner hydrophobic part of the protein might affect the correct folding of G9a. The C8 position is also in the proximity of Y1154, whose hydroxyl group participates in the catalysis, therefore a mutation at this position should be avoided. On the other hand, a C2 substitution would orientate the "bump" away from the binding pocket core. The G9a residues in the proximity of the SAH C2 position are the backbone of T1167 and the side chain of Q1169; the following section will focus on analyzing if those residues are safe to be mutated without disrupting G9a's SET domain.

6.3.1.3. Assessment of G9a mutations to fit C2-modified SAM analogues

A pLogo [318] evolutionary analysis was performed to analyse the conservation of the G9a residues that would interact with synthetic bumps on the SAM atom candidates considered in the previous section. The logic behind this design strategy is that avoiding the mutation of residues that are part of any highly-conserved motifs present in the KMT family of proteins should avoid any unexpected unfolding or significant structural change of the SET domain, and thus SAM binding and orientation for catalysis would be conserved. The KMTs that were considered in the analysis were G9a, SETDB1, SETDB2, SETMAR, SUV39H1, SUV39H2 and GLP due to their high evolutionary resemblance (please refer to Chapter 2, Figure 2-3). The pLogo webserver took as an input a multiple sequence alignment performed using Clustal Omega [317] prior the analysis. A summary of the results obtained following this approach is shown in Figure 6-6.

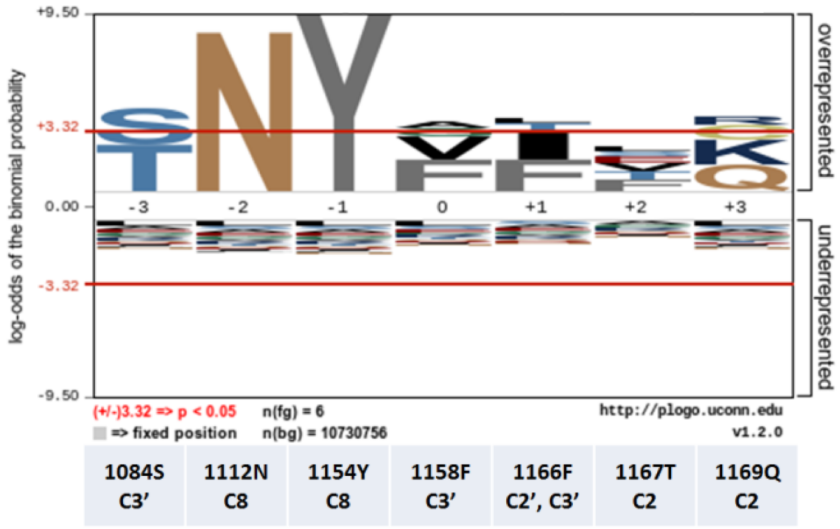


Figure 6-6. Statistical significance of the frequency of selected G9a residues located closer to each SAM atom candidate using pLogo. N(fg) indicates the number of aligned foreground sequences and n(bg) the number of aligned background sequences used to generate the image, respectively. The red horizontal bars on the pLogo correspond to P = 0.05.

According to the pLogo results, the most conserved residues for point mutations among those considered are, in decreasing order of conservation: Y1154, N1112, S1084, F1158, F1166, Q1169, and T1167. Therefore, the most-conserved G9a residues are found around the following SAM atoms: C8>C3'>C2'>C2. Thus, the worst point to insert a group in SAM appears to be C8 position, as the residues in its proximity are highly conserved. On the other hand, it can be concluded that the best site to insert a "bump" is the C2 atom of SAM, since its closest G9a residues are T1167 and Q1169 and these residues these residues show the most reduced conservation values among the atoms analysed. Therefore, we hypothesise that only small conformational changes will occur in G9a when these residues are mutated.

To present a visual representation of the conservation of G9a's SET domain residues, the B-value of each atom (temperature factor) was substituted in G9a's PDB file by the conservation values computed according to the BLOSUM62 matrix using our Matlab algorithm and the resulting structure was visualised in Pymol. As the B-value differs per atom, not per residue, every atom B-value was modified to the BLOSUM62 values of its corresponding residue. Figure 6-7 shows the results of applying this approach on G9a's structure (PDB: 2O8J). The residues located at the bottom of the binding pocket of SAM appear in red, indicating their higher conservation (e.g. residues surrounding SAM's N6 position). However, the T1167 and Q1169-containing post-SET domain appears in blue,

indicating a low conservation. This is a good indicator that the structure of G9a would not unfold when mutating the T1167 or Q1169 residues.

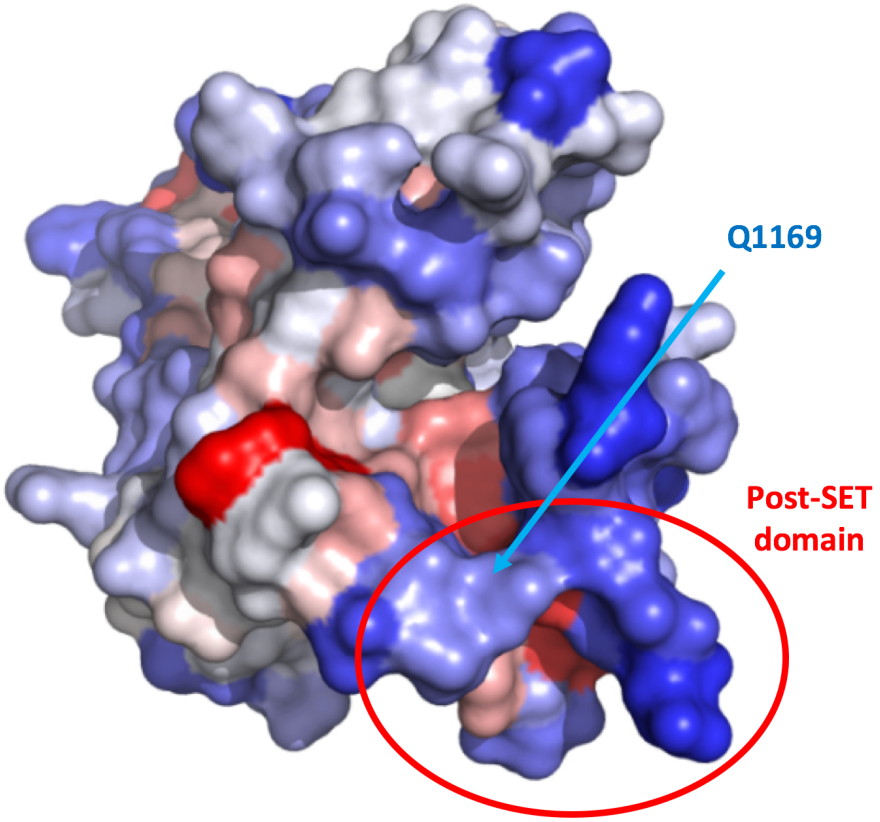


Figure 6-7. G9a's SAM binding pocket and hydrophobic channel residues according to their conservation state. Red: high conservation; white: average conservation; blue: low conservation. G9a's Q1169 residue and post-SET domain are indicated.

In summary, the combined computational analysis of the sequence, structure and evolutionary conservation of G9a's residues sets the C2 position of SAM as the preferential position to insert a "bump", with the corresponding "hole" generated by mutating the T1167 or Q1169 residues of G9a to smaller amino acids.

6.3.1.4. Candidate groups to incorporate on the C2 atom of SAM

To gain a first approximation of which moiety size would clash with the endogenous G9a structure but could fit into the "hole" generated by mutating T1167 or Q1169 side chains, methyl, ethyl, phenyl and benzyl groups were appended to the C2 positions of SAM based on G9a's structure (PDB: 2O8J) using PyMOL. The resulting clashes with the structure of G9a were visualised and are shown in Figure 6-8.

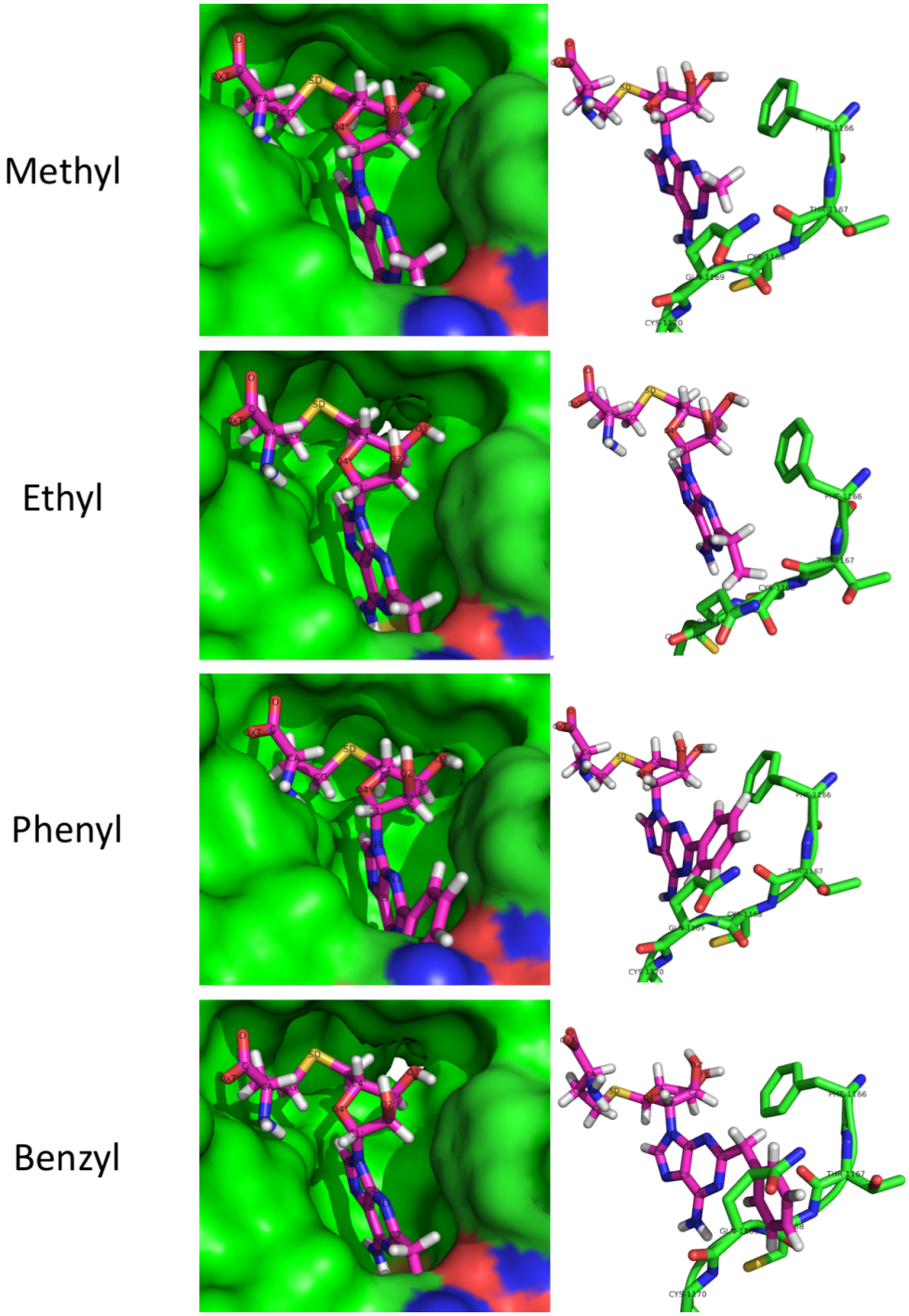


Figure 6-8. Insertion of groups at SAM's C2 position (magenta sticks) to analyse clashes with G9a's residues (green). Left: surface view. Right: sticks view.

Figure 6-8 shows that small groups such as methyl or ethyl group might not be sufficient to produce an adequate clash with the native structure of G9a, as the clashes visualised could be easily accommodated by slight changes in the protein structure. The insertion of a benzyl group seems to clash with the backbone of G9a's post-SET domain, making more difficult the generation of a suitable "hole". On the other hand, the insertion of a phenyl group

provides a rigid structure that clashes with the side chain of G9a's Q1169 residue, enabling the bump-and-hole application when mutating this residue to a flexible and small residue such as glycine or alanine. Therefore, the proposed moiety to be attached to the C2 position of SAM is a phenyl group, with the corresponding "hole" generated by mutating the Q1169 residue of G9a into a smaller amino acid.

6.3.1.5. Structural integrity of mutations to the post-SET domain: the ZnCys motif.

The proposed synthetic bump on the C2 position of SAM was found to contact Q1169 of G9a in its post-SET domain, which is not highly-conserved. It is important, however, to consider whether such a mutation still might lead to the unfolding of the SET domain and loss of function. This section provides evidence that the ZnCys motif of G9a should maintain the higher-order structure of the SET domain for the mutations considered in this work. G9a's ZnCys motif gathers 4 cysteine residues around a central zinc (Zn) ion. The G9a cysteines involved are C1115, located in the pseudoknot motif of the SET domain, and three cysteines located in the post-SET domain: C1168, C1170 and C1175. The structure of this ZnCys motif is shown in Figure 6-9. The coordination of the Zn ion with the cysteines of the post-SET domain makes the loop structure rigid (Figure 6-9, a), whereas the bond with C1115 makes this loop anchored to the pseudoknot motif in the SET domain (Figure 6-9, b). Together, these interactions "enclose" the binding pocket of G9a, making the binding pocket SAM-specific, theoretically disabling bulkier cofactors to bind (e.g. engineered SAM "bumped" cofactors at C2 position). Figure 6-10 presents an analysis of the presence of ZnCys motif in lysine and arginine methyltransferases.

It should be noted that the ZnCys motif does not appear in some SET domain-containing KMTs. Interestingly, it was found that KMTs SETD7 and SETD8 lack the presence of a ZnCys motif in their post-SET domains. This is the structural reason for why a natural "hole" was found in the SET domain of SETD7 in Chapter 4 that can accommodate a bump on the C2 position of SAM. It is obvious, therefore, that while the bump-and-hole approach proposed in this chapter would enable the differentiation of G9a activity when compared to all the ZnCys motif-containing KMTs, it would also be present in a background consisting of SETD7 and SETD8 methylation.

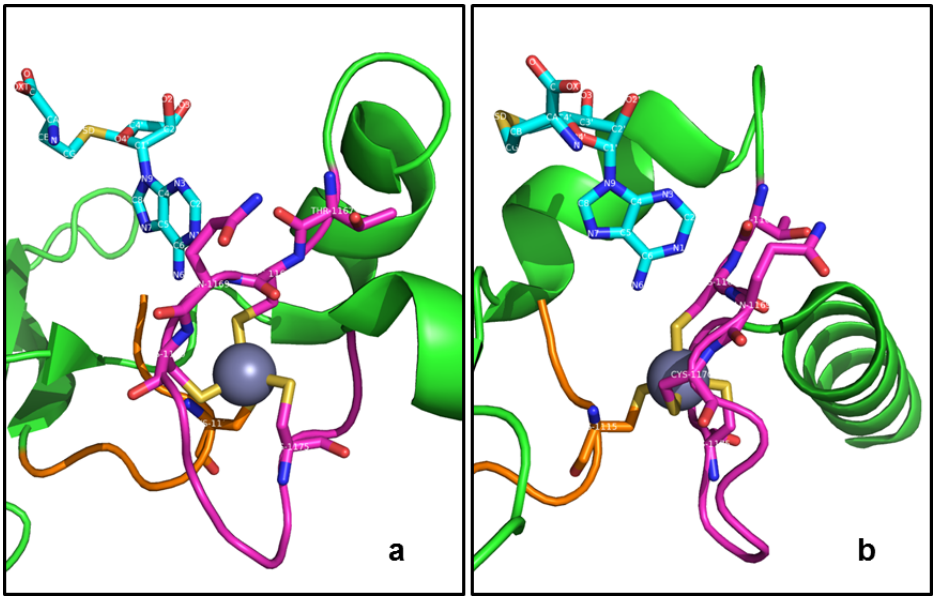


Figure 6-9. ZnCys motif of G9a. a) Front view. B) Side view. SAH is represented as blue sticks, Zn^{2+} as a grey sphere, G9a's pseudoknot C1115 as orange sticks and G9a's post-SET domain C1168, C1170 and C1175 as magenta sticks.

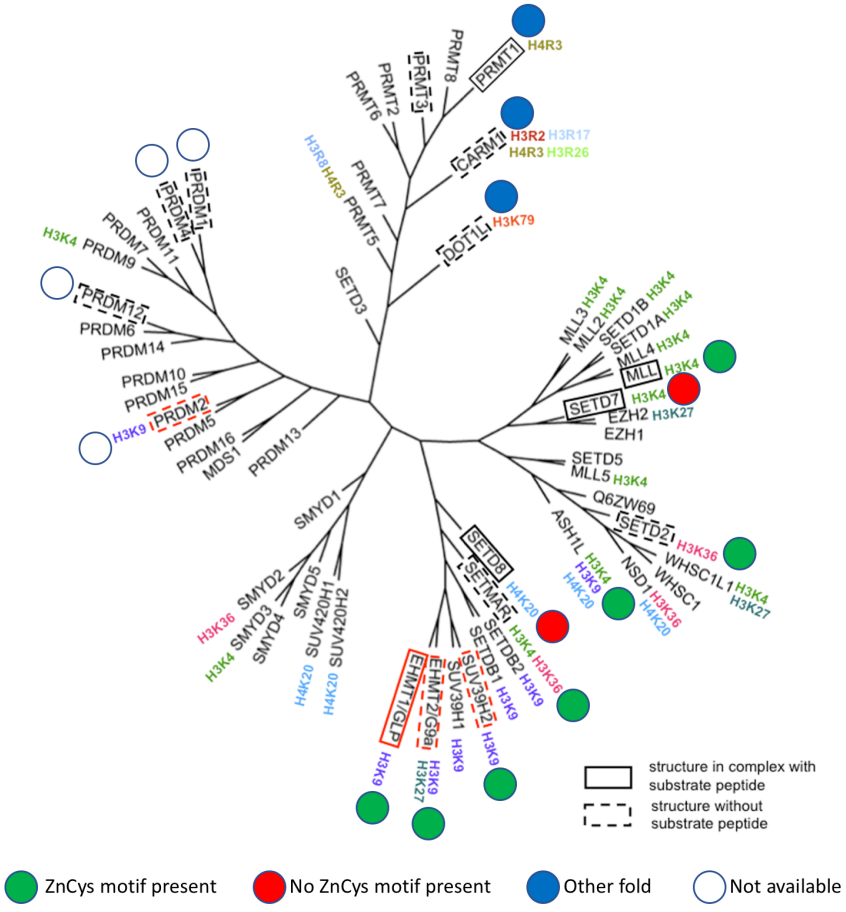


Figure 6-10. Analysis of the presence of a ZnCys motif in the post-SET domain of lysine and arginine methyltransferases based on available crystal structures. Adapted from a phylogenetic tree by Wu et al. [65].

6.3.1.6. Proposed mutations of G9a Q1169

In order to generate a hole for the application of the bump-and-hole approach, a point mutation on G9a's Q1169 to glycine (e.g. Q1169G) or alanine (e.g. Q1169A) is proposed. This residue does not appear to be conserved among H3K9 KMTs (see Figure 6-12), which makes it a good candidate to apply the bump-and-hole approach while preserving the overall fold. However, it is necessary to assess if the side chains of the residues located at the equivalent position of G9a's Q1169 occupy a similar space for other H3K9 methyltransferases to check if they would clash with SAM analogues modified at C2 position. Focusing on known H3K9 KMTs (G9a, GLP, SUV39H1, SUV39H2, SETDB1, SETDB2, ASH1L and PRDM2), crystallographic structures of the SET domain and post-SET domain were found for G9a (PDB:2O8J), GLP (PDB:2RFI), SUV39H2 (PDB: 2R3A) and ASH1L (PDB:4YNM). The rest of the enzymes in the analysis did not present available structures of their post-SET domains, therefore a sequence alignment was performed. A visual analysis of the post-SET domain of these enzymes is shown in Figure 6-11.

As can be observed in Figure 6-11, the ZnCys motif appears to be a recurrent structural feature in H3K9 KMTs. It also shows visually that the residue to mutate to accommodate bulky groups at SAM's C2 position is G9a's Q1169, located between two of the conserved cysteines (C1168 and C1170) of G9a's post-SET domain. For example, a mutation of this residue into glycine or alanine would generate the necessary space to accommodate the inserted groups into SAM's C2 position. This will be further evaluated in section 6.3.2. by generating comparative models of this mutant candidate.

In addition to the structural analysis, a sequence alignment was performed for H3K9 KMTs including those lacking a crystal structure to evaluate if the ZnCys motif cysteines were conserved in their SET domain. The result of this analysis is shown in Figure 6-12.

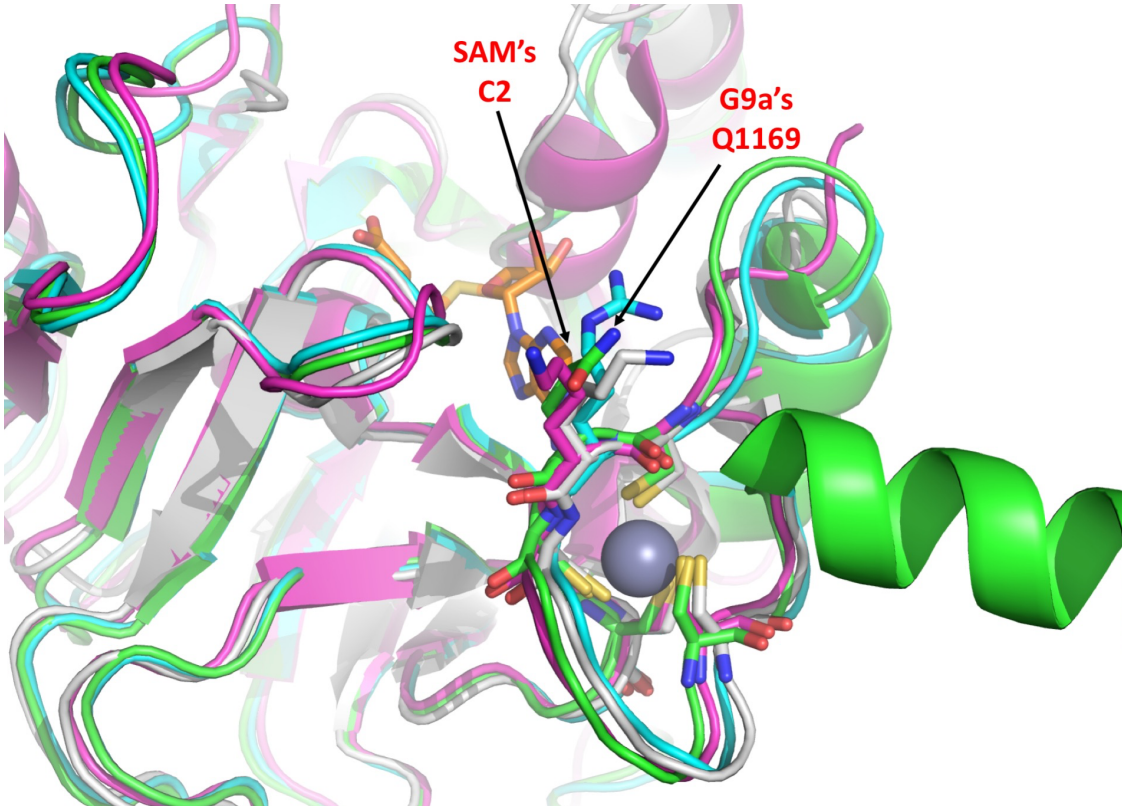


Figure 6-11. The ZnCys motif is frequent in H3K9 KMTs. G9a is shown in green, GLP in cyan, SUV39H2 in magenta and ASH1L in light grey. The C2 position of SAM is indicated, as well as the Q1169 residue of G9a and its equivalent residues in the other KMTs (depicted as sticks).

```
G9a       1096  VYCIDARYYGNISRFINHLCDPNIIPVRVFMLHQDLRFPRIAFFSSRDIRTGEELGFDYG   1155
GLP       1184  VYCIDARFYGNVSRFINHHCEPNLVPVRVFMAHQDLRFPRIAFFSTRLIEAGEQLGFDYG   1243
SUV39H1    307  VYTVDAAYYGNISHFVNHSCDPNLQVYNVFIDNLDERLPRIAFFATRTIRAGEELTFDYN    366
SUV39H2    314  EFTVDAARYGNVSHFVNHSCDPNLQVFNVFIDNLDTRLPRIALFSTRTINAGEELTFDYQ    373
SETDB1    1207  CYIIDAKLEGNLGRYLNHSCSPNLFVQNVFVDTHDLRFPWVAFFASKRIRAGTELTWDYN   1266
SETDB2     635  VFLLDATKEGNVGRFLNHSCCPNLLVQNVFVETHNRNFPLVAFFTNRYVKARTELTWDYG    694
ASH1L     2206  GMVIDSYRMGNEARFINHSCDPNCEMQKWSVNG----VYRIGLYALKDMPAGTELTYDYN   2261
PRDM2     1681  ------------SLRLASRCSPPAAPYITR-QYRKVKAPAAAQFQGPFFKE---------   1718
                            .  :   * *                    . :    .

G9a       1156  DRFWDI---------------------KSKYFTCQCGSEKCKHSAEAIALEQSRLARLDP   1194
GLP       1244  ERFWDI---------------------KGKLFSCRCGSPKCRHSSAALAQRQASAAQEAQ   1282
SUV39H1    367  MQVDPVDMESTRMDSNFGLAGLPGSPKKRVRIECKCGTESCRKYLF--------------    412
SUV39H2    374  MKGSGD-ISSDSID--------HSPAKKRVRTVCKCGAVTCRGYLN--------------    410
SETDB1    1267  YEVGSV----------------------EGKELLCCCGAIECRGRLL--------------   1291
SETDB2     695  YEAGTV---------------------PEKEIFCQCGVNKCRKKIL--------------    719
ASH1L     2262  FHSFNV----------------------EKQQLCKCGFEKCRGIIGGKSQRVNGLTSSKN   2299
PRDM2     1719  ------------------------------------------------------------   1718
```

Figure 6-12. Conservation of the ZnCys motif in human H3K9 methyltransferases. The cysteine residues involved in the ZnCys motif are highlighted in orange. The mG9a candidate Q1169 and its equivalent residues in other KMTs are highlighted in blue.

From Figure 6-11 and Figure 6-12, it appears that the residues at equivalent positions to Q1169 in G9a are also bulky (R, Q, K), also possibly reducing the binding specificity of these wild type KMTs for the proposed SAM analogues modified at C2 position. This is further evidence supporting the Q1169 mutation to alanine or glycine to perform the bump-and-hole for G9a. SETDB1 is the notable exception as it has a cysteine in the equivalent position; however, there is no available crystallographic structure of its SET domain to analyse this further. PRDM2 presents a very different SET domain and lacks a post-SET domain, but the structure of PRDM2 lacks a co-crystallised SAM or SAH (PDB: 2QPW), therefore there is no evidence of the exact binding mode of SAM to this protein [65] to validate if a C2-modified SAM could be utilised by it.

6.3.2. Computational screening of proposed mG9a candidates using Rosetta

Models of mG9a candidates were generated using the comparative modelling functionality of the Rosetta package as described in section 6.2.2.2. . First, mutations in conserved residues were performed to assess if they affect the fold of the pocket. Then, a structure of the mG9a Q1169G was generated to assess if the fold was conserved upon its mutation.

6.3.2.1. Negative control: mutations to conserved residues

A model was generated using Rosetta to assess the effect of C1168G, C1170G and C1175G mutations on the folding of the post-SET domain and therefore validate our hypothesis of avoiding mutations in conserved residues to preserve G9a's fold. The result of this study is shown in Figure 6-13.

As can be observed in Figure 6-13, the post-SET domain of G9a is modified when C1168, C1170 and C1175 are mutated to glycine, with the post-SET domain getting closer to SAM's binding pocket and making it smaller. This agrees with the evolutionary analysis in section 6.3.1.3. , which showed the high conservation of these residues and therefore their importance to preserve the correct fold of the post-SET domain. The SAH binding pocket is therefore altered, possibly representing a loss of G9a's activity due to its inability to bind

correctly to the SAM analogue, as seen in Figure 6-14. For this reason, it seems clear that mutations in highly-conserved residues should be avoided to preserve the fold and therefore the function of G9a.

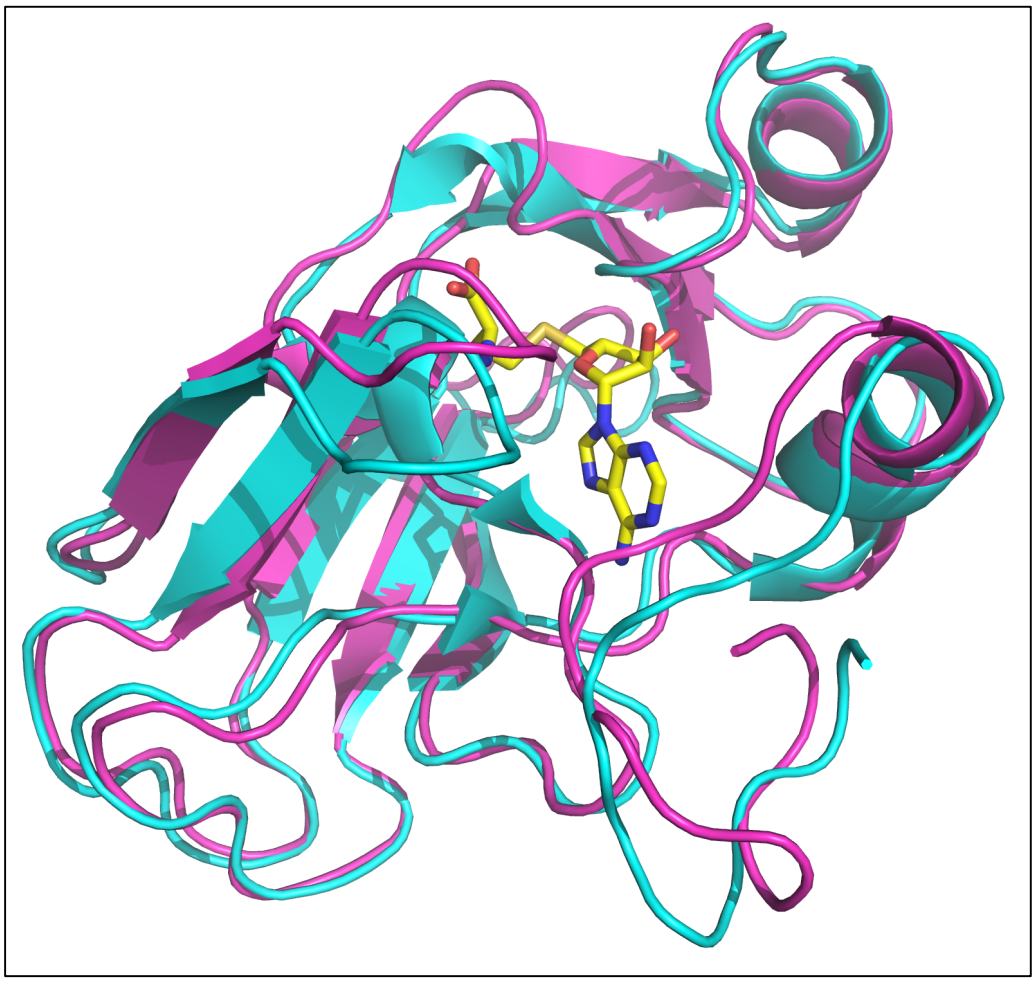


Figure 6-13. G9a's post-SET domain fold is altered (RMSD = 2.203 Å) when C1168, C1170 and C1175 are mutated to glycine. Endogenous G9a is depicted in cyan, the model of mG9a in pink and SAH in yellow.

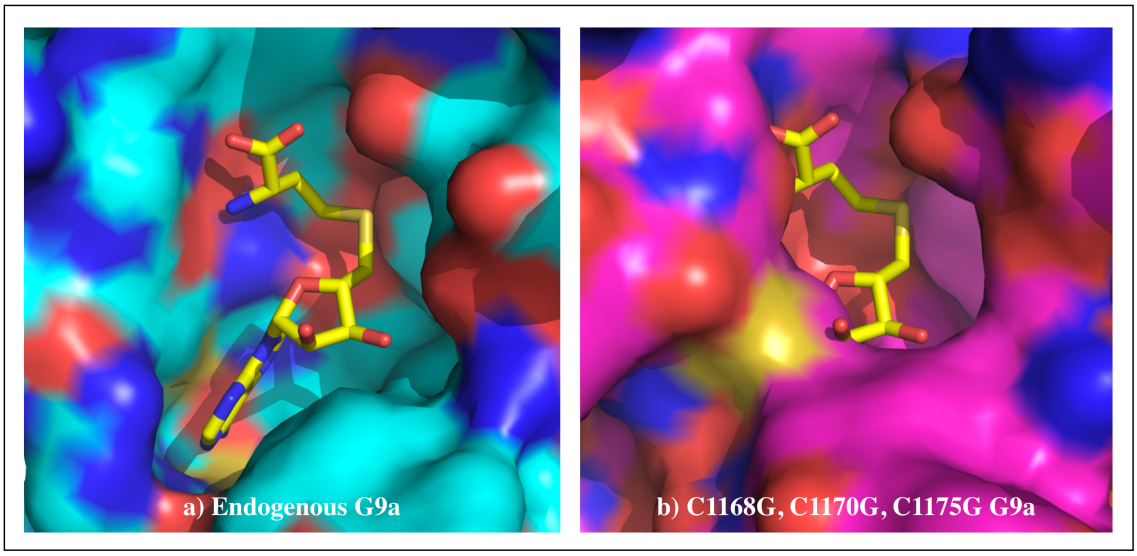


Figure 6-14. Surface representation of native SAH bound to endogenous G9a (a) and the same SAH structure superposed on an aligned model of C1168G, C1170G and C1175G G9a (b). The mutations in the highly-conserved cysteine residues strongly alter the binding pocket, possibly resulting in a loss of G9a's activity.

6.3.2.2. Rosetta folding predictions for proposed mG9a candidates

A model of a mutant Q1169G was generated with Rosetta to assess the stability of the G9a fold upon its mutation. A comparison of the model generated aligned with the structure of endogenous G9a is shown in Figure 6-15. This result is consistent with the expectation that there is only a relatively small change in the fold of G9a when Q1169 is mutated to glycine. This analysis showed the low conservation of this residue and therefore the higher probability of preserving the fold of G9a when mutated.

An analysis of the binding of Phe-SAH towards endogenous G9a and the G9a Q1169G was performed to assess if it clashes with the endogenous but not the mutant version. As seen in Figure 6-16a, there exists a steric clash between endogenous G9a and Phe-SAH that would likely impede the correct binding of this SAM analogue towards wild type G9a. However, for G9a Q1169G (Figure 6-16b), the mutation generates enough space to accommodate the phenyl moiety and therefore avoid steric clashes, facilitating the preservation of the original binding mode of native SAM. This evidences that the Phe-SAM cofactor analogue is likely to be specific towards G9a Q1169G.

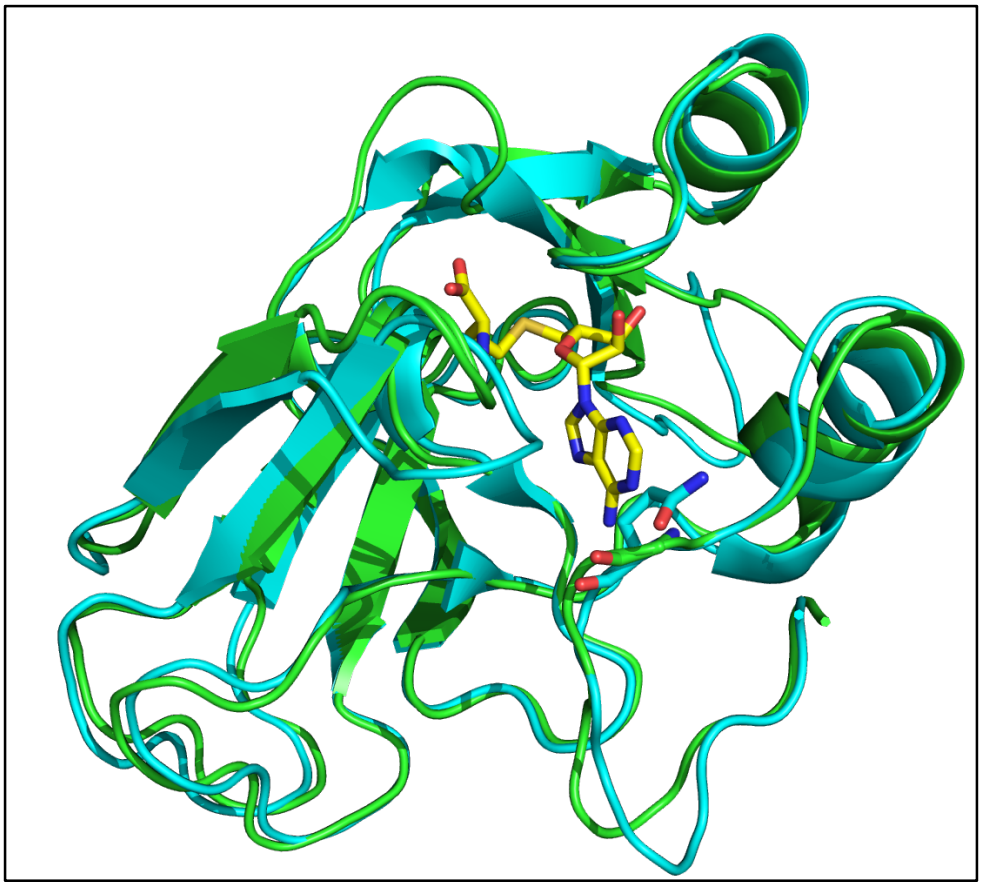


Figure 6-15. G9a's post-SET domain conserves its fold (RMSD = 0.938) when Q1169 (cyan sticks) is mutated to glycine (green sticks). Endogenous G9a is depicted in cyan, the model of Q1169G in green and SAH in yellow.

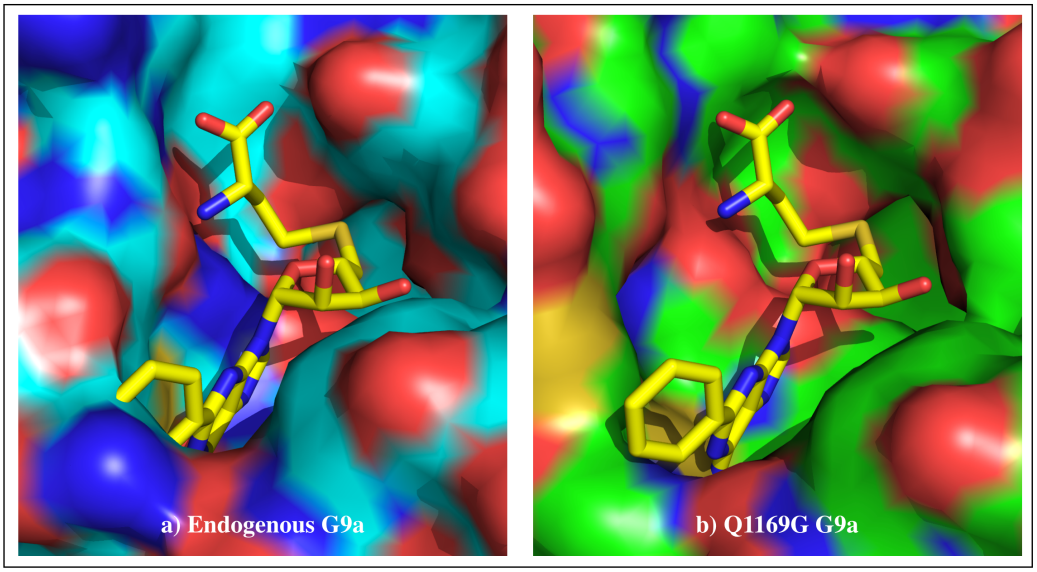


Figure 6-16. Surface representation of C2 Phe-SAH bound to endogenous G9a (a) and the same C2 Phe-SAH structure superposed on an aligned model of G9a Q1169G (b). While the phenyl group causes a steric clash with endogenous G9a, the Q1169G mutation generates an additional space that could suffice to accommodate the phenyl bump.

6.3.3. MD simulations of G9a Q1169A with SAM and SAM analogues

Alongside preserving the fold and avoiding clashes, an adequate binding of the SAM analogue to mG9a is required for the methyl transfer to occur. To analyse if the binding is preserved when the mutation is performed, MD-generated trajectories of SAM and the four C2-modified SAM analogues mentioned in Chapter 4 and Chapter 5, each bound in the SET domain of G9a Q1169A were obtained using the computational methodology described in Chapter 5. The structure of G9a Q1169A to use in the docking of the SAM cofactor analogues was obtained based on the structure of G9a (PDB: 2O8J) using the mutagenesis functionality of PyMOL [216]. The reason why the model generated in Rosetta was not used was to avoid any conformational change in G9a's side chains that could have been generated during the process and could alter the docking of SAM or the SAM analogues. The parametrization of the ZnCys motif gave as a result a distribution of the positive charge of the central Zn^{2+} ion into its surrounding cysteines so the actual partial charge of the Zn^{2+} ion resulted +0.140688.

The reason why the MD simulations were performed using G9a Q1169A and not G9a Q1169G is that substituting Q1169 by alanine retains the beta carbon while still generating a hole in the structure, due to the removal of the rest of the glutamine side chain. However, a mutation to glycine would remove this side chain beta carbon, providing additional flexibility in the peptide chain that could lead to stronger undesired conformational changes.

To obtain an estimate of when the systems become equilibrated and to assess the overall stability of the complex, the root-mean-square-deviation (RMSD) of backbone atoms (CA, C and N) was calculated with regards to the initial production frame for the different G9a Q1169A – SAM cofactor analogue pairs production runs as shown in Figure 6-17. This figure shows that the RMSD of the runs increased during the first 20 ns, but it remained quasi-stable for most of the cases from there on with the exception of Pip-SAM, where longer simulation times could be performed to ensure equilibration. The average RMSD values oscillate between 2.5 and 4.5 Å. Because of the systematic increase in the RMSD in the first 20 ns of each run and to compare equal ranges between the different KMTs, the

first 20 ns were discarded for the calculation of probability density distributions of distances and angles as shown in the next section.

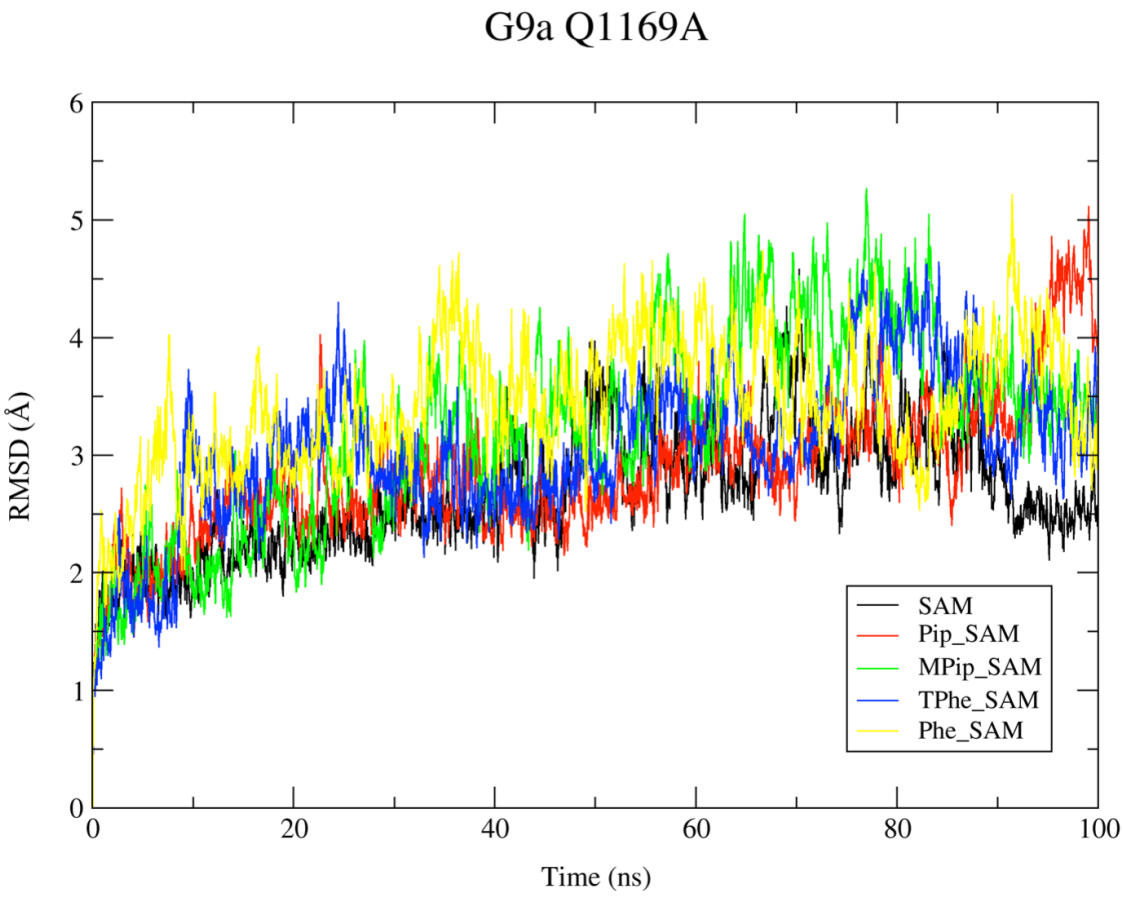


Figure 6-17. Root-mean-square-deviation (RMSD, Å) for the backbone atoms of G9a Q1169A with regards to the initial frame for the production trajectory.

6.3.3.1. G9a Q1169A distances and angles analysis

Figure 6-18 shows a similar analysis to the one performed for G9a in section 5.3.3.1. of Chapter 5 but for the proposed mG9a Q1169A. The distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 in G9a Q1169A is predominantly around 4.5 Å for native SAM. When this distance is analysed with the SAM cofactor analogues proposed, it is higher for most of them but not Phe-SAM: between 6 and 13 Å for Pip-SAM, between 10 and 20 Å for MPip-SAM, around 10 Å for TPhe-SAM but only 4 Å for Phe-SAM. Therefore, and as opposed to the analysis for endogenous G9a (section 5.3.3.1. of Chapter 5) where no cofactor resembled the distance

achieved with native SAM, the distance obtained with Phe-SAM is comparable, making it a good candidate to be a SAM cofactor analogue for G9a Q1169A.

A similar approach was carried out for the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of Y1067 in G9a Q1169A, and the results are shown in Figure 6-19. The angle formed between the aforementioned atoms with native SAM is predominantly around 145º, therefore this is assumed to be the angle enabling the methyl transfer. The predominant angles obtained with SAM analogues were distinct depending on the cofactor: angles around 50 or around 130º for Pip-SAM, around 75º for MPip-SAM, 50º for TPhe-SAM and 140º for Phe-SAM. Therefore, the only SAM cofactor analogue that seems to provide comparable angles for the methyl transfer compared to native SAM is Phe-SAM.

In conclusion, after analyzing both distance and angle results together, it seems that the best SAM cofactor analogue to resemble the efficient methyl transfer in native SAM is Phe-SAM, showing a distance of 4 Å and an angle of 140º for Phe-SAM vs a distance of 4.5 Å and an angle of 145º for native SAM. This also shows the positive effect of the point mutation on accepting the Phe-SAM analogue, as the results of Phe-SAM with endogenous G9a differed extensively from the ones of native G9a with native SAM. More specifically, results in section 5.3.3.1. of Chapter 5 showed a distance of 8 Å and an angle of 85º for Phe-SAM vs a distance of 5.5 Å and an angle of 150º for native SAM. In addition, for SETD7 the distances and angles also differed between Phe-SAM and native SAM, as seen in section 5.3.3.2. of Chapter 5, where Phe-SAM presented a distance of 9 Å vs 6 Å for native SAM, and an angle between 80 and 140º vs 145º for native SAM, also experimentally showing a reduction of around 40% SETD7 methylation activity with regards to native SAM as seen in section 4.3.2.4. . Therefore, Phe-SAM also appears to be a suitable cofactor to differentiate between the activities of G9a Q1169A and SETD7. In summary, Phe-SAM seems selective towards Q1169A with respect to both G9a and SETD7 based on the analysis of the relevant distances and angles to enable the methylation activity.

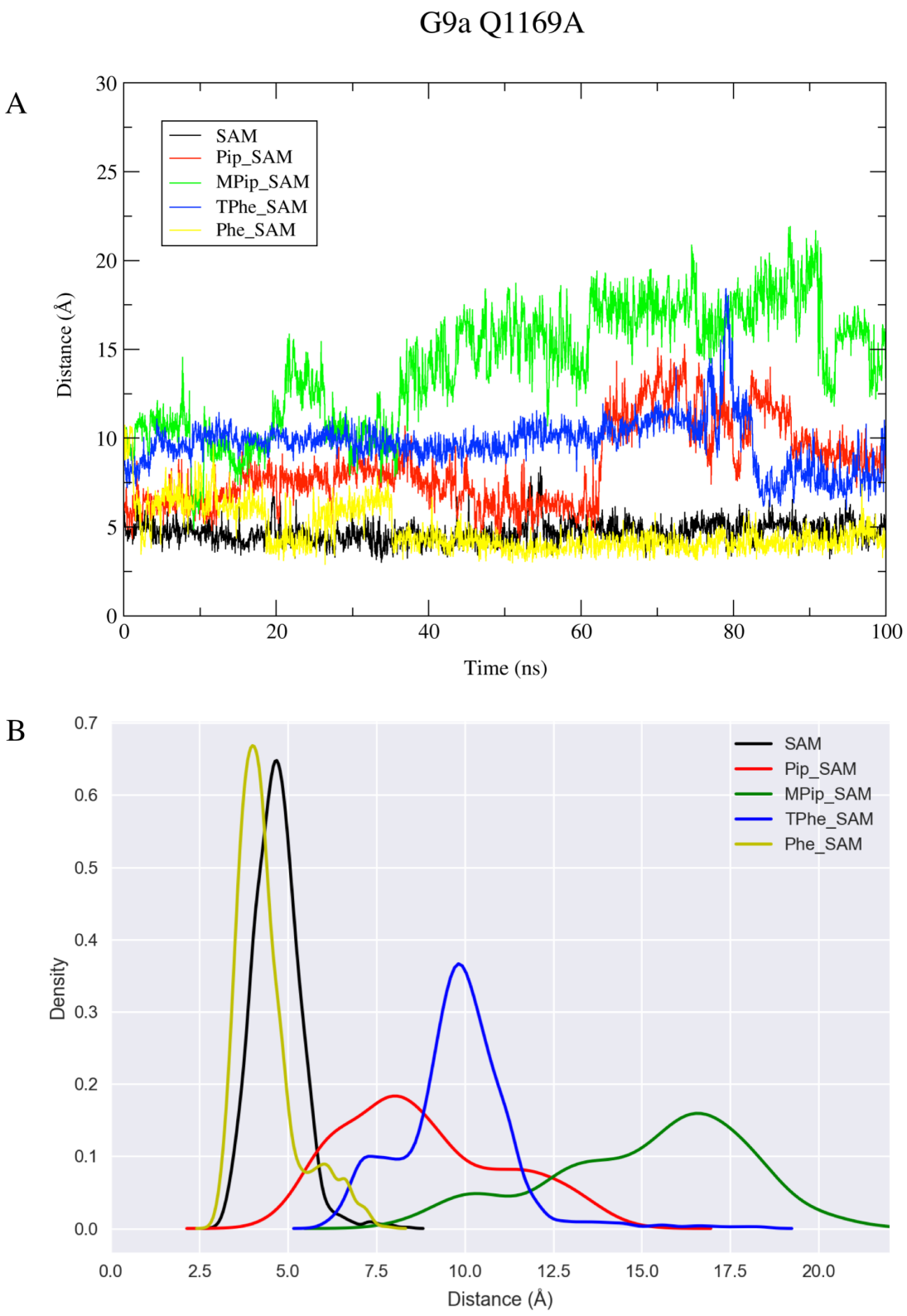


Figure 6-18. A) Evolution of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 during 100 ns of simulation. B) Probability distribution function of the distance between the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 for the last 80 ns of the simulation.

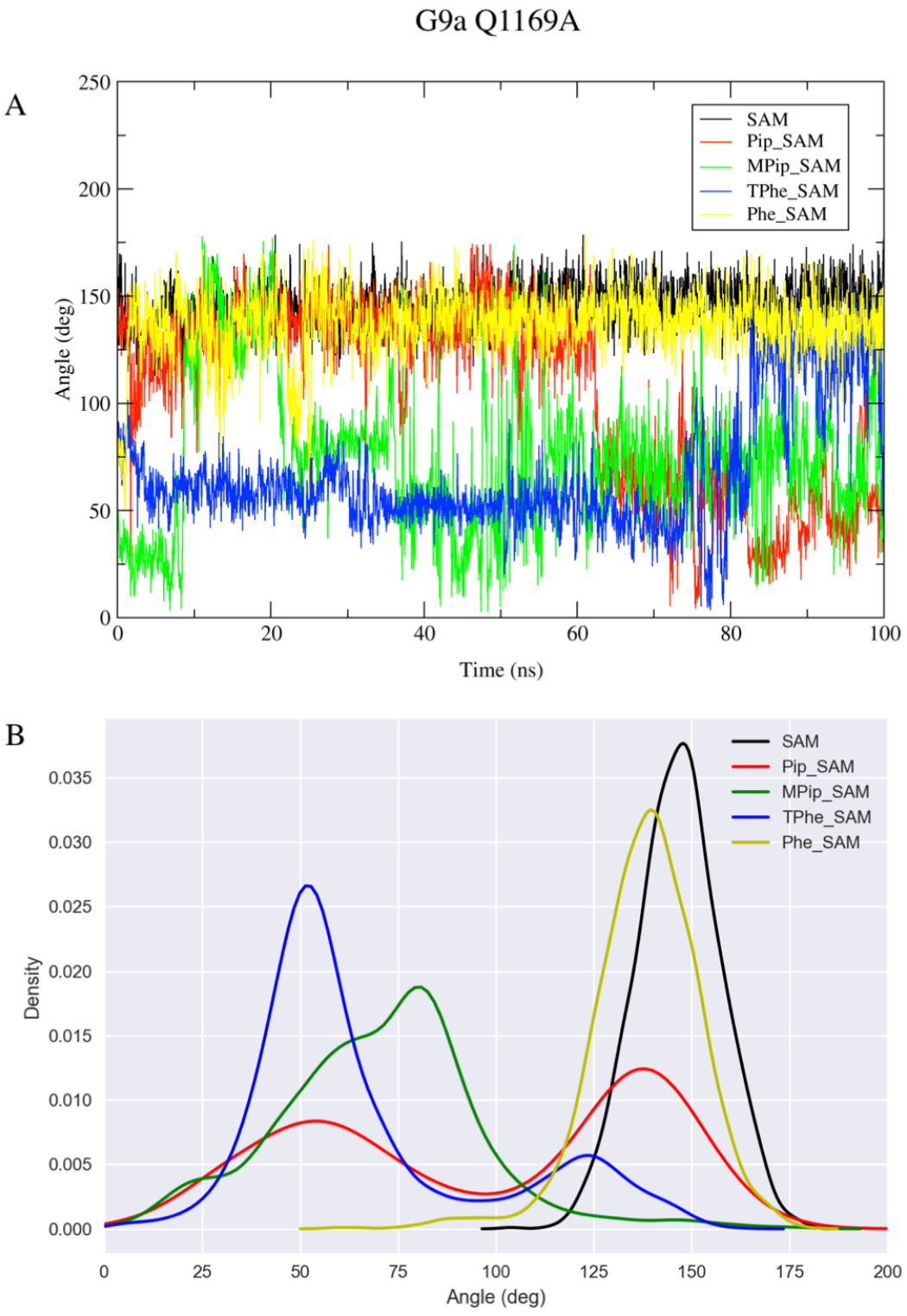


Figure 6-19. A) Evolution of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 during 100 ns of simulation. B) Probability distribution function of the angle between the sulphur atom of SAM, the carbon atom of the methyl transfer group of SAM and the oxygen atom of the hydroxyl group of G9a Q1169A's Y1067 for the last 80 ns of the simulation.

6.3.3.2. Representative structure of the major cluster

The representative structures of the major cluster of each MD simulation were analysed with a focus on the enzyme-cofactor pair that was previously shortlisted for its orthogonal application, Phe-SAM and G9a Q1169A. Figure 6-20 shows the differences in binding of Phe-SAM to G9a and G9a Q1169A.

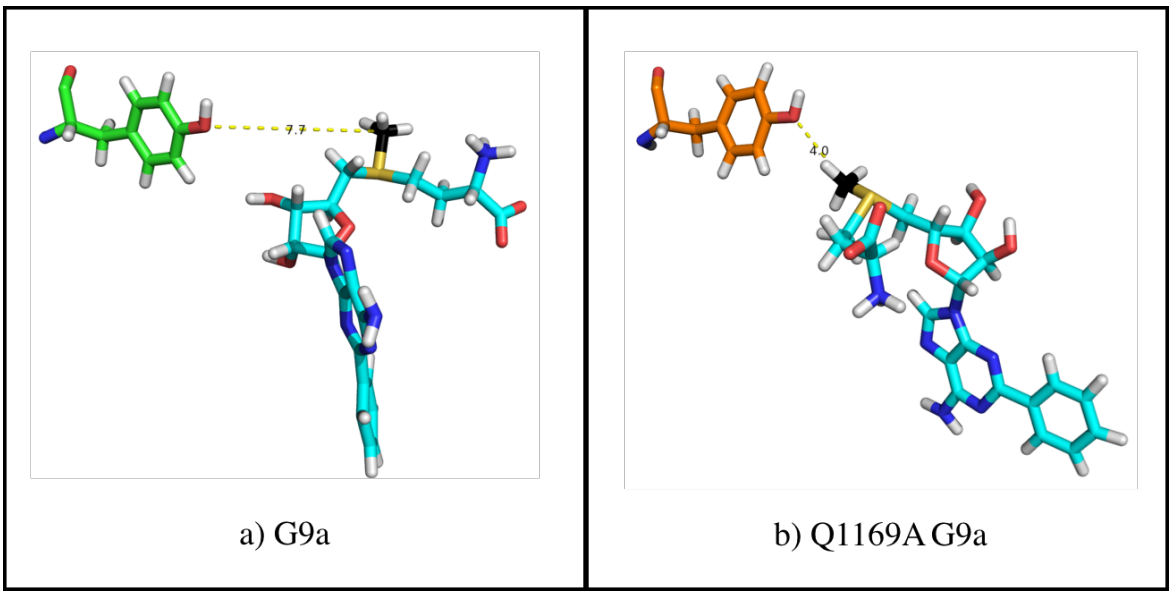


Figure 6-20. Differences in the relative positioning of Phe-SAM (cyan sticks) to a) G9a's Y1067 (green sticks) and b) G9a Q1169A's Y1067 (orange sticks).

The representative structures of the main clusters of the MD trajectories of G9a and G9a Q1169A with phenyl-SAM are shown in Figure 6-20. This figure indicates that both the angle and the distance between the methyl group (depicted in black) and Y1067 are not adequate in G9a, but in G9a Q1169A both distance and angle resemble the values obtained with native SAM that facilitate the transfer of the methyl group. This reveals a possible higher methylation activity of G9a Q1169A than G9a when using phenyl-SAM cofactor analogues. This information together with the distance and angle analyses shown on section 6.3.3.1. indicate that phenyl-SAM is the preferential SAM cofactor analogue to differentiate between the activities of G9a and G9a Q1169A, and therefore to implement the bump-and-hole approach on G9a.

The interactions of Phe-SAM with G9a Q1169A in the representative structure obtained from clustering were compared to the interactions between SAH and endogenous G9a in their crystallographic structure (PDB: 2O8J) using the program Ligplot [288] to analyse their preservation. This analysis is summarised in Figure 6-21.

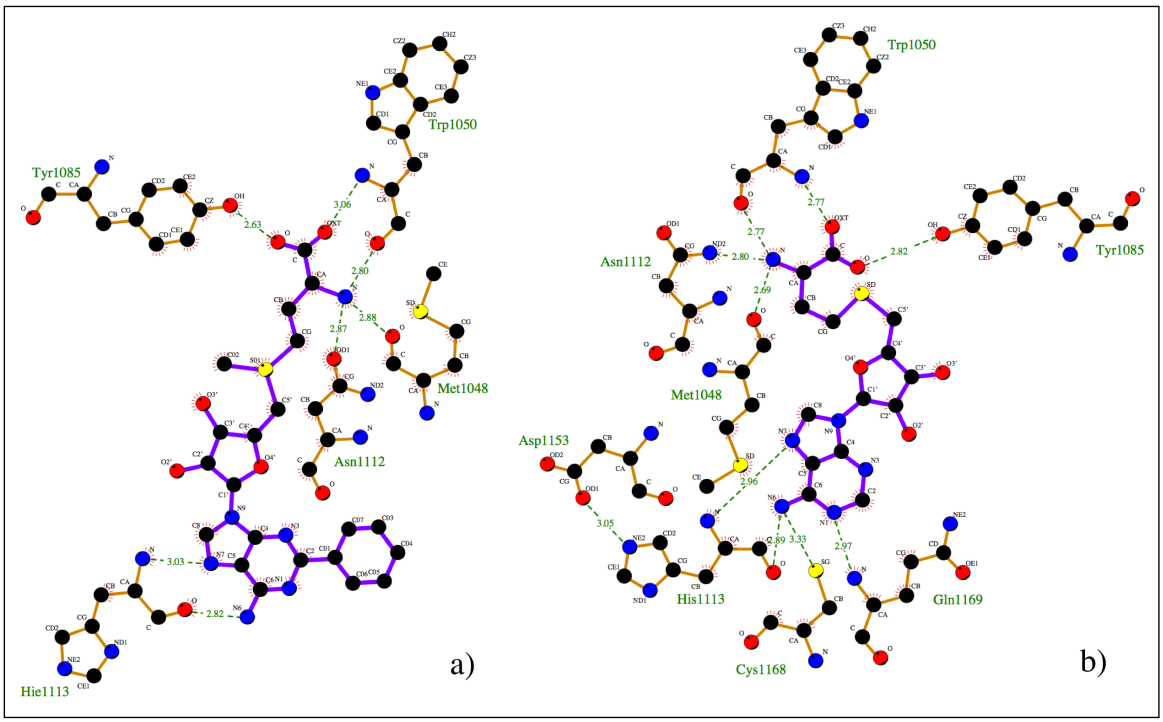


Figure 6-21. The interactions of Phe-SAM with G9a Q1169A (a) appear mostly preserved when compared to the interactions between SAH and endogenous G9a (b), except for the residues in the post-SET domain.

As can be observed in Figure 6-21, the interactions of Phe-SAM with G9a Q1169A are predominantly conserved when compared to the interactions between SAH and endogenous G9a. More specifically, the interactions with G9a that appear conserved when transitioning to the C2 phenyl-SAM - Q1169A system are: M1048, W1050, Y1085, N1112 and H1113. The two residues that lose the interactions are in the post-SET domain: C1168 and Q1169, the residue that was mutated to alanine. Overall, this indicates that there is an additional rearrangement of the post-SET domain to accommodate the inserted phenyl group, but the overall binding mode of SAM is preserved.

6.3.3.3. MM-PBSA calculation of the binding affinities

The binding affinities of SAM cofactor analogues towards G9a Q1169A were calculated with the program mmpbsa.py to elucidate if they were similar to the binding affinity of SAM. The calculation was done using the last 80 ns of the MD production trajectories, as specified in section 5.2.5.2. of Chapter 5. A special focus was on Phe-SAM as it appeared to be the best SAM cofactor analogue to resemble the characteristic distances and angles that allow an efficient methyl transfer with native SAM for each of the proteins (see sections

6.3.3.1. and 6.3.3.2. for reference). Table 6-5 shows a summary of the energetic analysis of SETD7 with different SAM cofactor analogues.

Table 6-5. Energetic analysis for the G9a Q1169A-SAM analogue complexes as obtained by MM–PBSA calculations.

Energy (kcal/mol)	M1	M2	M3	M4	M5
	SAM	Pip-SAM	MPip-SAM	TPhe-SAM	Phe-SAM
ΔE_{vdw}	-43.71	-40.33	-27.47	-38.21	-35.51
ΔE_{el}	-193.17	-191.22	-198.41	-223.26	-227.50
ΔE_{pol}	217.45	209.12	210.53	234.01	232.24
ΔE_{ap}	-4.32	-4.80	-3.46	-3.78	-4.00
ΔG_{gas}	-236.89	-231.55	-225.89	-261.47	-263.00
ΔG_{solv}	213.13	204.33	207.08	230.24	228.24
$\Delta G_{bind\ MM\text{-}PBSA}$	-23.76	-27.22	-18.81	-31.23	-34.76
$\Delta\Delta G_{MM\text{-}PBSA}$ (relative to SAM)	-	-3.46	4.95	-7.47	-11.00

where ΔE_{vdw} is the van der Waals energy, ΔE_{el} is the electrostatic energy, ΔE_{pol} is the polar solvation energy, ΔE_{ap} is the non-polar solvation energy, ΔG_{gas} is the total gas phase free energy, ΔG_{solv} is the total solvation free energy, ΔG_{bind} is the total binding energy and $\Delta\Delta G_{MM\text{-}PBSA}$ is calculated as the difference in binding energy between G9a Q1169A and SAM and G9a Q1169A and the corresponding SAM analogue. All the energies are shown in kcal/mol.

It was observed in sections 6.3.3.1. and 6.3.3.2. that the best SAM cofactor analogue for G9a Q1169A was Phe-SAM, based on Phe-SAM having similar distances and angles to the ones of native SAM obtained through MD simulations. The binding affinity estimate shows that the binding is effective, as the binding of G9a Q1169A to Phe-SAM is 11 kcal/mol stronger than for native SAM. Experimental data is necessary to discern if the binding is not excessive and therefore assess whether Phe-SAM acts as a cofactor and not as an inhibitor to G9a Q1169A. Besides, the results show that Pip-SAM and TPhe-SAM are bound with more similar values to SAM, whereas MPip-SAM binds with 4.95 kcal/mol less binding energy with respect to SAM.

In summary, MD simulations show that Phe-SAM might be an adequate SAM analogue to achieve selectivity towards G9a Q1169A. However, experimental validation is needed to assess if Phe-SAM does not act as an inhibitor due to its stronger estimated binding.

6.3.4. Experimental validation

The characterization of the G9a-SAM complex in section 6.3.1. , the computational screening of proposed G9a mutants using Rosetta in section 6.3.2. and the molecular dynamics screening in section 6.3.3. revealed that the ideal bump-and-hole pair applied to G9a is G9a Q1169A with C2-phenyl-SAM. To test the validity of this finding, recombinant G9a Q1169A was cloned, expressed and purified as described in section 6.2.3. . The samples of G9a Q1169A obtained after FPLC purification were analysed via SDS-PAGE analysis as described in chapter 4, section 4.2.2.3. . The resulting protein gel is shown in Figure 6-22.

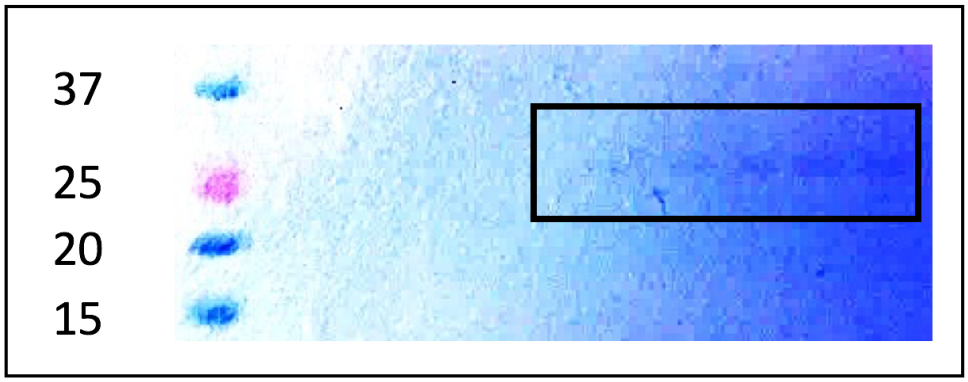


Figure 6-22. SDS-Page gel showing G9a Q1169A after FPLC purification.

As observed in Figure 6-22, the samples obtained after FPLC purification presented protein bands located between the 25 and 37 kDa marks; this fact provides a strong evidence that the purified protein is indeed G9a Q1169A as the estimated MW of this protein using ExPASy [209] is 34 kDa.

To further examine if the purified protein was indeed G9a Q1169A, an in-solution digestion and tandem mass spectrometry analysis was performed on the purified Q1169A samples as described in Chapter 4, section 4.2.2.4. . The peptide containing Q1169A was identified, therefore verifying that the point mutation had worked. However, G9a Q1169A had a coverage of 82.21% and a score of 594.36, the tenth highest of all the proteins found. The rest of the proteins found with higher scores than G9a Q1169A were from *E. coli*. This result indicates that although the protein was expressed, the purification was not as effective as for G9a, as the purification process did not manage to eliminate *E. coli* proteins. The purified samples of G9a Q1169A were flash frozen and stored in a solution 20 % glycerol at a

concentration of 0.53 mg/mL of protein as measured by a Nanodrop (Thermo Fisher Scientific).

In vitro assays were performed with G9a Q1169A under the same conditions described in Chapter 4, section 4.2.2.5. with heavy methyl SAM as the reference cofactor. However, there was no apparent heavy methylation appearing in the mass spectra. This is probably caused by the fact that the real concentration of G9a Q1169A might be much lower than the one measured by the Nanodrop as there are *E. coli* proteins with higher scores in the same sample, or a loss of function due to unfolding or aggregation issues during the purification. Future work should focus on finding the best conditions to optimise induction, expression, purification and handling of this protein to test if it indeed preserves similar activity levels independent of using native SAM or the C2-modified SAM analogues.

6.4. Conclusions and future work

This study provides a proof-of-concept for engineering a bump-and-hole system for KMT target identification. The case study focused on the G9a methyltransferase due to its importance in the development of several cancers and the availability of crystallographic structures detailing its binding interactions with the SAM cofactor.

The methodology involved a thorough structural and conservation analysis of G9a's binding pocket to propose mG9a candidates that could fit SAM cofactor analogues. This was followed by comparative modelling with energy refinement studies to assess the impact of the proposed mutations on the structure, and molecular dynamics simulations to test the suitability of SAM cofactor analogues. Lastly, *in vitro* assays using the recombinant mutant enzyme were performed to validate co-factor utilization in histone methylation.

Following the proposed design strategy, we identified that the Q1169 residue of G9a is a suitable candidate to mutate to alanine to generate a complementary "hole" for a synthetic bump on the C2 atom of SAM. The bump was proposed to be a phenyl group as it provides a rigid structure that clashes with the Q1169 residue side chain of G9a but does not clash when mutating this residue to flexible and small residues such as glycine or alanine.

Homology modelling with energy refinement simulations with Rosetta were then performed and predicted that G9a conserves its fold when Q1169 is mutated to glycine, whereas the binding pocket structure is lost when mutating more conserved residues such as the cysteines in the ZnCys motif.

The MD approach introduced in Chapter 5 was applied to G9a Q1169A, consisting in analyzing the conformational properties and binding modes of four SAM cofactor analogues (piperidine-SAM, methylpiperazine-SAM, thiophenyl-SAM and phenyl-SAM) and binding free-energy (MM–PBSA) calculations. Distance and angle analyses showed that phenyl-SAM is the best analysed SAM cofactor analogue to differentiate between the activity of G9a Q1169A and other KMTs such as G9a and SETD7. The reasoning to conclude this was that the characteristic distances and angles that allow an efficient methyl transfer differed greatly between phenyl-SAM and native SAM when G9a and SETD7 were considered. However, these distances and angles were similar for G9a Q1169A. In addition, when observing the representative structures of the major clusters of the MD trajectories of G9a

and G9a Q1169A with phenyl-SAM, the binding mode was not adequate for methyl transfer in G9a, but in G9a Q1169A the binding mode resembled the one of native SAM. Besides, the comparison of the interactions between SAH and endogenous G9a in the crystallographic structure (PDB: 2O8J) with the representative structure of a molecular dynamics simulation of Phe-SAM with G9a Q1169A demonstrated that the cofactor-enzyme interactions are mostly preserved, indicating the suitability of Phe-SAM with G9a Q1169A.

An MM-PBSA calculation was also performed on the generated MD trajectories, showing that the binding of Phe-SAM to G9a Q1169A was 11 kcal/mol stronger than for native SAM. Experimental data is required to discern if the tighter binding enables the use of Phe-SAM as a cofactor and not as an inhibitor for G9a Q1169A. Overall, these arguments suggest that, when using Phe-SAM, a more effective methyl transfer will occur with G9a Q1169A than with G9a or SETD7. Therefore, Phe-SAM seems the ideal candidate to implement the bump-and-hole approach of G9a. To our knowledge, this study also constitutes the first implementation of molecular dynamics simulations applied to a bump-and-hole design.

Regarding the experimental validation, recombinant G9a Q1169A was cloned, expressed and purified for *in vitro* assays with nucleosome substrates to test its activity with heavy methyl SAM. However, the assays were inconclusive as no signs of heavy methylation were detected. The expression, purification and handling of G9a Q1169A should be optimised to obtain an active version of the enzyme and test our MD simulation findings further. Future work should focus on finding the best conditions for G9a Q1169A expression and purification to get an active enzyme to test if it indeed presents similar methylations activity levels independently of using native SAM or the C2-phenyl SAM analogues.

One of the most significant novelties of this project is the methodical design strategy involving comparative modelling simulations, conservation studies and molecular dynamics prior to experimentation, which differs largely from existing bump-and-hole approaches where only steric clashes are considered or extensive experimental screening is the key driver. However, there are also certain limitations, such as the introduction of the SAM analogues into the cell membrane, which need to be overcome to achieve a truly cellular profiling method of PTMs.

Chapter 7

Conclusions and future work

7.1. Conclusions

The activity of lysine methyltransferases (KMTs) is crucial for the regulation of gene expression and plays an important role in the onset, progression and maintenance of epigenetically-regulated diseases, such as cancer. Being able to unambiguously differentiate between the activity of different KMTs is therefore critical for the development of more specific treatments for diseases such as brain or lung cancer [290]. Despite the interest of the pharmaceutical industry towards epigenetic therapies, with over 30 different epigenetic drugs undergoing clinical trials [54], there are currently no biologically-consistent methods in the literature to directly assess which KMT contributes to specific methylation events, as analytical techniques such as Western Blotting or mass spectrometry only provide evidence of cumulative and not individual KMT activity. This is because all KMTs utilise the same SAM cofactor to donate the same CH_3 methyl group to their respective protein targets; in the case of chromatin, there is significant redundancy in which KMTs can methylate the same histone lysine residues. This fact also limits the identification of non-histone targets for a particular KMT (e.g. methylation of transcription factors and other chromatin modifiers by G9a [295]).

To overcome the limitations of previous methods and address this technology gap, this thesis presented a "methylome" profiling technology based on engineering synthetic SAM cofactors to be specific towards certain KMTs. The SAM analogues were designed to contain a heavy methyl donor group (e.g. CD_3), thereby replacing the native CH_3 methyl to ensure that any new methylation events are distinguishable by MS proteomics and retain biological consistency as methylation-binding proteins (e.g. with chromodomains) can still "read" the heavy methylation mark on protein targets. This research specifically focused on the development of "bumped" heavy SAM cofactor analogues that can be utilised by only a subclass of KMTs. These enzyme-cofactor systems were engineered following two approaches, depending if the "holes" were either found or generated via mutagenesis: (1) the first approach identified naturally occurring "holes" in the structures of wild type KMTs that could potentially accommodate matching bumped SAM analogues; and (2) the second approach sought to engineer a complementary "hole" in a KMT structure using point mutation of the KMT to accommodate a bumped SAM analogue (i.e. generally known as the "bump-and-hole" method).

Regarding the identification of natural pockets in the structures of KMTs presented in Chapter 4, we developed a program that takes aligned KMT-cofactor structures as input and identifies pockets that are unique to one of the KMTs. Using this program, we identified a natural pocket in the SET domain of SETD7 which does not appear in post-SET containing KMTs, such as G9a, and we determined design atoms on SAM for incorporation of synthetic "bumps" to match the corresponding "hole" in the structure of SETD7. It was found that the C2 atom of SAM was the best to insert a "bump", as the groups inserted at this position did not generate any major clashes with the residues of SETD7 but clashed with G9a's post-SET domain Q1169 side chain. Four C2-modified SAM analogues (piperidine-SAM, methylpiperazine-SAM, thiophenyl-SAM and phenyl-SAM) were proposed and synthesised in collaboration with Dr Matthew J. Fuchter and Katie Addison (Department of Chemistry, Imperial College London) to selectively bind to SETD7.

The aforementioned SAM analogues were experimentally tested using *in vitro* assays with nucleosome substrates and recombinant SETD7 (and G9a as a control) to demonstrate their specificity towards SETD7. MS proteomics provided the primary readout for these assays in terms of SAM cofactor utilization and histone methylation. The assays revealed cofactor selectivity of the SAM cofactor analogues piperidine-SAM and methylpiperazine-SAM

towards SETD7. Notably, piperidine-SAM presented a similar (6% higher) methylation activity when compared to the one shown by native SAM with SETD7. In addition, piperidine-SAM presented a much lower (92% less) methylation activity when compared to the one shown by native SAM with G9a. These results suggest piperidine-SAM would be a suitable cofactor to help differentiating between the methylation activities of SETD7 and G9a.

In Chapter 5, an *in silico* approach based on molecular dynamics (MD) simulations was presented to computationally assess which of the SAM cofactor analogues could be utilised by the KMTs, with the motivation of using this tool to rank SAM analogues for synthesis. The variables investigated were the binding mode (distances, angles) and binding affinity of the different SAM analogue - KMT pairs, and comparing these results with those of native SAM to determine if the use of the cofactor analogues would result in the enzymatic transfer of a methyl group. The statistics of these variables were analysed using the MD trajectories via CPPTRAJ and the computation of the binding free-energy calculations using MM–PBSA. The four SAM cofactor analogues analysed (piperidine-SAM, methylpiperazine-SAM, thiophenyl-SAM and phenyl-SAM) were then ranked according to their capacity to be utilised specifically by G9a and SETD7, assuming that the more the similarities to the native SAM binding mode and affinity, the higher the compatibility.

Following the aforementioned approach, we found that piperidine-SAM presented the characteristic distances and angles that allow an efficient methyl transfer for SETD7, but differed extensively for G9a. The positioning of the catalytic tyrosine with regards to the methyl group of piperidine-SAM also agreed with the previous result, as it was conserved for SETD7 but not for G9a when compared to native SAM. Additionally, MM-PBSA calculations showed that piperidine-SAM was the SAM cofactor analogue presenting a more similar binding energy with regards to the one of native SAM for SETD7, only differing in a 0.35 kcal/mol tighter binding for piperidine-SAM. Collectively, these arguments suggest that piperidine-SAM enables selective methylation by SETD7. This is consistent with the results obtained experimentally (see Chapter 4, section 4.3.2.4.), which supports the feasibility of using of our molecular dynamics-based method as a screening tool to assess cofactor compatibility prior to experimentation.

In Chapter 6, this approach was extended to investigate the possibility of using a bump-and-hole technique for KMT target identification. G9a was selected for this case study, given the importance of its role in various cancers [294] and the availability of crystallographic structures with it bound to SAH [65]. For this proof-of-principle concept, we simplified the problem to focus on developing an mG9a that could selectively utilise one or more of the proposed bumped SAM analogues from Chapter 4. The proposed design methodology consisted of a bioinformatic analysis of the features of G9a's binding pocket to propose an mG9a candidate, which combined information relating to its structural characteristics and the evolutionary conservation of residues in the SAM binding pocket. We identified the Q1169 residue of G9a as a suitable candidate for mutagenesis to generate a "hole" to accommodate the bumps located at the C2 position of the SAM analogues.

The analysis and identification of the Q1169 residue was followed by comparative modelling with energy refinement simulations to assess the impact of the proposed mutation on the structure. Comparative modelling simulations with Rosetta showed that G9a conserved its fold when Q1169 was mutated to glycine (i.e. Q1169G), whereas the binding pocket structure was altered when more conserved residues such as the cysteines in the ZnCys motif were mutated to glycine (control). In addition, the interactions between SAH and endogenous G9a in the crystallographic structure (PDB: 2O8J) were proven to be highly similar to the interactions between phenyl-SAM and G9a Q1169A in the representative structure when applying the molecular dynamics tools in Chapter 5. In particular, it was found that phenyl-SAM presents the characteristic distances and angles that allow an efficient methyl transfer for G9a Q1169A, but differed extensively for wild type G9a or SETD7. The analysis of the representative structures of the major cluster of the MD trajectories of G9a and G9a Q1169A with phenyl-SAM revealed that the binding mode was not adequate for methyl transfer in G9a, but in G9a Q1169A it resembled the one of native SAM. Finally, an MM-PBSA calculation was performed, predicting that the phenyl-SAM binding to G9a Q1169A was 11 kcal/mol stronger than for native SAM.

The combined results suggest that phenyl-SAM could be utilised by G9a Q1169A for protein methylation. To validate this result experimentally, G9a Q1169A was cloned, expressed and purified, and *in vitro* assays were performed to test its basal activity with CD_3-SAM. The reaction products were analysed via liquid chromatography tandem mass spectrometry (LC-MS/MS), however no heavy methylation was observed. Future work

should focus on optimizing the conditions for its expression and purification to get an active mG9a Q1169A.

7.2. Implications and future work

This study presented the first evidence of the ability to design specific SAM cofactors to be selective towards a subclass of KMTs based solely on existing natural pockets to guide the cofactor design. One of the main novelties of this approach is the introduction of a program to geometrically identify structural "holes" that are present in a subset of KMTs. An important implication this method is that it can be easily extended to the design of other enzyme-cofactor systems (e.g. kinases and ATP analogues). The bump-and-hole approach has been successfully applied for target identification of kinases; however, the application of our methodology based on natural pockets avoids the necessity of generating point mutations in the enzymes and therefore eliminates the potential perturbations to enzyme functions that results from mutagenesis or overexpression in cellular models. Our method also avoids the need of a protein knockdown that could affect the activity of several cellular mechanisms related to the knockdowned enzyme

Notably, this assay could also be easily applied to differentiate between the activities of methyltransferases with similar targets. For example, between ASH1L and SETD7, which are both known to methylate histone H3K4. ASH1L was identified to contain a post-SET domain with a ZnCys motif (see Figure 6-11) in a similar way to G9a. Therefore, we hypothesise that our SAM analogue piperidine-SAM could also be used to differentiate between the activities of SETD7 and ASH1L as it should be selective towards SETD7. This could be tested experimentally in future work following from this thesis. In addition, if this technique is applied to other enzyme-cofactor analogue pairs, it could have broader implications in the field of systems biology, where further experimental data about the interactions between different proteins needs to be generated to develop more reliable models.

Compared to the currently most successful application of the bump-and-hole approach for KMTs based on G9a Y1154A, which was paired with sulfonium-alkyl SAM analogues [169], our approach presents several advantages to be applied in a cellular model. First, we

identified that G9a's Y1154 residue is highly conserved among KMTs, and it is known in the literature that its hydroxyl group of participates in the catalytic methyl transfer. The mutation of this residue into alanine could possibly alter the substrate specificity of G9a. In addition, the modified alkyl donor group holds no structural resemblance to a methyl mark; this implies that it cannot be recognised by the chromodomains of methylation-binding proteins, and therefore the downstream epigenetic functions of this mark are disrupted. Our approach for target identification of KMTs addresses both problems, since no catalytic residues need to be mutated (i.e. wild type SETD7 or G9a Q1169A, the latter mutation is not involved in catalysis and this position is not conserved across other KMTs) and our reporter group is a heavy methyl group (CD_3), which preserves all the downstream epigenetic functions and is therefore biologically compatible.

Starting from the results of this thesis, further research might well be conducted in some areas of improvement. First, our *in vitro* approach could be extended to cellular assays to facilitate the identification of non-histone targets. One possible challenge in achieving this goal is to overcome the low permeability of SAM cofactors through the cell membranes [211]. A possible solution to this challenge could be to synthesise these analogues inside the cells by mutating the human methionine adenosyltransferase (MAT) to accommodate bulky adenine analogues, in a similar way to the accommodation of bulky methionine analogues introduced by Luo's lab [212]. Additionally, we propose to extend the SETD7 analysis to obtain the conditions achieving higher heavy methylation to G9a, including replicates of the *in vitro* assays with different SAM analogues. This is because G9a *in vitro* assays were only successfully performed on very saturating conditions and with only one replicate due to technical difficulties and time constraints.

Regarding our application of the bump-and-hole approach to G9a, the final outcome (G9a Q1169A paired with a phenyl-SAM analogue) resembles in a way the results obtained for kinases years ago (for example, v-Src kinase V323A, I338A paired with N6-(cyclopentyl) [176], or c-Src kinase T338G paired with N6-(benzyl) ATP [174]). However, we believe that our work represents a step forward from these kinase approaches because of the proposed engineering strategy for designing the KMT-SAM systems, involving an extensive computational assessment (conservation analysis, homology modelling with energy

refinement, molecular dynamics simulations) instead of simply trial and error or extensive screening.

Future work should also focus on finding the best conditions for G9a Q1169A expression and purification to get an active enzyme at appropriate buffers and concentration levels. This would enable the experimental validation of the computational analysis that predicted similar activity levels independently of using native SAM or the C2-modified phenyl-SAM analogue. The novelty of our bump-and-hole approach comes with associated challenges, as protein structure prediction programs are still under development, and their documentation is limited. Nevertheless, our method could indeed accelerate the discovery process as it reduces the amount of experiments needed and it could be readily extended to find other suitable enzyme-cofactor systems.

Importantly, our research also revealed a correlation between the binding mode and binding affinities of SAM cofactor analogues obtained through molecular dynamics simulations with experimental results on the utilization of those SAM cofactor analogues by different KMTs. This could have implications in the computational design of cofactors for target identification, as it implies that molecular dynamics could be used as a screening tool to test the feasibility of enzyme-cofactor pairs prior to experimentation. This work demonstrates that the study of distances, angles and binding affinities can help discerning the compatibility of different SAM analogues towards endogenous KMTs, and it also constitutes the first implementation of molecular dynamics simulations towards the development of a bump-and-hole approach.

Future work should focus on the design and screening of further SAM cofactor analogues starting from current results. In addition, independent MD simulations could be run for each of the systems described in this thesis, each of them with different initial velocities, to confirm the reproducibility of the results. Besides, the proposed phenyl-SAM should be tested experimentally with a properly expressed and purified G9a Q1169A to evaluate if it can be indeed utilised by the enzyme and therefore extend the validation to the use of molecular dynamics simulations for the screening of mutant enzymes - cofactor analogue pairs.

The technology proposed here represents the first epigenetically consistent assay for profiling the targets of KMTs. Small molecule inhibitors of KMTs are currently highly

attractive for their potential roles in reversing aberrant epigenetic silencing as mediated through histone lysine methylation [322]. The proposed design methodology could provide insight into the substrates of KMTs and has the potential to track perturbations in methylation activity and specificity in response to KMT inhibitors. In summary, this method represents a novel engineering strategy to decipher the activity of methyltransferases and holds potential of a direct application on therapeutic research and diagnosis.

Bibliography

1. Waddington, C. H. The epigenotype. *Endeavor* **1,** 18–20 (1942).

2. Riggs, A. D. & Porter, T. N. Overview of epigenetic mechanisms. *Cold Spring Harb. Monogr. Arch. Vol. 32 Epigenetic Mech. Gene Regul.* 29–45 (1996).

3. Berger, S. L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. An operational definition of epigenetics. *Genes Dev.* **23,** 781–3 (2009).

4. Bartke, T., Borgel, J. & DiMaggio, P. A. Proteomics in epigenetics: new perspectives for cancer research. *Brief. Funct. Genomics* **12,** 205–18 (2013).

5. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).

6. Sidoli, S., Cheng, L. & Jensen, O. N. Proteomics in chromatin biology and epigenetics: Elucidation of post-translational modifications of histone proteins by mass spectrometry. *J. Proteomics* **75,** 3419–33 (2012).

7. Campos, E. I. & Reinberg, D. Histones: Annotating Chromatin. *Annu. Rev. Genet.* **43,** 559–599 (2009).

8. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–19 (2007).

9. Jenuwein, T. & Allis, C. D. Translating the Histone Code. *Science* **293,** 1074–1080 (2001).

10. Garcia, B. a, Shabanowitz, J. & Hunt, D. F. Characterization of histones and their post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **11,** 66–73 (2007).

11. Harshman, S. W., Young, N. L., Parthun, M. R. & Freitas, M. A. H1 histones: current perspectives and challenges. *Nucleic Acids Res.* **41,** 9593–609 (2013).

12. Jansen, A. & Verstrepen, K. J. Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* **75,** 301–20 (2011).

13. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* **8,** 286–298 (2007).

14. Leroy, G. *et al.* Proteogenomic characterization and mapping of nucleosomes decoded by Brd and HP1 proteins. *Genome Biol.* **13,** R68 (2012).

15. Clapier, C. R. & Cairns, B. R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78,** 273–304 (2009).

16. van Steensel, B. Chromatin: constructing the big picture. *EMBO J.* **30,** 1885–95 (2011).

17. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief. Bioinform.* **18,** 279–290 (2017).

18. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150,** 12–27 (2012).

19. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196,** 261–

282 (1987).

20. Illingworth, R. S. & Bird, A. P. CpG islands - 'A rough guide'. *FEBS Lett*. **583,** 1713–1720 (2009).

21. Hutter, B., Helms, V. & Paulsen, M. Tandem repeats in the CpG islands of imprinted genes. *Genomics* **88,** 323–332 (2006).

22. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol*. **28,** 1057–1068 (2010).

23. Kawashima, T. & Berger, F. Epigenetic reprogramming in plant sexual reproduction. *Nat. Rev. Genet*. **15,** 613–624 (2014).

24. Yang, X. *et al*. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26,** 577–590 (2014).

25. Meng, H. *et al*. DNA methylation, its mediators and genome integrity. *Int. J. Biol. Sci*. **11,** 604–617 (2015).

26. Du, Q., Luu, P.-L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics* **7,** 1051–73 (2015).

27. Medvedeva, Y. A. *et al*. Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* **15,** 119 (2013).

28. Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim. Biophys. Acta* **1839,** 1362–1372 (2014).

29. Hermann, A., Goyal, R. & Jeltsch, A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J. Biol. Chem*. **279,** 48350–48359 (2004).

30. Goelz, S. E., Vogelstein, B., Hamilton, S. R. & Feinberg, A. P. Hypomethylation of DNA from Benign and Malignant Human Colon Neoplasms. *Science* **228,** 187–190 (1985).

31. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31,** 27–36 (2010).

32. Irizarry, R. A. *et al*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet*. **41,** 178–186 (2009).

33. Iwata, A. *et al*. Altered CpG methylation in sporadic alzheimer's disease is associated with APP and MAPT dysregulation. *Hum. Mol. Genet*. **23,** 648–656 (2014).

34. Mastronardi, F. G., Noor, A., Wood, D. D., Paton, T. & Moscarello, M. A. Peptidyl Argininedeiminase 2 CpG Island in Multiple Sclerosis White Matter Is Hypomethylated. *J. Neurosci. Res*. **85,** 2006–16 (2007).

35. Richardson, B. DNA methylation and autoimmune disease. *Clin. Immunol*. **109,** 72–79 (2003).

36. Tropberger, P. & Schneider, R. Scratching the (lateral) surface of chromatin regulation by histone modifications. *Nat. Struct. Mol. Biol*. **20,** 657–61 (2013).

37. Tan, M. *et al*. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146,** 1016–28 (2011).

38. Zagni, C., Chiacchio, U. & Rescifina, A. Histone methyltransferase inhibitors: novel epigenetic agents for cancer treatment. *Curr. Med. Chem*. **20,** 167–85 (2013).

39. Rando, O. J. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.* **22,** 148–55 (2012).

40. Xiao, B. *et al.* Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature* **421,** 652–6 (2003).

41. Cohen, I., Poręba, E., Kamieniarz, K. & Schneider, R. Histone modifiers in cancer: friends or foes? *Genes Cancer* **2,** 631–47 (2011).

42. Helin, K. & Dhanak, D. Chromatin proteins and modifications as drug targets. *Nature* **502,** 480–8 (2013).

43. Kim, W., Choi, M. & Kim, J.-E. The histone methyltransferase Dot1/DOT1L as a critical regulator of the cell cycle. *Cell Cycle* **13,** 726–38 (2014).

44. Slany, R. K. The molecular biology of mixed lineage leukemia. *Haematologica* **94,** 984–993 (2009).

45. Tibshirani, M. *et al.* Cytoplasmic sequestration of FUS/TLS associated with ALS alters histone marks through loss of nuclear protein arginine methyltransferase 1. *Hum. Mol. Genet.* **24,** 773–786 (2015).

46. Glajch, K. E. & Sadri-Vakili, G. Epigenetic Mechanisms Involved in Huntington ' s Disease Pathogenesis. *J. Huntingtons. Dis.* **4,** 1–15 (2015).

47. Miao, F. *et al.* Lymphocytes From Patients With Type 1 Diabetes Display a Distinct Profile of Chromatin Histone H3 Lysine 9 Dimethylation: an epigenetic study in diabetes. *Diabetes* **57,** 3189–3198 (2008).

48. Ocampo, A. *et al.* In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming. *Cell* **167,** 1719–1733 (2016).

49. Dang, Y., Li, L., Guo, W., Xue, Z. & Liu, Y. Convergent Transcription Induces Dynamic DNA Methylation at disiRNA Loci. *PLoS Genet.* **9,** e1003761 (2013).

50. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133,** 523–536 (2008).

51. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16,** 519–532 (2015).

52. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 487–500 (2016).

53. Son, C.-H. *et al.* Combination treatment with decitabine and ionizing radiation enhances tumor cells susceptibility of T cells. *Sci. Rep.* **6,** 32470 (2016).

54. Jones, P. A., Issa, J.-P. J. & Baylin, S. Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.* **17,** 630–41 (2016).

55. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–57 (2012).

56. Dillon, S. C., Zhang, X., Trievel, R. C. & Cheng, X. The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biol.* **6,** 227 (2005).

57. Di Lorenzo, A. & Bedford, M. T. Histone arginine methylation. *FEBS Lett.* **585,** 2024–31 (2011).

58. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters

and enhancers in the human genome. *Nat. Genet.* **39,** 311–8 (2007).

59. Bannister, A. J. *et al.* Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.* **280,** 17732–6 (2005).

60. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441,** 349–53 (2006).

61. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–37 (2007).

62. Guo, H.-B. & Guo, H. Mechanism of histone methylation catalyzed by protein lysine methyltransferase SET7/9 and origin of product specificity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 8797–8802 (2007).

63. Smith, B. C. & Denu, J. M. Chemical mechanisms of histone lysine and arginine modifications. *Biochim. Biophys. Acta* **1789,** 45–57 (2009).

64. Xiao, B., Wilson, J. R. & Gamblin, S. J. SET domains and histone methylation. *Curr. Opin. Struct. Biol.* **13,** 699–705 (2003).

65. Wu, H. *et al.* Structural biology of human H3K9 methyltransferases. *PLoS One* **5,** e8570 (2010).

66. Del Rizzo, P. A. & Trievel, R. C. Molecular basis for substrate recognition by lysine methyltransferases and demethylases. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1839,** 1404–1415 (2014).

67. Wu, H. *et al.* Histone methyltransferase G9a contributes to H3K27 methylation in vivo. *Cell Res.* **21,** 365–7 (2011).

68. Weiss, T. *et al.* Histone H1 variant-specific lysine methylation by G9a/KMT1C and Glp1/KMT1D. *Epigenetics Chromatin* **3,** 7 (2010).

69. Tachibana, M. *et al.* G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes Dev.* **16,** 1779–91 (2002).

70. Nishida, M. *et al.* Identification of ZNF200 as a novel binding partner of histone H3 methyltransferase G9a. *Genes to Cells* **12,** 877–88 (2007).

71. Cheng, X., Collins, R. E. & Zhang, X. Structural and sequence motifs of protein (histone) methylation enzymes. *Annu. Rev. Biophys. Biomol. Struct.* **34,** 267–94 (2005).

72. Qian, C. & Zhou, M.-M. SET domain protein lysine methyltransferases: Structure, specificity and catalysis. *Cell. Mol. Life Sci.* **63,** 2755–63 (2006).

73. Collins, R. E. *et al.* The ankyrin repeats of G9a and GLP histone methyltransferases are mono- and dimethyllysine binding modules. *Nat. Struct. Mol. Biol.* **15,** 245–50 (2008).

74. Vedadi, M. *et al.* A chemical probe selectively inhibits G9a and GLP methyltransferase activity in cells. *Nat. Chem. Biol.* **7,** 566–74 (2011).

75. Shinkai, Y. & Tachibana, M. H3K9 methyltransferase G9a and the related molecule GLP. *Genes Dev.* **25,** 781–788 (2011).

76. Tachibana, M. *et al.* Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. *Genes Dev.* **19,** 815–26 (2005).

77. Mozzetta, C. *et al.* The histone H3 lysine 9 methyltransferases G9a and GLP regulate polycomb repressive complex 2-mediated gene silencing. *Mol. Cell* **53,** 277–89 (2014).

78. Fritsch, L. *et al*. A subset of the histone H3 lysine 9 methyltransferases Suv39h1, G9a, GLP, and SETDB1 participate in a multimeric complex. *Mol. Cell* **37,** 46–56 (2010).

79. Fang, S. *et al*. Coordinated recruitment of histone methyltransferase G9a and other chromatin-modifying enzymes in SHP-mediated regulation of hepatic bile acid metabolism. *Mol. Cell. Biol.* **27,** 1407–24 (2007).

80. Rathert, P. *et al*. Protein lysine methyltransferase G9a acts on non-histone targets. *Nat. Chem. Biol*. **4,** 344–6 (2008).

81. Barsyte-Lovejoy, D. *et al*. (R)-PFI-2 is a potent and selective inhibitor of SETD7 methyltransferase activity in cells. *Proc. Natl. Acad. Sci.* **111,** 12853–12858 (2014).

82. Schapira, M. Structural Chemistry of Human SET Domain Protein Methyltransferases. *Curr. Chem. Genomics* **5,** 85–94 (2011).

83. Scheufler, C. *et al*. Optimization of a Fragment-Based Screening Hit toward Potent DOT1L Inhibitors Interacting in an Induced Binding Pocket. *ACS Med. Chem. Lett*. **7,** 730–4 (2016).

84. Min, J., Feng, Q., Li, Z., Zhang, Y. & Xu, R. M. Structure of the catalytic domain of human Dot1L, a non-SET domain nucleosomal histone methyltransferase. *Cell* **112,** 711–723 (2003).

85. Lu, D. Epigenetic modification enzymes: catalytic mechanisms and inhibitors. *Acta Pharm. Sin. B* **3,** 141–149 (2013).

86. Yu, W. *et al*. Bromo-deaza-SAH: A potent and selective DOT1L inhibitor. *Bioorg. Med. Chem.* **21,** 1787–1794 (2013).

87. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447,** 407–412 (2007).

88. Allis, C. D. *et al*. New Nomenclature for Chromatin-Modifying Enzymes. *Cell* **131,** 633–636 (2007).

89. Morera, L., Lübbert, M. & Jung, M. Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy. *Clin. Epigenetics* **8,** 57 (2016).

90. Zhang, X., Huang, Y. & Shi, X. Emerging roles of lysine methylation on non-histone proteins. *Cell. Mol. Life Sci*. **72,** 4257–4272 (2015).

91. Beltrao, P., Bork, P., Krogan, N. J. & Van Noort, V. Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol*. **9,** 1–13 (2013).

92. Valekunja, U. K. *et al*. Histone methyltransferase MLL3 contributes to genome-scale circadian transcription. *Proc. Natl. Acad. Sci. U. S. A*. **110,** 1554–9 (2013).

93. Biggar, K. K. & Li, S. S.-C. Non-histone protein methylation as a regulator of cellular signalling and function. *Nat. Rev. Mol. Cell Biol*. **16,** 5–17 (2015).

94. Wu, Z., Connolly, J. & Biggar, K. K. Beyond histones - the expanding roles of protein lysine methylation. *FEBS J*. 14056 (2017).

95. West, L. & Gozani, O. Regulation of p53 function by lysine methylation. *Epigenomics* **3,** 361–369 (2011).

96. Huang, J. *et al*. Repression of p53 activity by Smyd2-mediated methylation. *Nature* **444,** 629–632 (2006).

97. Roy, S. *et al*. Structural Insight into p53 Recognition by the 53BP1 Tandem Tudor Domain. *J.*

Mol. Biol. **398,** 489–496 (2010).

98. Pascal, R., Pross, A. & Sutherland, J. D. Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol.* **3,** 130156 (2013).

99. Oudhoff, M. J. *et al.* Control of the Hippo Pathway by Set7-Dependent Methylation of Yap. *Dev. Cell* **26,** 188–194 (2013).

100. Yang, J. *et al.* Reversible methylation of promoter-bound STAT3 by histone-modifying enzymes. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21499–504 (2010).

101. Andreu-Perez, P. *et al.* Protein arginine methyltransferase 5 regulates ERK1/2 signal transduction amplitude and cell fate through CRAF. *Sci. Signal.* **4,** ra58 (2011).

102. Bikkavilli, R. K. & Malbon, C. C. Wnt3a-stimulated LRP6 phosphorylation is dependent upon arginine methylation of G3BP2. *J. Cell Sci.* **125,** 2446–56 (2012).

103. Xu, J. & Derynck, R. Does Smad6 methylation control BMP signaling in cancer? *Cell Cycle* **13,** 1209–1210 (2014).

104. Chin, H. G. *et al.* Automethylation of G9a and its implication in wider substrate specificity and HP1 binding. *Nucleic Acids Res.* **35,** 7313–7323 (2007).

105. Yu, Y. *et al.* Histone H3 lysine 56 methylation regulates DNA replication through its interaction with PCNA. *Mol. Cell* **46,** 7–17 (2012).

106. Trojer, P. *et al.* Dynamic Histone H1 Isotype 4 Methylation and Demethylation by Histone Lysine Methyltransferase G9a/KMT1C and the Jumonji Domain-containing JMJD2/KDM4 Proteins. *J. Biol. Chem.* **284,** 8395–405 (2009).

107. Sampath, S. C. *et al.* Methylation of a histone mimic within the histone methyltransferase G9a regulates protein complex assembly. *Mol. Cell* **27,** 596–608 (2007).

108. Martens, J. H. A., Verlaan, M., Kalkhoven, E. & Zantema, A. Cascade of distinct histone modifications during collagenase gene activation. *Mol. Cell. Biol.* **23,** 1808–16 (2003).

109. Francis, J., Chakrabarti, S. K., Garmey, J. C. & Mirmira, R. G. Pdx-1 links histone H3-Lys-4 methylation to RNA polymerase II elongation during activation of insulin transcription. *J. Biol. Chem.* **280,** 36244–36253 (2005).

110. Chuikov, S. *et al.* Regulation of p53 activity through lysine methylation. *Nature* **432,** 353–360 (2004).

111. Kouskouti, A., Scheer, E., Staub, A., Tora, L. & Talianidis, I. Gene-specific modulation of TAF10 function by SET9-mediated methylation. *Mol. Cell* **14,** 175–182 (2004).

112. Pradhan, S., Chin, H. G., Estève, P. O. & Jacobsen, S. E. SET7/9 mediated methylation of non-histone proteins in mammalian cells. *Epigenetics* **4,** 383–387 (2009).

113. Kassner, I., Barandun, M., Fey, M., Rosenthal, F. & Hottiger, M. O. Crosstalk between SET7/9-dependent methylation and ARTD1-mediated ADP-ribosylation of histone H1.4. *Epigenetics Chromatin* **6,** 1 (2013).

114. Nguyen, A. T. & Zhang, Y. The diverse functions of Dot1 and H3K79 methylation. *Genes Dev.* **25,** 1345–1358 (2011).

115. Feng, Q. *et al.* Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr. Biol.* **12,** 1052–1058 (2002).

116. Jones, B. *et al*. The histone H3K79 methyltransferase Dot1L is essential for mammalian development and heterochromatin structure. *PLoS Genet*. **4,** e1000190 (2008).

117. Frederiks, F. *et al*. Nonprocessive methylation by Dot1 leads to functional redundancy of histone H3K79 methylation states. *Nat. Struct. Mol. Biol*. **15,** 550–7 (2008).

118. Wong, M., Polly, P. & Liu, T. The histone methyltransferase DOT1L: regulatory functions and a cancer therapy target. *Am. J. Cancer Res*. **5,** 2823–2837 (2015).

119. Okada, Y. *et al*. hDOT1L links histone methylation to leukemogenesis. *Cell* **121,** 167–178 (2005).

120. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153,** 1194–1217 (2013).

121. De Vos, D. *et al*. Progressive methylation of ageing histones by Dot1 functions as a timer. *EMBO Rep*. **12,** 956–962 (2011).

122. Soria-Valles, C. *et al*. NF-[kappa]B activation impairs somatic cell reprogramming in ageing. *Nat. Cell Biol*. **17,** 1004–1013 (2015).

123. Soria-Valles, C., Osorio, F. G. & Lopez-Otin, C. Reprogramming aging through DOT1L inhibition. *Cell Cycle* **14,** 3345–3346 (2015).

124. Luo, M. *et al*. Identification of phenoxyacetamide derivatives as novel DOT1L inhibitors via docking screening and molecular dynamics simulation. *J. Mol. Graph. Model*. **68,** 128–139 (2016).

125. Horiuchi, K. Y. *et al*. Assay Development for Histone Methyltransferases. *Assay Drug Dev. Technol*. **11,** 227–236 (2013).

126. Moore, K. E. & Gozani, O. An unexpected journey: Lysine methylation across the proteome. *Biochim. Biophys. Acta - Gene Regul. Mech*. **1839,** 1395–1403 (2014).

127. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).

128. Wasinger, V. C. *et al*. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **16,** 1090–4 (1995).

129. Burkard, T. R. *et al*. Initial characterization of the human central proteome. *BMC Syst. Biol*. **5,** 1–13 (2011).

130. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312,** 212–7 (2006).

131. Wilkins, M. Proteomics data mining. *Expert Rev. Proteomics* **6,** 599–603 (2009).

132. Altelaar, a F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet*. **14,** 35–48 (2013).

133. Frenkel-Morgenstern, M. *et al*. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res*. **22,** 1231–42 (2012).

134. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–8 (2011).

135. González-Díaz, H., González-Díaz, Y., Santana, L., Ubeira, F. M. & Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **8,** 750–78 (2008).

136. Grammel, M. & Hang, H. C. Chemical reporters for biological discovery. *Nat. Chem. Biol*. **9,**

475–84 (2013).

137. Wang, R. *et al*. Profiling Genome-Wide Chromatin Methylation with Engineered Posttranslation Apparatus within Living Cells. *J. Am. Chem. Soc*. **135,** 1048–1056 (2013).

138. Zee, B. M., Levin, R. S., DiMaggio, P. A. & Garcia, B. A. Global turnover of histone post-translational modifications and variants in human cells. *Epigenetics Chromatin* **3,** 22 (2010).

139. Ibáñez, G. *et al*. A High Throughput Scintillation Proximity Imaging Assay for Protein Methyltransferases. *Comb. Chem. High Throughput Screen*. **15,** 359–371 (2012).

140. Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. *ACS Chem. Biol*. **7,** 443–63 (2012).

141. Dhayalan, A., Dimitrova, E., Rathert, P. & Jeltsch, A. A continuous protein methyltransferase (G9a) assay for enzyme activity measurement and inhibitor screening. *J. Biomol. Screen*. **14,** 1129–33 (2009).

142. Suh-Lailam, B. B. & Hevel, J. M. A fast and efficient method for quantitative measurement of S-adenosyl-L-methionine-dependent methyltransferase activity with protein substrates. *Anal. Biochem*. **398,** 218–24 (2010).

143. Wu, J., Xie, N., Feng, Y. & Zheng, Y. G. Scintillation proximity assay of arginine methylation. *J. Biomol. Screen*. **17,** 237–44 (2012).

144. Gauthier, N. *et al*. Development of homogeneous nonradioactive methyltransferase and demethylase assays targeting histone H3 lysine 4. *J. Biomol. Screen*. **17,** 49–58 (2012).

145. Quinn, A. M., Allali-Hassani, A., Vedadi, M. & Simeonov, A. A Chemiluminescence-based Method for Identification of Histone Lysine Methyltransferase Inhibitors. *Mol. Biosyst*. **6,** 782–788 (2010).

146. Bartke, T. *et al*. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143,** 470–84 (2010).

147. Moradian, A., Kalli, A., Sweredoski, M. J. & Hess, S. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. *Proteomics* **14,** 489–97 (2014).

148. Eliuk, S. M., Maltby, D., Panning, B. & Burlingame, A. L. High resolution electron transfer dissociation studies of unfractionated intact histones from murine embryonic stem cells using on-line capillary LC separation: determination of abundant histone isoforms and post-translational modifications. *Mol. Cell. Proteomics* **9,** 824–37 (2010).

149. Pesavento, J. J., Yang, H., Kelleher, N. L. & Mizzen, C. a. Certain and progressive methylation of histone H4 at lysine 20 during the cell cycle. *Mol. Cell. Biol*. **28,** 468–86 (2008).

150. Siuti, N. & Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **4,** 817–821 (2007).

151. Meyer, B., Papasotiriou, D. G. & Karas, M. 100% Protein Sequence Coverage: a Modern Form of Surrealism in Proteomics. *Amino Acids* **41,** 291–310 (2011).

152. Tian, Z. *et al*. Enhanced top-down characterization of histone post-translational modifications. *Genome Biol*. **13,** R86 (2012).

153. Kalli, A. & Håkansson, K. Electron capture dissociation of highly charged proteolytic peptides from Lys N, Lys C and Glu C digestion. *Mol. Biosyst*. **6,** 1668–81 (2010).

154. Garcia, B. A. What does the future hold for Top Down mass spectrometry? *J. Am. Soc. Mass Spectrom.* **21,** 193–202 (2010).

155. Arnaudo, A. M., Molden, R. C. & Garcia, B. A. Revealing histone variant induced changes via quantitative proteomics. *Crit. Rev. Biochem. Mol. Biol.* **46,** 284–294 (2011).

156. Syka, J. E. P. *et al.* Novel linear quadrupole ion trap/FT mass spectrometer: Performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J. Proteome Res.* **3,** 621–626 (2004).

157. Sidoli, S., Bhanu, N. V, Karch, K. R., Wang, X. & Garcia, B. A. Complete Workflow for Analysis of Histone Post-translational Modifications Using Bottom-up Mass Spectrometry: From Histone Extraction to Data Analysis. *J. Vis. Exp.* e54112 (2016).

158. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389,** 1017–31 (2007).

159. Knapp, A. R. *et al.* Quantitative profiling of histone post-translational modifications by stable isotope labeling. *Methods* **41,** 312–9 (2007).

160. Arnaudo, A. M. & Garcia, B. a. Proteomic characterization of novel histone post-translational modifications. *Epigenetics Chromatin* **6,** 1–7 (2013).

161. Li, S. *et al.* A liquid chromatography/mass spectrometry-based generic detection method for biochemical assay and hit discovery of histone methyltransferases. *Anal. Biochem.* **443,** 214–21 (2013).

162. Lakowski, T. M. & Frankel, A. Sources of S-adenosyl-L-homocysteine background in measuring protein arginine N-methyltransferase activity using tandem mass spectrometry. *Anal. Biochem.* **396,** 158–60 (2010).

163. Graves, T. L., Zhang, Y. & Scott, J. E. A universal competitive fluorescence polarization activity assay for S-adenosylmethionine utilizing methyltransferases. *Anal. Biochem.* **373,** 296–306 (2008).

164. Wang, C., Leffler, S., Thompson, D. H. & Hrycyna, C. A. A general fluorescence-based coupled assay for S-adenosylmethionine-dependent methyltransferases. *Biochem. Biophys. Res. Commun.* **331,** 351–6 (2005).

165. Ibáñez, G., McBean, J. L., Astudillo, Y. M. & Luo, M. An enzyme-coupled ultrasensitive luminescence assay for protein methyltransferases. *Anal. Biochem.* **401,** 203–10 (2010).

166. Wang, R. *et al.* Formulating a fluorogenic assay to evaluate S-adenosyl-L-methionine analogues as protein methyltransferase cofactors. *Mol. Biosyst.* **7,** 2970–81 (2011).

167. Lin, Q., Jiang, F., Schultz, P. G. & Gray, N. S. Design of Allele-Specific Protein Methyltransferase Inhibitors. *J. Am. Chem. Soc.* **123,** 11608–13 (2001).

168. Arrowsmith, C. H. *et al.* The promise and peril of chemical probes. *Nat. Chem. Biol.* **11,** 536–541 (2015).

169. Islam, K., Zheng, W., Yu, H., Deng, H. & Luo, M. Expanding Cofactor Repertoire of Protein Lysine Methyltransferase for Substrate Labeling. *ACS Chem. Biol.* **6,** 679–684 (2011).

170. Baud, M. G. J. *et al.* A bump-and-hole approach to engineer controlled selectivity of BET bromodomain chemical probes. *Science* **346,** 638–41 (2014).

171. Urich, R. *et al.* De novo design of protein kinase inhibitors by in silico identification of hinge region-binding fragments. *ACS Chem. Biol.* **8,** 1044–1052 (2013).

172. Liao, J. J. Molecular Recognition of Protein Kinase Binding Pockets for Design of Potent and Selective Kinase Inhibitors. *J. Med. Chem.* **50,** 409–24 (2007).

173. Fang, Z., Grütter, C. & Rauh, D. Strategies for the selective regulation of kinases with allosteric modulators: exploiting exclusive structural features. *ACS Chem. Biol.* **8,** 58–70 (2013).

174. Witucki, L. A. *et al.* Mutant Tyrosine Kinases with Unnatural Nucleotide Specificity Retain the Structure and Phospho-Acceptor Specificity of the Wild-Type Enzyme. *Chem. Biol.* **9,** 25–33 (2002).

175. Allen, J. J. *et al.* A semisynthetic epitope for kinase substrates. *Nat. Methods* **4,** 511–516 (2007).

176. Shah, K., Liu, Y., Deirmengian, C. & Shokat, K. M. Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proc. Natl. Acad. Sci.* **94,** 3565–3570 (1997).

177. Li, J., Wei, H. & Zhou, M.-M. Structure-guided design of a methyl donor cofactor that controls a viral histone H3 lysine 27 methyltransferase activity. *J. Med. Chem.* **54,** 7734–8 (2011).

178. Yao, Y. *et al.* Selective inhibitors of histone methyltransferase DOT1L: design, synthesis, and crystallographic studies. *J. Am. Chem. Soc.* **133,** 16746–9 (2011).

179. Osborne, T., Roska, R. L. W., Rajski, S. R. & Thompson, P. R. In situ generation of a bisubstrate analogue for protein arginine methyltransferase 1. *J. Am. Chem. Soc.* **130,** 4574–5 (2008).

180. Klimasauskas, S. & Weinhold, E. A new tool for biotechnology: AdoMet-dependent methyltransferases. *Trends Biotechnol.* **25,** 99–104 (2007).

181. Wang, R., Zheng, W. & Luo, M. A sensitive mass spectrum assay to characterize engineered methionine adenosyltransferases with S-alkyl methionine analogues as substrates. *Anal. Biochem.* **450,** 11–19 (2014).

182. Thomsen, M., Vogensen, S. B., Buchardt, J., Burkart, M. D. & Clausen, R. P. Chemoenzymatic synthesis and in situ application of S-adenosyl-L-methionine analogs. *Org. Biomol. Chem.* **11,** 7606–10 (2013).

183. Wang, R., Zheng, W., Yu, H., Deng, H. & Luo, M. Labeling Substrates of Protein Arginine Methyltransferase with Engineered Enzymes and Matched S-Adenosyl-L-methionine Analogues. *J. Am. Chem. Soc.* **133,** 7648–7651 (2011).

184. Slade, D. J., Subramanian, V., Fuhrmann, J. & Thompson, P. R. Chemical and biological methods to detect posttranslational modifications of arginine. *Biopolymers* **101,** 133–143 (2013).

185. Bothwell, I. R. *et al.* Se-Adenosyl-L-selenomethionine Cofactor Cofactor Analogue as a Reporter of Protein Methylation. *J. Am. Chem. Soc.* **134,** 14905–14912 (2012).

186. €2.7m for lung cancer biomarker research. *Pan European Networks* (2015). Available at: http://www.paneuropeannetworks.com/health/e2-7m-for-lung-cancer-biomarker-research/. (Accessed: 19th June 2017)

187. Heerboth, S. *et al.* Use of epigenetic drugs in disease: an overview. *Genet. Epigenetics* **6,** 9–19 (2014).

188. Rao, D. D., Vorhies, J. S., Senzer, N. & Nemunaitis, J. siRNA vs. shRNA: Similarities and differences. *Adv. Drug Deliv. Rev.* **61,** 746–759 (2009).

189. O'Connor, C. M., DiMaggio, P. A., Shenk, T. & Garcia, B. A. Quantitative Proteomic Discovery of Dynamic Epigenome Changes that Control Human Cytomegalovirus (HCMV) Infection. *Mol. Cell. Proteomics* **13,** 2399–2410 (2014).

190. Alemán, L. M., Doench, J. & Sharp, P. A. Comparison of siRNA-induced off-target RNA and protein effects. *RNA* **13,** 385–95 (2007).

191. Sandy, P., Ventura, A. & Jacks, T. Mammalian RNAi: A practical guide. *Biotechniques* **39,** 215–224 (2005).

192. Huang, H., Lin, S., Garcia, B. A. & Zhao, Y. Quantitative proteomic analysis of histone modifications. *Chem. Rev.* **115,** 2376–2418 (2015).

193. Ong, S.-E., Mittler, G. & Mann, M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods* **1,** 119–126 (2004).

194. Soldi, M., Cuomo, A., Bremang, M. & Bonaldi, T. Mass spectrometry-based proteomics for the analysis of chromatin structure and dynamics. *Int. J. Mol. Sci.* **14,** 5402–5431 (2013).

195. Bremang, M. *et al.* Mass spectrometry-based identification and characterisation of lysine and arginine methylation in the human proteome. *Mol. Biosyst.* **9,** 2231–47 (2013).

196. Fodor, B. D. *et al.* Jmjd2b antagonizes H3K9 trimethylation at pericentric heterochromatin in mammalian cells. *Genes Dev.* **20,** 1557–1562 (2006).

197. Zee, B. M. *et al.* In vivo residue-specific histone methylation dynamics. *J. Biol. Chem.* **285,** 3341–3350 (2010).

198. Rothbart, S. B. *et al.* An Interactive Database for the Assessment of Histone Antibody Specificity. *Mol. Cell* **59,** 502–511 (2015).

199. Olsen, J. B. *et al.* Quantitative Profiling of the Activity of Protein Lysine Methyltransferase SMYD2 Using SILAC-Based Proteomics. *Mol. Cell. Proteomics* **15,** 892–905 (2016).

200. Horning, B. D. *et al.* Chemical Proteomic Profiling of Human Methyltransferases. *J. Am. Chem. Soc.* **138,** 13335–13343 (2016).

201. Zhu, B. *et al.* In Situ Proteome Profiling and Bioimaging Applications of Small-Molecule Affinity-Based Probes Derived from DOT1L Inhibitors. *Chem. - A Eur. J.* **22,** 7824–7836 (2016).

202. Ahmed, H., Duan, S., Arrowsmith, C. H., Barsyte-Lovejoy, D. & Schapira, M. An integrative proteomic approach identifies novel cellular SMYD2 substrates. *J. Proteome Res.* **15,** 2052–2059 (2016).

203. Guo, A. *et al.* Immunoaffinity Enrichment and Mass Spectrometry Analysis of Protein Methylation. *Mol. Cell. Proteomics* **13,** 372–387 (2014).

204. Loriot, S., Cazals, F. & Bernauer, J. ESBTL: Efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules. *Bioinformatics* **26,** 1127–1128 (2010).

205. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290,** 253–66 (1999).

206. Kelley, K. D., Olive, L. Q., Hadziselimovic, A. & Sanders, C. R. Look and see if it is time to induce protein expression in escherichia coli cultures. *Biochemistry* **49,** 5405–5407 (2010).

207. Lin, S. & Garcia, B. A. *Examining histone posttranslational modification patterns by high-resolution mass spectrometry. Methods in Enzymology* **512,** (Elsevier Inc., 2012).

208. Trainor, N. Profiling the methylome targets of the histone lysine methyltransferase DOT1L. (Imperial College London, 2017).

209. Gasteiger, E. *et al*. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005).

210. Miyaura, N. & Suzuki, A. Palladium-Catalyzed Cross-Coupling Reactions of Organoboron Compounds. *Chem. Rev.* **95,** 2457–2483 (1995).

211. Fischle, W. & Schwarzer, D. Probing Chromatin-modifying Enzymes with Chemical Tools. *ACS Chem. Biol.* **11,** 689–705 (2016).

212. Wang, R. & Luo, M. A journey toward Bioorthogonal Profiling of Protein Methylation inside living cells. *Curr. Opin. Chem. Biol.* **17,** 729–37 (2013).

213. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9,** 71 (2011).

214. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **66,** 334–395 (2013).

215. Ou-Yang, S.-S. *et al*. Computational drug discovery. *Acta Pharmacol. Sin.* **33,** 1131–40 (2012).

216. Schrödinger. The PyMOL Molecular Graphics System, Version 1.8. (2015).

217. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383,** 66–93 (2004).

218. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59,** 4035–4061 (2016).

219. Trott, O. & Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **31,** 455–461 (2010).

220. Samsudin, F., Parker, J. L., Sansom, M. S. P., Newstead, S. & Fowler, P. W. Accurate Prediction of Ligand Affinities for a Proton-Dependent Oligopeptide Transporter. *Cell Chem. Biol.* **23,** 299–309 (2016).

221. Ponder, J. W. & Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **66,** 27–85 (2003).

222. González, M. A. Force fields and molecular dynamics simulations. *Collect. SFN* **12,** 169–200 (2011).

223. Dror, R. O., Jensen, M. Ø., Borhani, D. W. & Shaw, D. E. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J. Gen. Physiol.* **135,** 555–562 (2010).

224. Grubmüller, H., Heller, H., Windemuth, A. & Schulten, K. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions. *Mol. Simul.* **6,** 121–142 (1991).

225. Pearlman, D. A. *et al*. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91,** 1–41 (1995).

226. Steinbrecher, T., Case, D. A. & Labahn, A. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. *J. Med. Chem.* **49,** 1837–1844 (2006).

227. Tzoupis, H., Leonis, G., Mavromoustakos, T. & Papadopoulos, M. G. A comparative molecular dynamics, MM-PBSA and thermodynamic integration study of saquinavir complexes with wild-

type HIV-1 PR and L10I, G48V, L63P, A71V, G73S, V82A and I84V single mutants. *J. Chem. Theory Comput.* **9,** 1754–1764 (2013).

228. Chipot, C. in *New Algorithms for Macromolecular Simulation* 185–211 (2006).

229. Shirts, M. R., Mobley, D. L. & Brown, S. P. in *Drug Design: Structure- and Ligand-Based Approaches* 61–86 (2010).

230. Zwanzig, R. W. High Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **22,** 1420–1426 (1954).

231. Deng, Y. & Roux, B. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **113,** 2234–2246 (2009).

232. Gilson, M. K. & Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36,** 21–42 (2007).

233. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **30,** 2785–2791 (2009).

234. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267,** 727–48 (1997).

235. Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **39,** W270–W277 (2011).

236. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).

237. Spitzer, R. & Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput. Aided. Mol. Des.* **26,** 687–99 (2012).

238. Venkatachalam, C. M., Jiang, X., Oldfield, T. & Waldman, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **21,** 289–307 (2003).

239. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261,** 470–89 (1996).

240. Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B. & Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graph. Model.* **26,** 198–212 (2007).

241. Velec, H. F. G., Gohlke, H. & Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **48,** 6296–6303 (2005).

242. Todorov, N. P., Monthoux, P. H. & Alberts, I. L. The influence of variations of ligand protonation and tautomerism on protein-ligand recognition and binding energy landscape. *J. Chem. Inf. Model.* **46,** 1134–1142 (2006).

243. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **52,** 1757–1768 (2012).

244. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55,** 2324–2337 (2015).

245. Plewczynski, D., Łaźniewski, M., Augustyniak, R. & Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **32,** 742–755 (2011).

246. Aqvist, J., Medina, C. & Samuelsson, J.-E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng*. **7,** 385–391 (1994).

247. Wang, J., Dixon, R. & Kollman, P. A. Ranking ligand binding affinities with avidin: a molecular dynamics-based interaction energy study. *Proteins* **34,** 69–81 (1999).

248. Miller, B. R. *et al*. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput*. **8,** 3314–3321 (2012).

249. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the Performance of the MM / PBSA and MM / GBSA Methods . 1 . The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model*. **51,** 69–82 (2011).

250. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov*. **10,** 449–61 (2015).

251. Genheden, S. & Ryde, U. Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins* **80,** 1326–1342 (2012).

252. Singh, N. & Warshel, A. Absolute binding free energy calculations: On the accuracy of computational scoring of protein-ligand interactions. *Proteins* **78,** 1705–1723 (2010).

253. Fratev, F., Tsakovska, I., Al Sharif, M., Mihaylova, E. & Pajeva, I. Structural and dynamical insight into PPARγ antagonism: In Silico study of the ligand-receptor interactions of non-covalent Antagonists. *Int. J. Mol. Sci*. **16,** 15405–15424 (2015).

254. Gouda, H., Kuntz, I. D., Case, D. A. & Kollman, P. A. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers* **68,** 16–34 (2003).

255. Kuhn, B. & Kollman, P. A. QM-FE and molecular dynamics calculations on catechol O-methyltransferase: Free energy of activation in the enzyme and in aqueous solution and regioselectivity of the enzyme-catalyzed reaction. *J. Am. Chem. Soc*. **122,** 2586–2596 (2000).

256. Brooks, B. R., Brooks III, C. L. & Mackerell, A. D. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem*. **30,** 1545–1614 (2009).

257. Hu, P., Wang, S. & Zhang, Y. How do SET-domain protein lysine methyltransferases achieve the methylation state specificity? Revisited by Ab initio QM/MM molecular dynamics simulations. *J. Am. Chem. Soc*. **130,** 3806–3813 (2008).

258. Frisch, M. J. *et al*. Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford CT. (2004).

259. Shao, Y. *et al*. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys*. **8,** 3172–3191 (2006).

260. Yao, J., Chu, Y., An, R. & Guo, H. Understanding product specificity of protein lysine methyltransferases from QM/MM molecular dynamics and free energy simulations: The effects of mutation on SET7/9 beyond the Tyr/Phe switch. *J. Chem. Inf. Model*. **52,** 449–456 (2012).

261. Chu, Y., Yao, J. & Guo, H. QM/MM MD and free energy simulations of G9a-like protein (GLP) and its mutants: Understanding the factors that determine the product specificity. *PLoS One* **7,** e37674 (2012).

262. Kalinic, M., Zloh, M. & Eric, S. Structural insights into binding of small molecule inhibitors to Enhancer of Zeste Homolog 2. *J. Comput. Aided. Mol. Des*. **28,** 1109–1128 (2014).

263. Case, D. A. *et al*. AMBER 12, University of California, San Francisco. (2012).

264. Hong, W. *et al*. Investigating the binding preferences of small molecule inhibitors of human protein arginine methyltransferase 1 using molecular modelling. *J. Mol. Graph. Model*. **51,** 193–202 (2014).

265. Gathiaka, S., Boykin, B., Caceres, T., Hevel, J. M. & Acevedo, O. Understanding protein arginine methyltransferase 1 (PRMT1) product specificity from molecular dynamics. *Bioorganic Med. Chem*. **24,** 4949–4960 (2016).

266. Abraham, M. J. *et al*. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2,** 19–25 (2015).

267. Kumari, R., Kumar, R. & Lynn, A. g_mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model*. **54,** 1951–1962 (2014).

268. Chang, Y. *et al*. Structural basis for G9a-like protein lysine methyltransferase inhibition by BIX-01294. *Nat. Struct. Mol. Biol*. **16,** 312–7 (2009).

269. Case, D. A. *et al*. The Amber biomolecular simulation programs. *J. Comput. Chem*. **26,** 1668–1688 (2005).

270. Case, D. A. D. A. *et al*. *AMBER 2016 Reference Manual*. (2016).

271. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. a & Case, D. a. Development and testing of a general amber force field. *J. Comput. Chem*. **25,** 1157–1174 (2004).

272. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci*. **3,** 198–210 (2013).

273. Pronk, S. *et al*. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29,** 845–854 (2013).

274. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Antechamber: an accessory software package for molecular mechanical calculations. *Abstr. Pap. Am. Chem. Soc*. **222,** U403 (2001).

275. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model*. **56,** 599–604 (2016).

276. Roe, D. R. & Cheatham III, T. E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular synamics trajectory data. *J. Chem. Theory Comput*. **9,** 3084–3095 (2013).

277. ChemAxon. MarvinSketch. (2017).

278. Pettersen, E. F. *et al*. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem*. **25,** 1605–1612 (2004).

279. Anandakrishnan, R., Aguilar, B. & Onufriev, A. V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res*. **40,** W537–W541 (2012).

280. Frisch, M. J. *et al*. Gaussian 09, Revision A.02. (2016).

281. Pace, N. J. & Weerapana, E. Zinc-binding cysteines: diverse functions and structural motifs. *Biomolecules* **4,** 419–434 (2014).

282. Lin, F. & Wang, R. Systematic derivation of AMBER force field parameters applicable to zinc-containing systems. *J. Chem. Theory Comput*. **6,** 1852–1870 (2010).

283. Peters, M. B. *et al*. Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *J. Chem. Theory Comput*. **6,** 2935–2947 (2010).

284. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **11,** 1864–1874 (2015).

285. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).

286. Waskom, M. *et al.* seaborn: v0.7.0 (January 2016). (2016).

287. Turner, P. J. XMGRACE, Version 5.1.19. *Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR* (2005).

288. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. Ligplot - a Program To Generate Schematic Diagrams of Protein Ligand Interactions. *Protein Eng.* **8,** 127–134 (1995).

289. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410,** 116–20 (2001).

290. Casciello, F., Windloch, K., Gannon, F. & Lee, J. S. Functional role of G9a histone methyltransferase in cancer. *Front. Immunol.* **6,** 487 (2015).

291. Ke, X. X. *et al.* Inhibition of H3K9 methyltransferase G9A repressed cell proliferation and induced autophagy in neuroblastoma. *PLoS One* **9,** e106962 (2014).

292. Liu, F. *et al.* Discovery of an in vivo chemical probe of the lysine methyltransferases G9a and GLP. *J. Med. Chem.* **56,** 8931–8942 (2013).

293. Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H. & Shinkai, Y. G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. *EMBO J.* **27,** 2681–2690 (2008).

294. Huang, J. *et al.* G9a and Glp methylate lysine 373 in the tumor suppressor p53. *J. Biol. Chem.* **285,** 9636–9641 (2010).

295. Lee, J. S. *et al.* Negative Regulation of Hypoxic Responses via Induced Reptin Methylation. *Mol. Cell* **39,** 71–85 (2010).

296. Lee, J. S. *et al.* Hypoxia-induced methylation of a pontin chromatin remodeling factor. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 13510–13515 (2011).

297. Jin, W. *et al.* Involvement of MyoD and c-myb in regulation of basal and estrogen-induced transcription activity of the BRCA1 gene. *Breast Cancer Res. Treat.* **125,** 699–713 (2011).

298. Islam, K. *et al.* Defining efficient enzyme-cofactor pairs for bioorthogonal profiling of protein methylation. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 16778–83 (2013).

299. Baud, M. G. J. *et al.* A bump-and-hole approach to engineer controlled selectivity of BET bromodomain chemical probes. *Science* **346,** 638–641 (2014).

300. Das, R. & Baker, D. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77,** 363–82 (2008).

301. Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77 Suppl 9,** 89–99 (2009).

302. Ben-David, M. *et al.* Assessment of CASP8 structure predictions for template free targets. *Proteins* **77 Suppl 9,** 50–65 (2009).

303. Kinch, L. *et al.* CASP9 assessment of free modeling target predictions. *Proteins* **79 Suppl 1,** 59–73 (2011).

304. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins* **82 Suppl 2,** 1–6 (2014).

305. Leaver-Fay, A. *et al*. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*. **487,** 545–74 (2011).

306. Dill, K. a & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338,** 1042–6 (2012).

307. J. Rigden, D. *From Protein Structure to Function with Bioinformatics*. (Springer, 2017).

308. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*. **33,** W244-8 (2005).

309. Rohl, C. a, Strauss, C. E. M., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins* **55,** 656–77 (2004).

310. Canutescu, A. A. & Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci*. **12,** 963–972 (2003).

311. Barth, P., Wallner, B. & Baker, D. Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 1409–14 (2009).

312. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* **49,** 2987–98 (2010).

313. Morozov, A. V, Kortemme, T., Tsemekhman, K. & Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6946–51 (2004).

314. Young, A. R. J., Narita, M. & Narita, M. *Homology Modeling. Methods in Molecular Biology* **857,** (2012).

315. Gray, J. J. *et al*. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol*. **331,** 281–299 (2003).

316. Xu, D., Tsai, C. J. & Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*. **10,** 999–1012 (1997).

317. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol*. **7,** 539 (2011).

318. O'Shea, J. P. *et al*. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods* **10,** 1211–2 (2013).

319. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol*. **22,** 1035–1036 (2004).

320. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*. **32,** W526-31 (2004).

321. Zhang, Z. & Schreiner, P. R. (Thio)urea organocatalysis—What can be learnt from anion recognition? *Chem. Soc. Rev*. **38,** 1187–1198 (2009).

322. Spannhoff, A., Sippl, W. & Jung, M. Cancer treatment of the future: Inhibitors of histone methyltransferases. *The International Journal of Biochemistry and Cell Biology* **41,** 4–11 (2009).

Appendices

```
Minimise complex with 2 kcal/mol restraints in 10000 steps
 &cntrl
  imin=1 !Perform an energy minimization.
  maxcyc=10000 !Maximum number of cycles of minimization.
  ncyc=3000, !Switch from steepest descent to conjugate gradient after these cycles.
  cut=12.0, !Non-bonded cutoff in angstroms.
  ntb=1, !Constant volume.
  ntc=2, !Flag for SHAKE to perform bond length constraints = 2 bonds involving hydrogen are constrained.
  ntf=2, !Force evaluation bond interactions involving H-atoms omitted (use with NTC=2).
  ntpr=100, !Print mdout and mdinfo every X steps.
  ntr=1, !Flag for restraining atoms using an harmonic potential.
 /
Allowing only water and the counterions to move with 2 kcal/mol restraint
2.0
RES 1 273
END
END
```

Figure A-1. First minimization with 2 kcal/mol energy restraints.

```
Minimise complex with no restraints
 &cntrl
  imin=1, !Perform an energy minimization.
  maxcyc=10000, !Maximum number of cycles of minimization.
  ncyc=3000, !Switch from steepest descent to conjugate gradient after these cycles.
  cut=12.0, !Non-bonded cutoff in angstroms (VdW and electrostatics).
  ntb=1, !Constant volume.
  ntc=2, !Flag for SHAKE to perform bond length constraints = 2 bonds involving hydrogen are constrained.
  ntf=2, !Force evaluation bond interactions involving H-atoms omitted (use with NTC=2).
  ntpr=100, !Print mdout and mdinfo every X steps.
  ntr=0, !Flag for NOT restraining atoms using an harmonic potential.
 /
```

Figure A-2. Second minimization without energy restraints.

```
Heating the complex during 500 ps from 0 to 100K with 10 kcal mol restraint on complex at constant volume
 &cntrl
  imin=0, !Run MD with no minimisation.
  irest=0, !Do not restart the simulation.
  ntx=1, !Only coordinates will be read from the inpcrd file.
  nstlim=250000, !Number of MD steps to be performed.
  dt=0.002, !Time step in ps.
  ntc=2, !Flag for shake to perform bond length constraints.
  ntf=2, !Force evaluation bond interactions involving H-atoms omitted (use with NTC=2).
  tol=0.0000001, !Relative geometrical tolerance for coordinate resetting in shake.
  cut=10.0, !Nonbonded cutoff, angstroms.
  ntb=1, !Constant volume.
  ntpr=5000, !Write mdout every X steps.
  ntwx=5000, !Write rst every X steps.
  ntt=3, !Switch for temperature scaling using Langevin dynamics with gamma_ln collision frequency
  gamma_ln=1.0, !Collision frequency.
  ig=-1, !Seed for the pseudo random number generator.
  ntr=1, !Restrain atoms using harmonic potential.
  ntp=0, !No pressure scaling.
  ioutfm=1, !Write coordinate and velocity files.
  nmropt=1, !NMR restraints will be read.
  ntxo=2, !NetCDF file format for the final coordinates.
 /
 &wt TYPE='TEMP0', istep1=0, istep2=250000,
  value1=0.0, value2=100.0 /
 &wt TYPE='END' /
Hold protein and ligand fixed
10.0
RES 1 273
END
END
```

Figure A-3. First heating for 500 ps from 0 to 100K with 10 kcal/mol energy restraints on the complex at constant volume.

```
Heating the complex during 500 ps from 100K to 310K without constraints at constant pressure
 &cntrl
  imin=0, !Run MD with no minimisation.
  irest=1, !Restart the simulation.
  ntx=5, !Coordinates and velocities will be read.
  nstlim=250000, !Number of MD steps to be performed.
  dt=0.002, !Time step in ps.
  ntc=2, !Flag for shake to perfoem bond length constraints.
  ntf=2, !Force evaluation bond interactions involving H-atoms omitted (use with NTC=2).
  tol=0.0000001, !Relative geometrical tolerance for coordinate resetting in shake.
  cut=10.0, !Nonbonded cutoff (vdW and electrostatics), angstroms.
  ntb=2, !Constant pressure.
  ntpr=5000, !Print mdout mdinfo printed frequency.
  ntwx=5000, !Print mdcrd file frequency.
  ntt=3, !Switch for temperature scaling using Langevin dynamics with gamma_ln collision frequency
  gamma_ln=1.0, !Collision frequency.
  ig=-1, !Seed for the pseudo number generator.
  ntr=0, !Do not restrain atoms using harmonic potential.
  ntp=1, !Pressure scaling.
  ioutfm=1, !Write coordinate and velocity files.
  nmropt=1, !NMR restraints will be read.
  taup=2.0, !Pressure relaxation time (in ps), when NTP > 0.
  ntxo=2, !NetCDF file format for the final coordinates.
 /
 &wt TYPE='TEMP0', istep1=0, istep2=250000,
  value1=100.0, value2=310.0 /
 &wt TYPE='END' /
END
```

Figure A-4. Second heating for 500 ps from 100K to 310K without energy restraints at constant pressure.

```
Production run for 50 ns at 310K
 &cntrl
  imin=0, !Run MD with no minimisation
  irest=1, !Restart the simulation
  ntx=5, !Coordinates and velocities will be read
  nstlim=25000000, !Number of MD steps to be performed
  dt=0.002, !Time step in ps
  ntc=2, !Flag for shake to perform bond length constraints
  ntf=2, !Force evaluation bond interactions involving H-atoms omitted (use with NTC=2)
  cut=10.0, !Nonbonded cutoff
  ntb=2, !Constant pressure
  ntp=1, !Constant pressure dynamics
  ntpr=5000, !Print mdout mdinfo frequency
  ntwx=5000, !Print mdcrd file frequency
  ntwr=100000, !Print coordinates restart file frequency
  ntt=3, !Switch for temperature scaling
  gamma_ln=1.0, !Collision frequency
  temp0=310.0, !Reference temperature to keep the system at (K)
  ig=-1, !Seed for the pseudo number generator
  barostat=2, !Monte carlo barostat
  ioutfm=1,  !Write as binary netcdf trajectory
  ntxo=2, !Format of the final coordinates as netcdf
 /
```

Figure A-5. Production run for 50 ns at 310K at constant pressure.

```
# Cluster analysis with cpptraj.
# Load topology trajectory
parm G9a_solv_nowater.prmtop
trajin 5_Prod1_nowater.nc
trajin 5_Prod2_nowater.nc
# Remove ions so they do not appear in output structures.
strip :Na+ outprefix noions
# Cluster analysis command:
#  C0: Cluster output data set(s) name.
# CLUSTERING OPTIONS:
#  dbscan: Use the DBSCAN (density based) clustering algorithm.
#    minpoints: Minimum # of points to form a cluster,  epsilon: Distance cutoff for forming cluster.
#    sievetoframe: Restore sieved frames by comparing to all cluster frames, not just centroid.
# DISTANCE METRIC OPTIONS:
#  rms <mask>: Use RMSD of atoms in <mask> as distance metric.
#  sieve 10  : Use <total> / 10 initial frames for clustering.
# OUTPUT OPTIONS:
#  out <file>: Write cluster number versus time to file.
#  summary <file>: Write overall clustering summary to file.
#  info <file>: Write detailed cluster results (including DBI, pSF etc) to file.
#  cpopvtime <file> normframe: Write cluster population vs time to <file>, normalised by # frames.
# COORDINATE OUTPUT OPTIONS:
#  repout <file prefix> repfmt pdb: Write cluster representatives to files with PDB format.
#  singlerepout <file> singlerepfmt netcdf: Write cluster representatives to single file with NetCDF format.
#  avgout <file> avgfmt restart: Write average over all frames in each cluster to separate files with Amber restart file format.
cluster C0 \
        dbscan minpoints 25 epsilon 1.5 sievetoframe \
        rms @CA,C,O,N,H \
        sieve 10 random \
        out cnumvtime.dat \
        summary summary.dat \
        info info.dat \
        cpopvtime cpopvtime.agr normpop \
        repout rep repfmt pdb \
        singlerepout singlerep.nc singlerepfmt netcdf \
        avgout Avg avgfmt restart
run
```

Figure A-6. CPPTRAJ input to perform the clustering of MD trajectories.

A

```
blosum=dlmread('blosum.txt');
blosum_names=['A' ,'R' , 'N', 'D' , 'C', 'Q' , 'E' , 'G' , 'H' , 'I' , ...
    'L' , 'K' , 'M' , 'F' , 'P' , 'S' , 'T' , 'W' , 'Y' , 'V' , 'B' , 'Z' , 'X', '*'];

%values=dlmread('values.txt');

ff=fopen('alignment.txt');
alignment=textscan(ff,'%s');
%values{1}
fclose(ff);

rows=7;
columns=202;

g9a=alignment{1}(1:columns);

values(1:rows-1,1:columns)=0;
for i=1:rows-1
    for j=1:columns
        comparison=alignment{1}(i*columns+j);
        ref=alignment{1}(j);
        values(i,j)=finder(blosum,blosum_names,comparison,ref);
    end
end

for i=1:rows-1
    dlmwrite('values.txt',values(i,:),'-append','newline','pc');
end
```

B

```
function [evolution]=finder(blosum,blosum_names,a,b)
%Find the number where the columns A and B cross in the matrix blosum

%Find the line of A
for i=1:length(blosum_names)
    if blosum_names(i)==a{1}
        a_index=i;
        break
    end
end

%Find the line of B
for i=1:length(blosum_names)
    if blosum_names(i)==b{1}
        b_index=i;
        break
    end
end

evolution=blosum(a_index,b_index);
```

Figure A-7. A) MATLAB code to assign conservation values based on the frequency of each residue at each position in a multiple sequence alignment according to the BLOSUM62 matrix. B) Finder function definition.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure A-8. BLOSUM62 matrix.

```
/rvol/cpse-apps/RHEL5-x86_64/RosettaCommons/rosetta-3.5/rosetta_source/bin/minirosetta.linuxgccrelease
@/home/ps/cefs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/ro
setta_inputs/comparative_modeling.args

-loops:frag_files
/home/ps/cefs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/roset
ta_inputs/FragPsiPredG9a/aat000_09_05.200_v1_3
/home/ps/cefs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/roset
ta_inputs/FragPsiPredG9a/aat000_03_05.200_v1_3 none

-in:file:psipred_ss2
/home/ps/cefs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/roset
ta_inputs/FragPsiPredG9a/t000_.psipred_ss2

-in:file:alignment

/home/ps/cefs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/exp9
_fixed.grishin

-in:file:template_pdb

/home/ps/ce-
fs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/*.pdb

-database /rvol/cpse-apps/RHEL5-x86_64/RosettaCommons/rosetta-3.5/rosetta_database

-out:file:silent HomModG9aexp9NEW.out

-in:file:fasta

/home/ps/cefs1/da1712/Diego/RosettaExperimentsJuly/4_HomologyModellingWithAnExtension/hommodexp9/Quer
ySequenceExp9.fasta.txt

-nstruct 10000
```

Figure A-9. Example of Rosetta homology modelling script.

```
/rvol/cpse-apps/RHEL5-x86_64/RosettaCommons/rosetta-3.5/rosetta_source/bin/cluster.linuxgccrelease

-database /rvol/cpse-apps/RHEL5-x86_64/RosettaCommons/rosetta-3.5/rosetta_database

-in:file:silent HomModG9aexp9NEW.out

-out:file:silent

-in::file::fullatom

-cluster:radius 1.5

-cluster:limit_cluster_size_percent 2.0

-native 3hnaA.pdb > clusterexp9_2percent.log
```

Figure A-10. Example of clustering algorithm operation in Rosetta.

Table A-1. Copyright permissions.

Element	Source	Copyright Holder	Permission Granted	License Number
Figure 1-2	Jansen, A. & Verstrepen, K. J. Nucleosome positioning in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. 75, 301–20 (2011).	Nature Publishing Group	Yes	4178351399850
Figure 2-1	Helin, K. & Dhanak, D. Chromatin proteins and modifications as drug targets. Nature 502, 480–8 (2013).	Nature Publishing Group	Yes	4178360578878
Figure 2-2	Tan, M. et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 146, 1016–28 (2011).	Elsevier	Yes	4178361009941
Figure 2-3	Helin, K. & Dhanak, D. Chromatin proteins and modifications as drug targets. Nature 502, 480–8 (2013).	Nature Publishing Group	Yes	4178361324137
Figure 2-4	Xiao, B., Wilson, J. R. & Gamblin, S. J. SET domains and histone methylation. Curr. Opin. Struct. Biol. 13, 699–705 (2003).	Elsevier	Yes	4178370084752
Figure 2-5	Wu, H. et al. Structural biology of human H3K9 methyltransferases. PLoS One 5, e8570 (2010).	Open-access article distributed under the terms of the Creative Commons Attribution License	Yes	-
Figure 2-6	Wu, H. et al. Structural biology of human H3K9 methyltransferases. PLoS One 5, e8570 (2010).	Open-access article distributed under the terms of the Creative Commons Attribution License	Yes	-
Figure 2-7	Collins, R. E. et al. The ankyrin repeats of G9a and GLP histone methyltransferases are mono- and dimethyllysine binding modules. Nat. Struct. Mol. Biol. 15, 245–50 (2008).	Nature Publishing Group	Yes	4178370663394
Figure 2-8	Xiao, B. et al. Structure and catalytic mechanism of the human histone methyltransferase SET7/9. Nature 421, 652–6 (2003).	Nature Publishing Group	Yes	4178370949302
Figure 2-9	Biggar, K. K. & Li, S. S.-C. Non-histone protein methylation as a regulator of cellular signalling and function. Nat. Rev. Mol. Cell Biol. 16, 5–17 (2014).	Nature Publishing Group	Yes	4178371170875
Figure 3-1	Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. ACS Chem. Biol. 7, 443–63 (2012).	American Chemical Society	Yes	-
Figure 3-2	Ibanez, G. et al. A High Throughput Scintillation Proximity Imaging Assay for Protein Methyltransferases. Combinatorial Chemistry & High Throughput Screening 15, 359–371 (2012).	Bentham Science Publishers	Yes	4180130850542
Figure 3-3	Gauthier, N. et al. Development of homogeneous nonradioactive methyltransferase and demethylase assays targeting histone H3 lysine 4. J. Biomol. Screen. 17, 49–58 (2012).	Society for Laboratory Automation and Screening	Yes	-
Figure 3-4	Moradian, A., Kalli, A., Sweredoski, M. J. & Hess, S. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. Proteomics 14, 489–97 (2014).	John Wiley & Sons, Inc.	Yes	4178400598542
Figure 3-5	Bartke, T., Borgel, J. & DiMaggio, P. a. Proteomics in epigenetics: new perspectives for cancer research. Brief. Funct. Genomics 12, 205–18 (2013).	This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License	Yes	-
Figure 3-6	Arrowsmith, C. H. et al. The promise and peril of chemical probes. Nat. Chem. Biol. 11, 536–541 (2015).	Nature Publishing Group	Yes	4178401013069
Figure 3-9	Urich, R. et al. De novo design of protein kinase inhibitors by in silico identification of hinge region-binding fragments. ACS Chem. … 1044–1052 (2013). Further permissions related to this material should be directed to the ACS. Link for further copyright permissions: http://pubs.acs.org/doi/abs/10.1021%2Fcb300729y.	American Chemical Society	Yes	-
Figure 3-11	Luo, M. Current chemical biology approaches to interrogate protein methyltransferases. ACS Chem. Biol. 7, 443–63 (2012).	American Chemical Society	Yes	-
Figure 3-12	Wang, R. et al. Profiling Genome-Wide Chromatin Methylation with Engineered Posttranslation Apparatus within Living Cells. J. Am. Chem. Soc. 135, 1048–1056 (2013).	American Chemical Society	Yes	-
Figure 5-1	Ponder, J. W. & Case, D. A. Protein Simulations. Advances in Protein Chemistry 66, (Elsevier, 2003).	Elsevier	Yes	4196030166451
Figure 5-2	Steinbrecher, T., Case, D. A. & Labahn, A. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. J. Med. Chem. 49, 1837–1844 (2006).	American Chemical Society	Yes	-
Figure 6-12	Wu, H. et al. Structural biology of human H3K9 methyltransferases. PLoS One 5, e8570 (2010).	Open-access article distributed under the terms of the Creative Commons Attribution License	Yes	-