Identifying the DNA Binding Specificity of Mixed Lineage Leukaemia in Leukaemia

Catherine Chahrour
St. Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2025

Abstract

Chromosomal translocations of the *KMT2A* gene generate MLL fusion proteins that drive aberrant gene expression and define an aggressive, poor-prognosis subset of acute leukaemia. Although MLL binds unmethylated CpG-island promoters through its CXXC domain, only a subset of these potential sites is occupied in the fusion context. This thesis investigates the determinants of that selectivity across three regulatory layers: DNA sequence, co-factors and chromatin environment, and DNA methylation, including 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC).

Attribution analyses from DNA-sequence models revealed a shared CpG baseline across cell types, characterised by strong positive correlations with CG-rich motifs such as VEZF1, KLFs, and MAZ. In the non-fusion MLL context, this CpG signal was refined by negatively attributed motifs, including MYC and KLF6, restricting binding to a selective subset of CpG islands. In MLL-fusion cells, CpG density contributed less strongly to prediction, and negatively attributed lineage-specific motifs such as RUNX and SP-family transcription factors highlighted cell-type specificity, suggesting that transcription-factor context along with CpG content, determines occupancy.

Models based on co-binding proteins and histone marks revealed two regimes: MLL co-localised with Menin, LEDGF, and initiation machinery at promoters, whereas MLL-AF4 was enriched along active transcription units with elongation-associated complexes, including PAF1, the super-elongation complex, BRD proteins, RNA polymerase II, and DOT1L with H3K79me2. These co-binding patterns indicate that chromatin context directs selectivity within CpG-rich regions.

Integrating DNA methylation with hydroxymethylation showed that unmethylated CpG provides a baseline requirement for binding, but differential cytosine modification states sharpen site selection. Across cell types, MLL-AF4 cells displayed elevated promoter 5mC and globally reduced 5hmC. Inclusion of 5hmC features improved prediction of MLL occupancy. Local 5hmC at the +2 nucleosome emerged as a salient predictor in a non-fusion context. In MLL-AF4 correlations with TET2 binding suggest that TET2-driven demethylation facilitates recruitment at specific elongation-linked sites.

Together, these results support a model in which CpG recognition defines baseline potential, while transcription-factor context and chromatin environment tune site selection. TET2-mediated modulation of methylation state adds a further regulatory layer that determines which CpG-rich regions are ultimately bound.

Acknowledgements

In memory of my Mum, Colleen Tresnan (1956-2024), and my Dad, Richard Squire (1950-2015). Love you to the moon and back, thank you for making me the person I am today.

First and foremost, I would like to thank my supervisors, Tom Milne and Alastair Smith. Tom, thank you for being a thoughtful and kind mentor throughout providing useful insights, many informative anecdotes, endless memes, and pastries. Alastair, I can acknowledge that you are indeed my supervisor! Thank you for all the back-and-forth, the invaluable instruction on, well all of it (including egg-sucking), and for your friendship and support through all the bugs and chaos.

To the wonderful friends I've made in the lab (a.k.a. the SeqNado beta-testers) in no particular order but firstly Vassi, thank you for all the welfare, the snackies, and the plot appreciation. This would have been much sadder without you. Ana, you are truly a storm and will change the world, a special thank you for being my rock in Madrid. Disha, thank you for being a delight. Gianna, thank you for your kindness and being a sensible influence. Lucia, I can't wait to meet Alba. Rebecca, even though you just got here it's been lovely getting to know you. Grace, best collaborator ever! Nicole, it's been a whirlwind, I hope Australia is everything I told you it would be and more. A special thanks to Joe Harman for reading the SeqNado chapter and for being you.

To my GMS besties; Mo, Dan, and Bana it's been a journey. Even though we had to go off and do separate projects (rude!), each of you has been a constant support throughout. I can't wait to see what comes next for you all.

To my siblings; Claire (and Dan, Carter, Ellie, and Emmett) and Paul (and Rachel) thank you for your love and pride, especially through the loss of Mum. Thank you for dealing with everything I will be forever grateful. To Paul and Rach, congratulations on your wedding, sorry I can't make it, may all your toast be non-soggy.

To the Chahrours; Hussein, Mariam, Maria (and Ahmad, Adam, and Malek), Rima (and Adam and Magnus), and Ali. Each of you holds a special place in my heart and you constantly inspire me.

To my oldest and bestest friends; Ahmed, Daria, and Shoaib, sorry for moving Hassan and Ruby to Oxford, but thank you for visiting us so much (don't get excited Ahmed, the list is alphabetical).

And to my husband Hassan, I can't believe all we've managed over these years, including building a home. I can't wait to find out what's next. "Thank you" is nowhere near strong enough for all your support, without you I am less.

And finally, and most importantly, to Ruby, welfare dog extraordinaire.

This work was funded by a Wellcome Trust PhD Studentship.

Declaration

I declare that the work presented in this thesis is my own unless otherwise stated, for example where my analyses used data from published work or data from collaborators. This thesis has not been submitted, either partially or in full, for another qualification of this University, or for a qualification at any other institution.

The following experiments were conducted in collaboration with others:

1. CUT&Tag experiments in SEM and patient samples were conducted by Dr. Alastair L. Smith.
2. CUT&Tag experiments in OCI-AML3 cells were conducted by Dr. Rebecca Maynard.
3. ChIPmentation experiments in patient samples were conducted by Dr. Alastair L. Smith.
4. ChIP-seq experiments in SEM cells were conducted by members of the laboratory of Prof. Thomas Milne.
5. ATAC-seq experiments in RCH-ACV cells were conducted by Ana Dopico-Fernandez.
6. ATAC-seq experiments in SEM cells were obtained from GSE117862 (Godfrey, L. et al. 2019).
7. TAPS experiments in SEM and RCH-ACV cells were obtained from the laboratory of Prof. Skirmantas Kriaucionis.
8. Patient samples were obtained from the laboratories of Prof. Anindita Roy and Dr. Ronald W. Stam.

List of Publications

Lau, I.-J., Bloye, G., Smith, A. L., Harman, J. R., Hamley, J. C., Sharlandjieva, V., Denny, N., **Chahrour, C.**, Li, H., Jackson, N. E., Vyas, P., Davies, J. O. J., Hughes, J. R., Crump, N. T., Milne, T. A. (2026). 'MYB activity drives emergent enhancer activation and enhancer-promoter interactions in acute lymphoblastic leukemia'. *Blood*. (in press)

Dopico-Fernandez, A. M., Li, H., **Chahrour, C.**, Dalgleish, J. L. T., Davies, J. O. J., Beagrie, R. A., Milne, T. A. (2026). 'FACT safeguards promoter topology by maintaining nucleosomes and restricting chromatin factor spreading'. *bioRxiv*, https://doi.org/10.64898/2026.02.18.706382

Sharlandjieva, V., **Chahrour, C.**, Lassen, F. H., Hamley, J. C., Damianou, A., Denny, N., Smith, A. L., Hester, S. S., Vendrell, I., Stam, R. W., Konopleva, M., Roy, A., Davies, J. O. J., Crump, N. T., Kessler, B. M., Milne, T. A. (2026). 'Menin maintains enhancer-promoter interactions in a leukemia-specific manner'. *bioRxiv*, https://doi.org/10.64898/2026.01.16.698179

Ciaurro, V., Sharlandjieva, V., Skwarska, A., **Chahrour, C.**, Baran, N., Zeng, Z., Ramage, C. L., Daver, N., Carter, B. Z., Chaudhry, S., Thandapani, P., Martelli, M. P., Milne, T. A., Konopleva, M. (2025). 'Menin inhibitor DS-1594b drives differentiation and induces synergistic lethality in combination with venetoclax in acute myeloid leukemia cells with rearranged Mixed-lineage Leukemia and mutated Nucleophosmin-1'. *Haematologica*, https://doi.org/10.3324/haematol.2024.286833

Meaker, G. A., Nicholls, M., **Chahrour, C.**, Hsu, I., Smith, A. L., Bozhilov, Y., Leung, M., Vassort, H., Olender, L., Beaven, O., Huang, X., Brown, E. J., Vanden Bempt, M., Khoo, H. M., Bhadury, J., Milne, T. A., Wilkinson, A. C. (2025). 'Genome-wide screen identifies *Runx2* as a novel regulator of haematopoietic stem cell expansion'. *Blood*. https://doi.org/10.1182/blood.2025029115

Smith, A. L., Denny, N. D. R., **Chahrour, C.**, Sharp, K., Arachi, M., Dopico-Fernandez, A. M., Elliott, N., Harman, J. R., Jackson, T., Geng, H., Smith, O., Bond, J., Roberts, I., Stam, R. W., Crump, N. T., Davies, J. O. J., Roy, A., Milne, T. A. (2025). 'Enhancer heterogeneity in acute lymphoblastic leukemia drives differential gene expression in patients'. *Blood*, https://doi.org/10.1182/blood.2024028019

Ling, R., Jackson, T., Elliott, N., Cross, J., Hamer, L., Wuppalapati, A., Smith, A. L., **Chahrour, C.**, Sevim, O., Iskander, D., Wang, G., Rice, S., O'Byrne, S., Harman, J., Psaila, B., Morgan, R., Roberts, I., Milne, T. A., Roy, A. (2024). 'The fetal specific gene *LIN28B* is essential for human fetal B-lymphopoiesis and initiation of KMT2A::AFF1 infant leukemia'. *bioRxiv*, https://doi.org/10.1101/2024.09.18.613730

Crump, N. T., Smith, A. L., Godfrey, L., Dopico-Fernandez, A. M., Denny, N., Harman, J. R., Hamley, J. C., Jackson, N. E., **Chahrour, C.**, Riva, S., Rice, S., Kim, J., Basrur, V., Fermin, D., Elenitoba-Johnson, K., Roeder, R. G., Allis, C. D., Roberts, I., Roy, A., Geng, H., Davies, J. O. J., Milne, T. A. (2023). 'MLL-AF4 cooperates with PAF1 and FACT to drive high density enhancer interactions in leukemia'. *Nature Communications*, https://doi.org/10.1038/s41467-023-40981-9

Wiggins, B. G., Wang, Y.-F., Burke, A., Grunberg, N., Vlachaki Walker, J. M., Dore, M., **Chahrour, C.**, Pennycook, B. R., Sanchez-Garrido, J., Vernia, S., Barr, A. R.,

Frankel, G., Birdsey, G. M., Randi, A. M., Schiering, C. (2023). 'Endothelial sensing of AHR ligands regulates intestinal homeostasis'. *Nature*, https://doi.org/10.1038/s41586-023-06508-4

Conference Abstracts

Chahrour, C., Smith, A. L., Milne, T. A. (2024) 'Identifying the DNA binding specificity of chromatin complexes in leukaemia'. *Molecular Haemopoiesis*, London, UK.

Chahrour, C., Smith, A. L., Milne, T. A. (2024) 'DNA binding specificity of MLL-AF4 in leukemia'. *Genome Regulation and Cellular Fates in Homeostasis and Disease*, Madrid, Spain.

Contents

Table of Contents v

List of Figures xii

List of Tables xiv

List of Abbreviations xv

1 Introduction 1

1.1 Mixed Lineage Leukaemia (MLL) 1

1.1.1 The MLL complex and cofactors 2

1.1.2 MLL as a histone methyltransferase 2

1.1.3 Developmental role of MLL 2

1.2 MLL Rearrangements and Leukaemia 2

1.2.1 Fusion mechanisms 3

1.2.2 Epigenetic dysregulation in leukaemia 4

1.3 Chromatin and Histone Context 4

1.3.1 Chromatin remodelling and accessibility 5

1.3.2 Transcription Factors and Chromatin Interactions 5

1.3.3 Histone modifications as regulatory signals 6

1.3.4 Chromatin dynamics in development and haematopoiesis . . 7

1.4 DNA Methylation and Hydroxymethylation 7

1.4.1 The Role of DNA methylation in gene regulation 9

5-methylcytosine: canonical and context-dependent roles . . 9

5-hydroxymethylcytosine: beyond a demethylation intermediate 10

1.4.2 DNA Methylation in leukaemia 10

1.5 The Problem of Binding Specificity 11

1.5.1 Recruitment to target sites 12

1.5.2 Intrinsic sequence features and motifs 12
1.5.3 The MLL complex 12
1.5.4 MLL fusion partners and the Super Elongation Complex . . . 12
1.5.5 Histone marks as signposts of regulatory activity 13
1.5.6 MLL binding and DNA methylation 14
1.5.7 Towards clarity of MLL recruitment logic 14
1.6 Machine Learning Approaches in Genomics 15
1.6.1 From Statistical Models to Deep Learning and Beyond . . . 15
1.6.2 Choosing the right model for the data 17
1.6.3 Model Interpretability 17
1.7 Thesis Aim and Objectives 18

2 Materials and Methods 19
2.1 Cell Culture 19
2.1.1 Culturing THP-1, SEM, RCH-ACV and RS4;11 cell lines . . . 19
2.1.2 Patient samples 19
2.2 Chromatin Immunoprecipitation Methods 20
2.2.1 ChIPmentation 20
2.2.2 Cleavage Under Targets and Tagmentation (CUT&Tag) . . . 20
2.3 Simultaneous 5mC, 5hmC, and gDNA sequencing 21
2.3.1 gDNA Extraction, Fragmentation and Quantification 21
2.3.2 Library Preparation 22
2.3.3 Methylome Enrichment and Sequencing 22
2.4 Whole Genome Sequencing 23
2.4.1 Library Preparation and Sequencing 23

3 SeqNado: Uniform and Reproducible Data Processing for ML-Ready Outputs 24
3.1 Introduction 24
3.2 Pipeline framework and workflow language 26
3.2.1 Pipelines 26
3.2.2 Workflow languages 26
3.3 SeqNado 28
3.3.1 SeqNado Implementation 28
3.3.2 User Workflow 29
Installation 29

Step 1: Initialisation (one-time) . . . 30
Step 2: Configure the run . . . 30
Step 3: Generate the design file . . . 31
Step 4: Run the pipeline . . . 32
Outputs . . . 33
3.3.3 Summary . . . 35
3.3.4 Supported Assays . . . 35
RNA-Seq . . . 35
ATAC-Seq . . . 36
ChIP-seq and CUT&Tag . . . 36
Whole Genome sequencing . . . 36
Methylation . . . 37
CRISPR screen . . . 37
MCC . . . 37
3.3.5 Outputs . . . 37
UCSC hub . . . 37
GEO submission . . . 37
Plotting . . . 38
Quality control metrics . . . 38
Machine Learning Integration . . . 38
3.4 Pipelines Benchmarked . . . 40
3.4.1 nf-core . . . 40
3.4.2 ENCODE-DCC . . . 41
3.4.3 snakePipes . . . 41
3.5 Benchmarking Results . . . 43
3.5.1 Computational Performance . . . 43
3.5.2 Output Quality . . . 43
3.6 Case Study: High-Throughput CUT&Tag Processing . . . 46
3.7 Discussion . . . 47

4 DNA Sequence specificity in MLL binding 49
4.1 Introduction . . . 49
4.1.1 A motif for MLL . . . 50
4.1.2 DNA Sequence and Machine Learning . . . 51
4.1.3 Chapter Aims . . . 52
4.2 Methods . . . 54

4.2.1 Data preparation 54
Regions and signal extraction 54
Normalisation, binarisation, and data partitioning 55
Sequence composition and motif enrichment 55
4.2.2 Model architectures 55
4.2.3 Fine-tuning and evaluation 59
Convolutional neural network baseline 59
Transformer models and LoRA fine-tuning 59
Evaluation and threshold calibration 60
Embedding analysis 60
4.2.4 Attribution and motif discovery 61
Genomic annotation of predictions 61
Integration with gene expression 61
4.2.5 Software versions and environments 62
4.3 Results 63
4.3.1 Dataset Design and Predictive Task Formulation 63
Assay choice and data labelling 63
Region choice 64
Windowing and normalisation 65
Binarisation and label validation 66
Regression versus classification 66
Data partitioning and leakage control 66
Class balance and negative sampling 67
4.3.2 Dataset Distribution and labelling 67
4.3.3 Model Benchmarking 70
Model performance after fine-tuning for MLL-N binding (multi-label classification) 70
4.3.4 MLL Binding Across Cell Lines 73
MLL-N Predictions per label for methylome generalised to chr9 tiled 1024bp regions 73
Model Embeddings 78
Feature Extraction 80
Validation with RUNX1 80
Token level attribution 81
Sequence-level attributions and motif discovery 83
4.3.5 Sequence composition 84

4.3.6 Motif enrichment . . . 86
4.4 Discussion . . . 88
4.4.1 MLL binding is shaped by CpG-island context . . . 88
4.4.2 Dataset composition influences predictive challenge . . . 88
4.4.3 DNA language models recover MLL binding with high accuracy 88
4.4.4 Token-level attributions reveal GC-driven features . . . 88
4.4.5 Seqlet-level motifs highlight conserved CpG factors and diverse suppressors . . . 89
4.4.6 Limitations and future directions . . . 89
4.4.7 Conclusions . . . 90

5 MLL binding in the context of cooperating factors 91
5.1 Introduction . . . 91
5.1.1 Co-factor binding and chromatin context . . . 91
5.1.2 Histone marks as the regulatory landscape . . . 92
5.1.3 MLL biology and transcriptional dysregulation . . . 93
5.1.4 Previous Machine Learning Approaches to Predicting TF Binding . . . 94
5.1.5 Chapter Aims . . . 95
5.2 Methods . . . 96
5.2.1 Datasets and Preprocessing . . . 96
5.2.2 Modelling Approaches . . . 96
GANDALF . . . 96
Hyperparameter sweep . . . 97
Model selection and final training . . . 97
Evaluation and logging . . . 98
XGBoost . . . 98
5.2.3 Feature Extraction . . . 98
5.2.4 Software versions and environments . . . 99
5.3 Results . . . 100
5.3.1 MLL Correlates with Distinct Co-factor Networks Across Genomic Contexts . . . 100
MLL Correlation in RCH-ACV Cells . . . 100
MLL Correlation in SEM Cells . . . 102
Comparing datasets and scaling . . . 102
5.3.2 Model benchmarking . . . 105
5.3.3 Co-factors which recruit MLL at promoters . . . 106

Model evaluation . . . 106
Feature Importance at Promoters . . . 107
5.3.4 Co-factors which recruit MLL at methylome regions . . . 109
Feature importance at methylome regions . . . 113
5.4 Discussion . . . 116
5.4.1 Mechanistic interpretation of MLL recruitment . . . 116
5.4.2 Limitations . . . 119
5.4.3 Conclusion . . . 120

6 The Role of DNA Methylation in MLL Recruitment 122
6.1 Introduction . . . 122
6.1.1 Detecting DNA methylation . . . 122
6.1.2 DNA methylation and DNA binding proteins . . . 124
6.1.3 Prediction of DNA binding proteins from methylation state . . . 125
6.1.4 Chapter Aims . . . 127
6.2 Methods . . . 128
6.2.1 Samples and sequencing . . . 128
6.2.2 Differential binding and differential methylated regions . . . 128
6.2.3 Modelling MLL from methylation features . . . 128
Dataset preparation . . . 128
Model specification and training . . . 129
Model interpretation . . . 129
Software . . . 130
6.3 Results . . . 131
6.3.1 Simultaneous 5mC and 5hmC profiling . . . 131
6.3.2 Methylation Landscape and MLL Binding . . . 134
Relationship between methylation and binding . . . 134
Differential methylation between MLL-AF4 and MLL contexts 137
6.3.3 Predicting MLL Binding from Methylation State . . . 140
Positional DNA methylation and 5hmC at TSSs . . . 141
6.3.4 TET2, MLL, and Promoter Methylation Dynamics . . . 144
6.4 Discussion . . . 148
Predicting MLL binding from methylation state . . . 148
Mechanistic implications and future directions . . . 149
Conclusions . . . 149

7 Discussion **151**

7.0.1 DNA Sequence as a Baseline Layer 151
7.0.2 Chromatin Cofactors and Context 152
7.0.3 DNA Methylation as a Modulatory Layer 153
7.0.4 Limitations and Future Directions 153
7.0.5 Concluding Remarks 154

References **156**

Appendices **172**

A Data preparation **173**

A.1 Cell lines used 173
A.2 6-letter sequencing data 173
A.3 CUT&Tag data 174
A.4 ChIPmentation data 174
A.5 ChIP-seq data 175

B Software Environments **177**

B.1 Notebook and Jupyter Packages 177
B.2 Environment: Crested 177
B.3 Environment: Transformer Datasets 178
B.4 Environment: Transformer Training 178
B.5 Environment: PyTorch Tabular 179
B.6 Environment: Methylation Model 179
B.7 Environment: XGBoost 179

C Supplementary Tables **180**

C.1 GEO metadata for CUT&Tag samples 180

List of Figures

1.1 MLL and MLL fusion schematic. 3
1.2 CpG methylation cycle and MLL binding. 8
1.3 Conceptual model of MLL recruitment and regulation. 14

3.1 Overview of SeqNado workflow from initialisation to processed data outputs. 29
3.2 Example SeqNado initialisation. 30
3.3 Example SeqNado configuration questionnaire for ChIP-seq. 31
3.4 Example SeqNado run. 33
3.5 Example output file directory from SeqNado 34
3.6 Example MultiQC report from SeqNado 35
3.7 Example SeqNado ML-ready h5ad file output. 39
3.8 Workflow benchmarking summary. 43
3.9 Output data quality metrics from pipeline benchmarking of ATAC-seq data. 46
3.10 SeqNado Runtime visualisation. 46

4.1 Wild-type and fusion MLL at the leukaemia breakpoint with antibody epitope. 49
4.2 HOCOMOCO v12 MLL motif logo 50
4.3 CUT&Tag compared to ChIP-seq 64
4.4 Effect of input sequence length on model performance. 65
4.5 Feature annotation of dataset regions. 68
4.6 Distribution and binarisation of MLL-N RPKM signal over methylome and promoter regions. 69
4.7 Model performance ROC AUC (macro) over test data sets and chr9 tiled 1024 bp. 72
4.8 GROVER training and validation loss for MLL-N binding prediction on methylome regions. 73

4.9 GROVER predictions of MLL-N binding across chromosome 9 tiled 1,024 bp regions. . . . 76
4.10 Examples of GROVER predictions of MLL-N binding across chromosome 9 tiled 1,024 bp regions. . . . 77
4.11 UMAP projection of GROVER sequence embeddings across chromosome 9. . . . 79
4.13 Token-level feature extraction from GROVER attributions of MLL binding across tiled chr9 regions. . . . 83
4.14 Motif enrichment from nucleotide-level attributions across cell types. 84
4.15 Sequence composition of promoter and methylome datasets. . . . 85
4.16 Motif enrichment in MLL-N binding sites stratified by methylation status in RCH-ACV and SEM cells. . . . 87

5.1 MLL-N correlations with co-factors differ between promoters and methylome regions in RCH-ACV cells. . . . 101
5.2 Correlation structure of MLL-N and co-binding factors across promoter and methylome regions in SEM cells. . . . 104
5.3 Model Benchmarking. . . . 106
5.4 Promoter GANDALF prediction performance. . . . 107
5.5 SHAP feature importance and associations for the promoter GANDALF model. . . . 109
5.6 Methylome GANDALF prediction performance. . . . 111
5.7 Genome browser snapshot of the *GNAQ* locus (chr9:77,714,000–78,046,000). . . . 113
5.8 SHAP feature importance and associations for the methylome GANDALF model. . . . 115

6.1 Simultaneous methylation sequencing assay overview. . . . 124
6.2 Methylation assay Comparison. . . . 132
6.3 Overview of capture performance and methylation data quality control. 133
6.4 Relationship between DNA methylation and MLL-N binding. . . . 135
6.5 MLL-N binding and DNA methylation at the *MEIS1* locus. . . . 137
6.6 Differential methylation between MLL-AF4 and MLL contexts . . . 140
6.7 XGBoost multi-target regression of MLL-N signal. . . . 144
6.8 Interplay between TET2, MLL-N, and promoter methylation in SEM cells. . . . 147

List of Tables

3.1 Comparison of Pipeline Setup, Assay Support, and Key Features Across Four Bioinformatics Workflows. 42

4.1 Summary of model architectures evaluated for MLL binding prediction. 58

4.2 Comparison of model parameter counts and reduction achieved via Low-Rank Adaptation (LoRA) fine-tuning. 60

4.3 Summary of dataset sizes and chromosome-based splits for training, evaluation, and testing. 70

5.1 Summary of dataset sizes and chromosome-based splits for training, evaluation, and testing. 96

A.1 Cell lines and patient samples used 173

A.2 6-letter sequencing data 173

A.3 CUT&Tag data 174

A.4 ChIPmentation data 174

A.4 ChIP-seq data 176

B.1 Jupyter and notebook packages used in all environments 177

B.2 Conda environment for Crested 177

B.3 Environment for transformer datasets 178

B.4 Environment for transformer training 178

B.5 Environment for PyTorch Tabular 179

B.6 Environment for methylation model 179

B.7 Environment for XGBoost 179

C.1 GEO metadata produced using SeqNado 180

List of Abbreviations

5caC 5-carboxylcytosine.
5fC 5-formylcytosine.
5hmC 5-hydroxymethylcytosine.
5mC 5-methylcytosine.

A Adenine.
ALL Acute Lymphoblastic Leukaemia.
AML Acute Myeloid Leukaemia.
AP Average Precision.
APOBEC3A Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3A.
ATAC-seq Assay for Transposase-Accessible Chromatin Using Sequencing.
AUC Area Under the Curve.
AWS Amazon Web Services.
AWS EC2 Amazon Web Services Elastic Compute Cloud.

BCE Binary Cross-Entropy.
BER Base Excision Repair.
BERT Bidirectional Encoder Representations from Transformers.
bp base pairs.
BPE Byte-Pair Encoding.

C Cytosine.
ChIP-seq Chromatin Immunoprecipitation with Sequencing.
CNN Convolutional Neural Network.
CpG Cytosine-Phosphate-Guanine.
CPU Central Processing Unit.
CREsted Cis Regulatory Element Sequence Training, Explanation and Design.
CTCF CCCTC-binding factor.
CUT&Tag Cleavage Under Targets and Tagmentation.
CXXC Cysteine-X-X-Cysteine.

DMR Differential Methylation Region.
DNA Deoxyribonucleic Acid.
DNA-LM DNA Language Models.
DNABERT Bidirectional Encoder Representations from Transformers model for DNA-language.
DNABERT-2 DNABERT-2–117M.
DNMT DNA Methyltransferase.
DNMT1 DNA Methyltransferase 1.

DNMT3A DNA Methyltransferase 3A.
DNMT3B DNA Methyltransferase 3B.
DNMT5 DNA Methyltransferase 5.

F1 Harmonic Mean of Precision and Recall.
FN False Negative.
FP False Positive.

G Guanine.
GANDALF Gated Adaptive Network for Deep Automated Learning of Features for Tabular Data.
gDNA Genomic DNA.
GFLU Gated Feature Learning Unit.
GPU Graphics Processing Unit.
GROVER Genome Rules Obtained Via Extracted Representations.

H3K27ac Histone H3 Lysine 27 Acetylation.
H3K27me3 Histone H3 Lysine 27 Trimethylation.
H3K36me3 Histone H3 Lysine 36 Trimethylation.
H3K4me3 Histone H3 Lysine 4 Trimethylation.
H3K79me Histone H3 Lysine 79 Methylation.
H3K79me2 Histone H3 Lysine 79 Dimethylation.
H3K79me3 Histone H3 Lysine 79 Trimethylation.
HATs Histone Acetyltransferases.
HDACs Histone Deacetylases.
HPC High-Performance Computing Cluster.
HSPCs Haematopoietic Stem and Progenitor Cells.

IP Immunoprecipitation.

KMT2A Histone-lysine N-methyltransferase 2A.

log2FC Log2 Fold Change.
LoRA Low-Rank Adaptation.

MCC Micro Capture-C.
ML Machine Learning.
MLL Mixed Lineage Leukaemia.
MLL-AF4 MLL fused to AF4.
MLL-AF9 MLL fused to AF9.
MLL-FP MLL fusion protein.
MLL-N MLL N-terminus.
MLLr MLL-rearranged.
MSE Mean Squared Error.

N Any Nucleotide.
NT-HUMAN Nucleotide Transformer 500m Human Reference.
NT2-MULTI Nucleotide Transformer v2 500m Multi Species.

OVR One-Versus-Rest.

p-TEFb Positive Transcription Elongation Factor b.

PCA Principal Component Analysis.
PEFT Parameter-Efficient Fine-Tuning.
PR Precision-Recall.
PTMs Post-Translational Modifications.

ROC Receiver Operating Characteristic.
RPKM Reads Per Kilobase Per Million Mapped Reads.
R^2 Coefficient of Determination.

SAF Simplified Annotation Format.
SEC Super Elongation Complex.
SHAP SHapley Additive exPlanations.
SW Smith-Waterman.

T Thymine.
TAPS Tet-assisted Pyridine Borane Sequencing.
TEFb Transcription Elongation Factor b.
TET Ten-Eleven Translocation.
TET1 Ten-Eleven Translocation 1.
TET2 Ten-Eleven Translocation 2.
TET3 Ten-Eleven Translocation 3.
TF Transcription Factor.
TN True Negative.
TP True Positive.
TPM Transcripts Per Million.
TrxG Trithorax Group.
TSS Transcription Start Site.

U Uracil.
uCpG Unmethylated Cytosine-Phosphate-Guanine.
UMAP Uniform Manifold Approximation and Projection.

WDL Workflow Description Language.
WGBS Whole Genome Bisulphite Sequencing.
WGS Whole Genome Sequencing.

XGBoost eXtreme Gradient Boosting.

βGT β-Glucosyltransferase.

1 Introduction

Precise control of gene expression arises from the interplay of Deoxyribonucleic Acid (DNA) sequence, chromatin, transcription factors, and epigenetic modifications (Spitz, F. 2012; Lambert, S. A. et al. 2018; Klemm, S. L. et al. 2019). Mixed Lineage Leukaemia (MLL), encoded by the *Histone-lysine N-methyltransferase 2A (KMT2A)* gene, normally targets Cytosine-Phosphate-Guanine (CpG)-rich promoters of developmental genes (Birke, M. 2002; Milne, T. A. et al. 2005a); however, in leukaemia, MLL fusions redirect recruitment and drive aberrant gene expression (Ziemin-van der Poel, S. et al. 1991; Krivtsov, A. V. and Armstrong, S. A. 2007; Meyer, C. et al. 2023). Yet in both non-fusion and fusion contexts, MLL occupies only a selective subset of potential sites *in vivo* (Milne, T. A. et al. 2005a; Okuda, H. et al. 2014). This thesis examines how DNA sequence, chromatin cofactors, and DNA methylation govern that selectivity by applying machine-learning models to provide an interpretable, hypothesis-generating framework that prioritises specific predictions about MLL binding for future experimental validation.

1.1 Mixed Lineage Leukaemia (MLL)

MLL was first identified in the early 1990s as a recurrent translocation breakpoint in acute leukaemias associated with particularly poor clinical outcomes (Ziemin-van der Poel, S. et al. 1991). Comparative studies revealed strong homology with *Drosophila* Trithorax, a founding member of the Trithorax Group (TrxG) proteins that maintain active gene expression states during development (Djabali, M. et al. 1992). This evolutionary conservation suggested that MLL plays a fundamental role in safeguarding developmental programmes. Early genetic experiments confirmed this: MLL knockout mice displayed embryonic lethality with widespread Hox gene dysregulation and profound defects in patterning and haematopoiesis (Yu, B. D. et al. 1995).

1.1.1 The MLL complex and cofactors

MLL functions within a large multi-protein complex containing the WRAD core subunits—WDR5, RBBP5, ASH2L, and DPY30—which are essential for full catalytic activity and stabilise MLL-chromatin interactions (Dou, Y. et al. 2006). This complex serves as an integration hub linking chromatin recognition, histone modification, and transcriptional machinery. MLL is further stabilised by cofactors Menin and LEDGF: Menin bridges MLL to LEDGF, whose PWWP domain recognises Histone H3 Lysine 36 Trimethylation (H3K36me3) and whose AT-hooks bind DNA, anchoring the complex at actively transcribed genes (Yokoyama, A. et al. 2002; Milne, T. A. et al. 2005a; Allen, M. D. et al. 2006). Together, these interactions tether MLL to its target promoters and integrate it with elongation-competent RNA polymerase II through the PAF1 complex (Milne, T. A. et al. 2010).

1.1.2 MLL as a histone methyltransferase

Biochemical studies demonstrated that MLL has intrinsic histone methyltransferase activity, catalysing methylation of histone H3 at lysine 4, a hallmark of transcriptionally active promoters (Milne, T. A. et al. 2002; Nakamura, T. et al. 2002). Through deposition of Histone H3 Lysine 4 Trimethylation (H3K4me3), MLL acts as a transcriptional co-activator, maintaining expression of developmental regulators such as Hox genes at appropriate levels during differentiation.

1.1.3 Developmental role of MLL

The critical developmental function of MLL is underscored by its regulation of Hox genes, which govern anterior-posterior patterning and haematopoietic lineage specification. MLL activity at these loci ensures that once developmental programmes are established, they are faithfully maintained through successive cell divisions. In haematopoietic stem and progenitor cells, MLL helps balance self-renewal with differentiation, a balance that is frequently disrupted in leukaemia.

1.2 MLL Rearrangements and Leukaemia

Chromosomal translocations involving *KMT2A*/*MLL* define a molecularly distinct subset of acute leukaemias with characteristic partner usage and transcriptional programmes (Ziemin-van der Poel, S. et al. 1991; Meyer, C. et al. 2023). These

rearrangements occur across ages but are disproportionately frequent in infants, can arise *in utero*, and are also observed after treatment with topoisomerase II inhibitors, consistent with multiple routes to fusion formation (Felix, C. A. et al. 1995; Gale, K. B. et al. 1997; Andersson, A. K. et al. 2015; Meyer, C. et al. 2023). As schematised in Figure 1.1, MLL is targeted to unmethylated CpG-island promoters via its N-terminal Cysteine-X-X-Cysteine (CXXC) domain and is stabilised by Menin-LEDGF interactions, depositing H3K4 methylation to maintain promoter activity at developmental genes (Birke, M. 2002; Yokoyama, A. et al. 2002; Milne, T. A. et al. 2005a; Allen, M. D. et al. 2006).

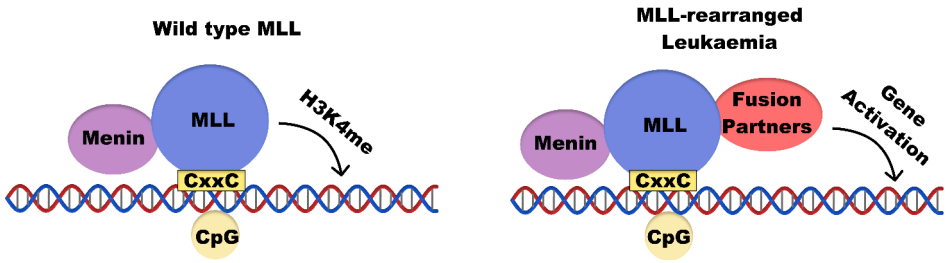


Figure 1.1: MLL and MLL fusion schematic. MLL binds unmethylated CpG islands via its CXXC domain, stabilised by Menin, and deposits H3K4 methylation to maintain promoter activity (Left). In leukaemia, chromosomal translocation generates an MLL fusion protein that retains the CXXC targeting module but gains additional activation capacity through its fusion partner, resulting in inappropriate gene activation (Right).

1.2.1 Fusion mechanisms

In wild-type form, MLL is targeted to unmethylated CpG-island promoters via its N-terminal CXXC domain and is stabilised by Menin-LEDGF interactions, maintaining promoter activity through H3K4 methylation (Birke, M. 2002; Milne, T. A. et al. 2005a). In MLL fusions, the N-terminal targeting modules are retained, whereas partner-derived sequences replace the C-terminal SET domain and introduce constitutive activation interfaces (Meyer, C. et al. 2023). Common partners—AF4 (AFF1), AF9 (MLLT3), ENL (MLLT1), and ELL—link MLL to the Super Elongation Complex (SEC), promoting RNA Pol II pause release and productive elongation at tethered loci (Lin, C. et al. 2010; Okuda, H. et al. 2014). Several partners also recruit DOT1L, establishing Histone H3 Lysine 79 Dimethylation (H3K79me2)/3 and a reinforcing elongation-associated feedback loop (Mueller, D. et al. 2009).

Mechanistically, this partner-mediated recruitment reprogrammes genomic occupancy, driving inappropriate activation of developmental regulators such as HOXA9 and MEIS1 that sustain self-renewal and block differentiation (Krivtsov, A. V. and Armstrong, S. A. 2007; Okuda, H. et al. 2014). Importantly, MLL and its fusion

proteins can co-occupy loci within the same cell, with wild-type MLL maintaining promoter activity while the fusion provides aberrant elongation capacity, thereby reinforcing oncogenic transcriptional circuits (Mueller, D. et al. 2009; Okuda, H. et al. 2014).

1.2.2 Epigenetic dysregulation in leukaemia

Beyond MLL fusions, leukaemias commonly feature mutations in genes regulating cytosine modification, such as DNA Methyltransferase 3A (DNMT3A) and Ten-Eleven Translocation 2 (TET2), or histone modifiers including EZH2, CREBBP, and ASXL1 (Fong, C. Y. et al. 2014). These lesions reconfigure chromatin landscapes, alter accessibility, and reshape enhancer logic to sustain aberrant transcription (Fang, C. et al. 2020). Aberrant DNA methylation is particularly pervasive: methylated CpGs can occlude transcription factor motifs, whereas unmethylated CpGs permit recruitment. Thus, methylation can shape transcription factor occupancy by favouring certain binding profiles over others (Yin, Y. et al. 2017).

In de novo Acute Myeloid Leukaemia (AML), recurrent lesions in DNMT3A and TET2 highlight methylation imbalance as a shared axis of leukaemogenesis (The Cancer Genome Atlas Research Network 2013; Tyner, J. W. et al. 2018). These modifications also constrain the genomic sites accessible to MLL fusion complexes, as their CXXC domains cannot bind methylated DNA.

1.3 Chromatin and Histone Context

MLL interacts with DNA within the context of chromatin—the dynamic complex of DNA and histones remodelled by ATP-dependent enzymes and histone modifiers that together govern genomic accessibility and transcription.

Eukaryotic genomes contain vast amounts of DNA that must be compacted to fit within the nucleus while remaining accessible for essential processes such as transcription, replication, recombination, and repair. This is achieved by packaging DNA into chromatin, a hierarchical and dynamic scaffold that serves both as a physical support and a regulatory platform. The fundamental unit of chromatin is the nucleosome, consisting of 147 bp of DNA wrapped around a histone octamer containing two copies each of histones H2A, H2B, H3, and H4 (Luger, K. et al. 1997).

1.3.1 Chromatin remodelling and accessibility

Control of DNA accessibility is mediated by ATP-dependent chromatin remodelling complexes, which reposition, restructure, or evict nucleosomes to regulate access to regulatory elements. These enzymes fall into several major families, including SWI/SNF, ISWI, CHD, and INO80, each with distinct biochemical functions and developmental roles. For example, SWI/SNF complexes can evict nucleosomes to expose previously inaccessible DNA and facilitate Transcription Factor (TF) binding at promoters—sites that initiate transcription—and at enhancers, which help regulate gene expression through promoter interactions. In contrast, ISWI complexes maintain regular nucleosome spacing, promoting more repressive chromatin domains (Clapier, C. R. and Cairns, B. R. 2009). In addition to remodellers, a subset of TFs act as pioneer factors, binding target sites even within condensed chromatin and initiating local opening by recruiting co-factors that stabilise accessibility. This hierarchical recruitment is particularly important during developmental transitions, where new transcriptional programmes must be activated as cells commit to specific lineages (Zaret, K. S. and Carroll, J. S. 2011). Genome-wide profiling methods such as Assay for Transposase-Accessible Chromatin Using Sequencing (ATAC-seq) (Buenrostro, J. D. et al. 2013) and Cleavage Under Targets and Tagmentation (CUT&Tag) (Kaya-Okur, H. S. et al. 2019) have revealed that accessibility patterns are highly cell-type-specific, reflecting the distinct regulatory networks active in different developmental and disease contexts.

1.3.2 Transcription Factors and Chromatin Interactions

TFs are sequence-specific DNA-binding proteins that recognise short motifs within regulatory elements such as promoters and enhancers, recruiting the transcriptional machinery and co-regulators to activate or repress transcription (Spitz, F. 2012; Lambert, S. A. et al. 2018). The human genome encodes over 1,600 TFs (Vaquerizas, J. M. et al. 2009), which act in combinatorial assemblies that integrate developmental and environmental signals to generate precise, cell-type-specific transcriptional programmes (Heinz, S. et al. 2015). While most TFs bind pre-accessible chromatin, a subset termed pioneer factors can recognise motifs in nucleosomal DNA, initiating local chromatin opening and enabling subsequent recruitment of other factors (Zaret, K. S. and Carroll, J. S. 2011; Zaret, K. S. and Mango, S. E. 2016). TF occupancy varies across cell types and developmental stages in step with accessibility, rather than being fixed (Klemm, S. L. et al. 2019).

During haematopoietic differentiation, lineage-defining TFs such as RUNX1, PU.1 (SPI1), GATA factors, and TAL1/SCL establish and maintain blood cell identity by activating lineage-specific gene expression programmes while repressing alternative fates (Wilson, N. K. et al. 2010; Lichtinger, M. et al. 2012). Changes in their occupancy coincide with large-scale remodelling of chromatin accessibility and histone marks, reflecting progressive restriction of developmental potential (Heinz, S. et al. 2010; Lara-Astiaso, D. et al. 2014). Dysregulation of these TF networks through mutation or mis-targeting is a common driver of haematological malignancies, including acute leukaemia (Tsankov, A. M. et al. 2015; Fang, C. et al. 2020). MLL relies on co-factors such as Menin and LEDGF to stabilise its binding at target promoters (Yokoyama, A. et al. 2004; Milne, T. A. et al. 2005a). In MLL-rearranged (MLLr) leukaemias, fusion partners hijack these interactions to recruit elongation machinery such as SEC and DOT1L, redirecting transcriptional activation to inappropriate sites (Lin, C. et al. 2010; Deshpande, A. J. et al. 2013). Thus, TF-chromatin interactions create the context in which MLL operates, linking intrinsic DNA sequence to higher-order chromatin regulation and setting the stage for the selective binding patterns explored in later chapters.

1.3.3 Histone modifications as regulatory signals

Beyond physical remodelling, chromatin is regulated by a diverse array of Post-Translational Modifications (PTMs) to histone proteins. These modifications include acetylation, methylation, phosphorylation, and ubiquitination, each of which can alter nucleosome structure or serve as a signal for the recruitment of specialised effector proteins (Kouzarides, T. 2007; Zhao, S. et al. 2021). PTMs act combinatorially to define distinct chromatin states, providing a layer of epigenetic information that integrates multiple regulatory inputs.

For example, histone acetylation, catalysed by Histone Acetyltransferases (HATs), neutralises the positive charge on lysine residues, weakening histone-DNA interactions and promoting an open, transcriptionally permissive state. Conversely, deacetylation by Histone Deacetylases (HDACs) leads to chromatin compaction and gene repression (Bannister, A. J. and Kouzarides, T. 2011). Histone methylation is more context-dependent H3K4me3 is a hallmark of active promoters, whereas Histone H3 Lysine 27 Trimethylation (H3K27me3) is associated with Polycomb-mediated gene silencing (Bernstein, B. E. et al. 2005).

These modifications form a regulatory language interpreted by effector proteins that read specific combinations of marks to recruit downstream complexes (Strahl, B. D.

and Allis, C. D. 2000). For instance, Histone H3 Lysine 27 Acetylation (H3K27ac) at distal regulatory elements signals active enhancers, while H3K27me3 marks repressed developmental genes poised for future activation. The interplay of these marks enables precise and flexible control of gene expression.

1.3.4 Chromatin dynamics in development and haematopoiesis

Chromatin structure plays a central role in development, where cell fate decisions depend on the coordinated activation of lineage-specific gene programmes and repression of alternative fates. In early embryonic cells, chromatin is generally open and accessible, reflecting a high degree of developmental plasticity. As differentiation progresses, chromatin becomes increasingly specialised, with lineage-defining enhancers gaining active marks such as H3K27ac, while genes associated with alternative lineages are silenced through repressive marks like H3K27me3 (Bernstein, B. E. et al. 2005; Lara-Astiaso, D. et al. 2014).

In the haematopoietic system, stem and progenitor cells must balance self-renewal with the production of diverse mature blood cell types. Haematopoietic stem cells (HSCs) exhibit broad, permissive enhancer landscapes that allow rapid activation of multiple transcriptional programmes. As cells commit to specific lineages, these landscapes are progressively pruned and refined, stabilising lineage identity (Heinz, S. et al. 2010; Lara-Astiaso, D. et al. 2014). Disruption of these chromatin transitions can lead to aberrant gene expression and leukaemic transformation, as developmental pathways are co-opted or dysregulated (Fang, C. et al. 2020; Tsankov, A. M. et al. 2015).

Many leukaemia-associated mutations occur in chromatin-modifying enzymes or their regulators, leading to altered accessibility and inappropriate activation of self-renewal programmes (Fong, C. Y. et al. 2014; Cabal-Hierro, L. et al. 2020). Among these regulators, MLL has emerged as a key factor in both normal development and disease (Milne, T. A. et al. 2005a). These chromatin features collectively establish the regulatory landscape that determines where transcription factors and chromatin modifiers, such as MLL, can bind and function.

1.4 DNA Methylation and Hydroxymethylation

Because MLL binds preferentially to unmethylated DNA, cytosine modification provides a direct epigenetic constraint on its recruitment.

DNA methylation is a dynamic, reversible epigenetic modification involving the covalent addition of a methyl group to the 5' carbon of cytosine, most commonly at CpG dinucleotides, where it plays a central role in regulating genome function in mammals (Greenberg, M. V. C. and Bourc'his, D. 2019; Smith, Z. D. et al. 2025b) (Figure 1.2). This reaction is catalysed by the DNA Methyltransferase (DNMT) family. De novo methylation, which establishes new methylation patterns during development, is primarily carried out by DNMT3A and DNA Methyltransferase 3B (DNMT3B), while DNA Methyltransferase 1 (DNMT1) maintains existing patterns during replication by copying methylation marks to daughter strands (Li, E. and Zhang, Y. 2014). CpG-rich regions, known as CpG islands, are typically unmethylated at active promoters, whereas methylation at these regions is linked to transcriptional repression (Deaton, A. M. and Bird, A. 2011).

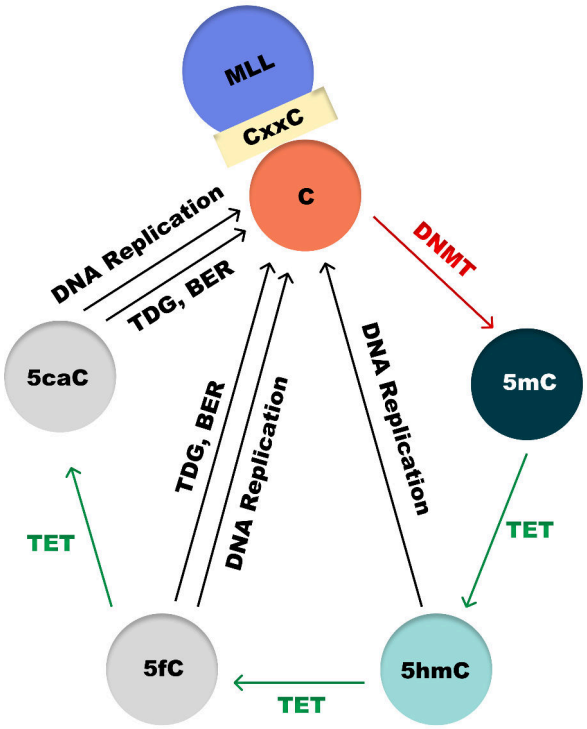


Figure 1.2: CpG methylation cycle and MLL binding. Cytosines within CpG dinucleotides can be methylated by DNA methyltransferases (DNMTs) to form 5-methylcytosine (5mC). TET enzymes oxidise 5mC to 5hmC and further to 5fC and 5caC. Thymine DNA glycosylase (TDG) recognises 5fC and 5caC and excises them, after which BER restores an unmodified cytosine. MLL binds to unmethylated CpGs via its CXXC domain, a process that is blocked by CpG methylation, while 5hmC is associated with dynamic, transcriptionally active regions that may facilitate MLL access. Adapted from (Wu, X. and Zhang, Y. 2017)

Methylation marks can be lost passively when DNMT1 fails to maintain them or actively removed by the Ten-Eleven Translocation (TET) family of enzymes. Ten-Eleven Translocation 1 (TET1), TET2, and Ten-Eleven Translocation 3 (TET3)

sequentially oxidise 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), then to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Tahiliani, M. et al. 2009; Wu, X. and Zhang, Y. 2017). These final oxidation products are excised by thymine DNA glycosylase (TDG) and replaced with unmodified cytosine through the Base Excision Repair (BER) pathway, fully restoring the original base (He, Y.-F. et al. 2011). This cyclical methylation-demethylation process enables precise regulation of gene expression and ensures developmental flexibility while maintaining genomic stability.

1.4.1 The Role of DNA methylation in gene regulation

DNA methylation influences transcription by modulating chromatin accessibility and TF binding, functioning as a key layer of epigenetic control. The roles of 5mC and 5hmC depend on both modification type and genomic context.

5-methylcytosine: canonical and context-dependent roles

Traditionally, 5mC is associated with transcriptional repression, particularly at promoters. Methylated CpGs can directly block TF binding when recognition motifs contain CpGs (Yin, Y. et al. 2017) and recruit methyl-CpG-binding domain (MBD) proteins such as MeCP2, MBD1, and MBD2, which assemble co-repressor complexes with histone deacetylases (HDACs) and chromatin remodellers, leading to compaction and stable silencing (Nan, X. et al. 1998; Baubec, T. et al. 2013). This underpins key processes such as genomic imprinting and X-chromosome inactivation, historically framing DNA methylation as a stable repressive mark (Bird, A. 2002).

However, more recent work suggests this model may be overly simplistic. Some TFs are insensitive to methylation, or even preferentially bind methylated CpGs, indicating that the impact of 5mC is protein-specific (Yin, Y. et al. 2017). Moreover, methylation is not always inversely correlated with expression. In certain cancers, including prostate cancer, hypermethylated promoters can paradoxically associate with increased transcription (Rauluseviciute, I. et al. 2020). Likewise, MeCP2, a classic methyl-binding protein, can act as either an activator or repressor depending on its interacting partners, such as CREB1 in neuronal tissue (Chahrour, M. et al. 2008). These findings show that 5mC acts as a context-sensitive signal, whose outcome depends on the local chromatin environment and available co-factors rather than being inherently repressive.

5-hydroxymethylcytosine: beyond a demethylation intermediate

5hmC was initially viewed as a transient demethylation intermediate, but it is now recognised as a stable, functional epigenetic mark in its own right (Tahiliani, M. et al. 2009; Wu, X. and Zhang, Y. 2017). Genome-wide mapping shows 5hmC enrichment at gene bodies, enhancers, and active regulatory elements, particularly in embryonic stem cells, neurons, and Haematopoietic Stem and Progenitor Cells (HSPCs) (Lister, R. et al. 2013; Pastor, W. A. et al. 2011; Yu, M. et al. 2012). This pattern links 5hmC to dynamic chromatin states and active transcription rather than simple demethylation.

Recent sequencing advances now distinguish 5mC from 5hmC at base resolution, revealing extensive 5hmC landscapes that were previously underestimated, especially in low-CpG or repetitive regions (Liu, Y. et al. 2019; Füllgrabe, J. et al. 2023). This has clarified that 5hmC is biologically significant, not merely a transient step in demethylation.

Functionally, 5hmC is read by distinct protein factors. While 5mC recruits repressive MBD proteins such as MeCP2, 5hmC is recognised by different readers, including MBD3 and hydroxymethylation-specific co-activators (Spruijt, C. G. et al. 2013). This reader switching mechanism actively counteracts repression, promoting open chromatin and transcription. In haematopoiesis, TET2 is highly expressed in HSPCs and generates 5hmC at enhancers of lineage-specific genes. Loss-of-function TET2 mutations, common in clonal haematopoiesis and myeloid malignancies, cause global 5hmC reduction, inappropriate enhancer hypermethylation, and repression of differentiation pathways, leading to aberrant self-renewal (Ko, M. et al. 2011; Rasmussen, K. D. et al. 2015; Cimmino, L. et al. 2017).

Thus, 5hmC plays a dual role: as a dynamic intermediate enabling active demethylation and as an independent regulatory signal marking transcriptionally active enhancers and gene bodies. This duality gives cells the flexibility to remodel epigenetic states while preserving lineage fidelity and developmental potential.

1.4.2 DNA Methylation in leukaemia

Aberrant DNA methylation is a hallmark of haematological malignancies, where disrupted cytosine modification patterns reshape transcription-factor binding and chromatin structure. Methylated CpGs can occlude factor binding sites, whereas unmethylated CpGs facilitate recruitment and activation, rendering methylation a direct regulator of transcriptional networks (Yin, Y. et al. 2017). MLL interfaces

with methylation via its CXXC domain, which recognises unmethylated CpG-island promoters (including HOX loci) and supports H3K4 methylation and transcriptional maintenance (Birke, M. 2002; Milne, T. A. et al. 2002).

In adult de novo AML, integrated genomic studies have established recurrent lesions in DNMT3A and TET2, alongside broader disruption of 5mC/5hmC homeostasis, underscoring DNA methylation pathways as a common axis of leukaemogenesis (The Cancer Genome Atlas Research Network 2013; Tyner, J. W. et al. 2018). These alterations reconfigure methylation landscapes and, indirectly, the accessibility and occupancy of transcriptional machinery.

Complementing this, paediatric Acute Lymphoblastic Leukaemia (ALL) studies demonstrate that genome-wide DNA methylation profiles carry independent prognostic information beyond cytogenetic subtype. In a cohort of 763 patients, methylation-based models improved risk stratification; notably, the mortality-risk predictor remained prognostic across validation cohorts and performed best in standard and infant risk groups, indicating that methylation state refines outcome prediction even in settings enriched for MLL fusions (Mosquera Orgueira, A. et al. 2024; Meyer, C. et al. 2023). In this light, methylation and hydroxymethylation do not merely correlate with transcriptional state but help define locus capacity for MLL recruitment: loss of an unmethylated CpG signal weakens canonical MLL targeting, whereas permissive methylation contexts can modulate the genomic engagement of both MLL and MLL fusion protein (MLL-FP) complexes.

1.5 The Problem of Binding Specificity

Despite strong enrichment of MLL at unmethylated CpG-island promoters, genome-wide maps show that only a subset of candidate CpG islands is occupied *in vivo* (Kerry, J. et al. 2017). In MLLr disease, fusion proteins display a reprogrammed occupancy pattern and sustained transcriptional activation at selected loci (Okuda, H. et al. 2014; Mueller, D. et al. 2009; Meyer, C. et al. 2023). The central question is therefore what distinguishes bound from unbound sites for both MLL and MLL-FP given that CpG density and promoter status alone are insufficient to explain selectivity.

1.5.1 Recruitment to target sites

As outlined earlier, MLL recruitment reflects coordinated inputs from DNA, chromatin, and transcription. Direct CXXC-mediated recognition of unmethylated CpG islands provides intrinsic specificity (Birke, M. 2002). Menin-LEDGF stabilises occupancy at active genes (LEDGF engaging H3K36me3 via its PWWP domain and DNA via AT-hooks) (Yokoyama, A. et al. 2002; Ayton, P. M. et al. 2004; Milne, T. A. et al. 2005a; Allen, M. D. et al. 2006). Coupling to the transcription apparatus, including PAF1C, links MLL to elongation-competent RNA Polymerase II, focusing H3K4 methylation at actively transcribed promoters (Milne, T. A. et al. 2010). These mechanisms concentrate activity at developmental loci yet still leave many CpG-island promoters unbound, motivating the analyses below.

1.5.2 Intrinsic sequence features and motifs

DNA sequence sets the baseline for recruitment, but CpG richness alone does not determine binding. Numerous CpG-island promoters with suitable sequence composition remain unoccupied (Kerry, J. et al. 2017), implying additional positive and negative sequence contexts—including local motif environments that favour or dissuade recruitment and higher-order features (e.g., nucleosome positioning signals) that interact with chromatin state. Sequence cues therefore operate alongside cofactors and cytosine modifications to produce selective occupancy.

1.5.3 The MLL complex

MLL acts within a multi-subunit complex whose WRAD core (WDR5, RBBP5, ASH2L, DPY30) is essential for full catalytic activity and efficient H3K4me3 deposition at active promoters (Milne, T. A. et al. 2002; Nakamura, T. et al. 2002; Dou, Y. et al. 2006). Recruitment integrates CXXC-CpG recognition and Menin-LEDGF tethering with PAF1C/Pol II interactions, thereby coupling promoter-proximal chromatin to transcriptional maintenance while avoiding promiscuous activation elsewhere (Birke, M. 2002; Yokoyama, A. et al. 2002; Milne, T. A. et al. 2005a; Milne, T. A. et al. 2010).

1.5.4 MLL fusion partners and the Super Elongation Complex

In MLL fusions, the N-terminal targeting modules (CXXC; Menin-LEDGF interfaces) are retained, while the C-terminal SET domain is replaced by a partner—

commonly AFF1/AF4, MLLT3/AF9, MLLT1/ENL, or ELL—that recruits the SEC and confers pause-release capability to tethered RNA polymerase II (Lin, C. et al. 2010; Yokoyama, A. et al. 2010; Okuda, H. et al. 2014). SEC constituents include AFF proteins (scaffolds), ELL (enhances processivity), and Positive Transcription Elongation Factor b (p-TEFb) (CDK9/Cyclin T), which phosphorylates the Pol II CTD to promote productive elongation. Constitutive SEC engagement at fusion-occupied loci shifts promoters from poised to persistently active states and underpins robust expression of oncogenic programmes (Okuda, H. et al. 2014). At a subset of targets, fusion occupancy spreads from promoter-proximal regions into gene bodies, accompanied by co-spreading of Menin, increased transcription, redistribution of H3K36me3, and pronounced gains in Histone H3 Lysine 79 Methylation (H3K79me) (both H3K79me2 and Histone H3 Lysine 79 Trimethylation (H3K79me3)); these spreading sites are molecularly distinct from super-enhancers and predict sensitivity to DOT1L inhibition (EPZ-5676) (Kerry, J. et al. 2017). Several fusion partners also recruit DOT1L, catalysing H3K79me within gene bodies, a modification associated with elongation and markedly enriched at fusion targets (Mueller, D. et al. 2009; Nguyen, A. T. and Zhang, Y. 2011). DOT1L activity contributes to a positive feedback loop that stabilises elongation and high transcriptional output; pharmacological DOT1L inhibition reduces HOXA9/MEIS1 expression and shows preclinical efficacy in MLLr models (Mueller, D. et al. 2009; Daigle, S. R. et al. 2011). Consistent with a functional requirement for this axis, perturbing H3K79me2/3 disrupts promoter-regulatory DNA communication and diminishes expression at fusion-occupied loci (Godfrey, L. et al. 2021).

1.5.5 Histone marks as signposts of regulatory activity

Chromatin modifications correlate with regulatory state and help predict (though do not by themselves determine) MLL occupancy patterns. H3K4me3 marks active promoters (and is catalysed by MLL/SET1 family complexes), H3K27ac delineates active regulatory elements, and H3K79me2/3 tracks elongation and is accentuated at fusion targets (Bernstein, B. E. et al. 2005; Creyghton, M. P. et al. 2010; Mueller, D. et al. 2009). These marks participate in feedback circuits: active marks facilitate factor recruitment and transcriptional stability, whereas repressive environments oppose binding. In MLL and fusion contexts, the combination of marks refines which loci are permissive versus refractory.

1.5.6 MLL binding and DNA methylation

MLL's CXXC domain binds unmethylated CpG and is inhibited by 5mC, making promoter methylation a potent veto on recruitment (Birke, M. 2002; Milne, T. A. et al. 2005a; Deaton, A. M. and Bird, A. 2011). 5hmC, generated by TET enzymes, is enriched at transcriptionally active, dynamic chromatin and is associated with contexts more permissive for factor engagement (Tahiliani, M. et al. 2009; Pastor, W. A. et al. 2011; Yu, M. et al. 2012; Lister, R. et al. 2013; Wu, X. and Zhang, Y. 2017). In fusion settings, retention of the CXXC module preserves a bias toward unmethylated DNA, while partner-mediated tethering (SEC/DOT1L) increases dependence on the cofactor and elongation environment, allowing occupancy at loci that MLL accesses less efficiently (Lin, C. et al. 2010; Okuda, H. et al. 2014; Mueller, D. et al. 2009).

1.5.7 Towards clarity of MLL recruitment logic

Selective targeting of MLL and MLL-FP emerges from the integration of multiple regulatory layers that each encode information about local chromatin context. Understanding how these layers interact clarifies why MLL binds a subset of CpG-island promoters, and how fusion proteins subvert this logic to activate inappropriate transcription in leukaemia as illustrated in Figure 1.3.

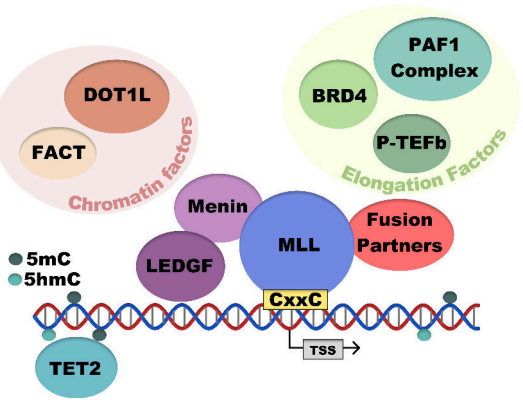


Figure 1.3: Conceptual model of MLL recruitment and regulation. MLL binds to unmethylated CpG-rich regions of DNA through its CXXC domain, with Menin and LEDGF stabilising this interaction. TET2 oxidises 5mC to 5hmC, promoting a dynamic and transcriptionally permissive state that facilitates MLL access. MLL acts as a scaffold for recruiting chromatin factors such as DOT1L and FACT, as well as elongation factors including P-TEFb, BRD4, and the PAF1 complex. In MLL-rearranged leukaemia, fusion partners bring additional activation domains, redirecting this regulatory network to drive inappropriate gene expression and leukaemogenesis. Adapted from (Smith, A. et al. 2025a).

At any given locus, MLL occupancy reflects non-linear interactions among sequence features, chromatin marks/co-factors, and cytosine modification. In normal cells, these layers cooperate to confine MLL to a subset of CpG-island promoters of developmental genes (Birke, M. 2002; Milne, T. A. et al. 2005a; Bernstein, B. E. et al. 2005). In MLLr disease, partner-mediated coupling to elongation machinery (SEC, DOT1L) reprogrammes occupancy and sustains high output at selected loci (Mueller, D. et al. 2009; Lin, C. et al. 2010; Okuda, H. et al. 2014). Prior studies typically examine a single layer at a time, identifying CpG-linked sequence features (Milne, T. A. et al. 2005a), linking MLL to specific histone marks and co-factors (Bernstein, B. E. et al. 2005; Tie, F. et al. 2009; Godfrey, L. et al. 2019), or demonstrating that 5mC blocks and 5hmC associates with permissive recruitment (Tahiliani, M. et al. 2009; Pastor, W. A. et al. 2011; Yu, M. et al. 2012); the joint relationships remain challenging to resolve by descriptive analysis alone.

1.6 Machine Learning Approaches in Genomics

1.6.1 From Statistical Models to Deep Learning and Beyond

The increasing availability of high-throughput genomic data has transformed our ability to study transcriptional regulation, but it has also created new challenges. Traditional statistical approaches, such as linear regression or motif enrichment analyses, have been invaluable for identifying simple relationships between DNA sequence features and binding events. However, these methods are limited in their ability to capture the non-linear, context-dependent interactions that characterise complex regulatory systems.

Machine Learning (ML) methods have emerged as powerful tools to address these limitations. Early applications in genomics used algorithms such as support vector machines (SVMs) and random forests, which could model non-linear relationships and handle large feature spaces (Libbrecht, M. W. and Noble, W. S. 2015). These approaches improved predictive accuracy compared to purely statistical models, particularly when dealing with heterogeneous datasets such as Chromatin Immunoprecipitation with Sequencing (ChIP-seq) or ATAC-seq.

The advent of deep learning marked a further leap forward. Convolutional neural networks (CNNs) enabled direct learning from raw DNA sequence data, automatically extracting motif representations and higher-order features without the need for manual engineering (Alipanahi, B. et al. 2015). However, CNNs are inherently lim-

ited by their fixed receptive fields, making it difficult to capture long-range genomic interactions.

Transformers replaced fixed receptive fields with self-attention mechanisms, enabling the model to consider relationships between any positions in the input. In genomics, transformers have rapidly been adopted for tasks ranging from transcription factor binding prediction to gene expression modelling because of their ability to learn complex dependencies directly from raw DNA sequence (Dalla-Torre, H. et al. 2024; Consens, M. E. et al. 2025).

Pretrained DNA language models (DNA-LMs) apply masked language modelling objectives to genomic data, learning rich internal representations of sequence patterns without explicit supervision (Dalla-Torre, H. et al. 2024). Fine-tuning adapts a pretrained DNA-LM to specific prediction tasks (here, MLL and MLL fused to AF4 (MLL-AF4) binding) by attaching a prediction head and updating parameters with labelled examples, often using smaller learning rates for pretrained layers and partial freezing for efficiency (Howard, J. and Ruder, S. 2018; Consens, M. E. et al. 2025).

To reduce compute, parameter-efficient fine-tuning (PEFT) such as LoRA adapts a small subset of parameters while keeping the pretrained model largely frozen (Houlsby, N. et al. 2019; Hu, E. J. et al. 2021). In this thesis, Low-Rank Adaptation inserts small trainable matrices into attention layers so that only these adapters and the classification head are trained, reducing memory while retaining the benefits of a large pretrained DNA-LM.

Not all genomic data are raw sequence. Many datasets are region-level, tabular features (e.g., ChIP/CUT&Tag intensities, accessibility, methylation fractions). For these, gradient-boosted decision trees (GBDTs) such as eXtreme Gradient Boosting (XGBoost), LightGBM, and CatBoost remain highly effective. Recent deep-tabular architectures (e.g., attention/gating/masking-based networks such as Gated Adaptive Network for Deep Automated Learning of Features for Tabular Data (GANDALF)) learn sparse feature interactions end-to-end (Joseph, M. and Raj, H. 2022). As a practical baseline, a calibrated GBDT is employed for tabular modalities, reserving deep-tabular architectures for settings with pronounced combinatorial feature structure or multi-task setups.

1.6.2 Choosing the right model for the data

The most reliable gains come from matching model class to data type and question. Model size and context help, but data curation and interpretability choices are equally important for trustworthy conclusions (Consens, M. E. et al. 2025).

The recruitment of MLL and its fusion proteins reflects the interplay of sequence, cofactors, and cytosine modifications. An integrated multimodal model is attractive but presents computational and experimental challenges. Instead, this thesis applies three independent ML frameworks, each optimised for a specific regulatory layer:

1. Sequence-based model, using transformer architectures, to capture the DNA-encoded rules of MLL and MLL-AF4 binding.
2. Cofactor-based model, using statistical and ensemble methods, to model the contribution of transcription factors, histone marks, and chromatin proteins.
3. Methylation-based model, using gradient-boosted decision trees and interpretability techniques, to assess how patterns of 5mC and 5hmC shape MLL binding across the genome.

Training separate models for each modality applies the most appropriate strategy while avoiding premature integration. The results can then be interpreted side-by-side to build a comprehensive picture of how sequence features, cofactor networks, and methylation landscapes collectively determine binding specificity.

1.6.3 Model Interpretability

As ML models grow in complexity, understanding why they make particular predictions becomes increasingly important. Interpretability is critical in biological applications, where the goal is not only to predict outcomes but also to generate mechanistic insights. For transformer-based sequence models, Layer Integrated Gradients (LIG) assigns importance scores to individual input bases (Sundararajan, M. et al. 2017). This enables identification of key motifs and sequence features that drive predictions. For tabular models, SHapley Additive exPlanations (SHAP) quantify feature importance and interactions (Lundberg, S. and Lee, S.-I. 2017). Together, these tools bridge prediction and understanding, allowing the models to serve as both analytical tools and hypothesis generators.

1.7 Thesis Aim and Objectives

The work presented in this thesis is motivated by a central hypothesis: MLL binding specificity emerges from the interplay of intrinsic DNA sequence features, local chromatin environment, and DNA methylation state. By dissecting each layer independently, it is possible to reconstruct the logic by which MLL and its fusion partners target specific genomic loci.

To explore this hypothesis, three complementary aims were pursued:

1. Which intrinsic DNA sequence features distinguish genomic regions that recruit MLL from those that do not?
2. To what extent do local chromatin states and cooperating factors explain MLL recruitment beyond sequence alone?
3. How do 5mC and 5hmC distributions modulate accessibility and co-factor engagement to permit or preclude MLL binding?

By addressing these aims, this thesis seeks to provide a comprehensive, multi-layered understanding of MLL recruitment in both normal and leukaemic contexts.

2 Materials and Methods

2.1 Cell Culture

All cell lines were grown in a 37 °C humidified incubator with 5% CO_2 and were confirmed to be mycoplasma free.

2.1.1 Culturing THP-1, SEM, RCH-ACV and RS4;11 cell lines

SEM cells, a MLL-AF4 B-cell ALL line (Greil, J. et al. 1994), were purchased from DSMZ (www.dsmz.de), and cultured in Iscove's modified Dulbecco's medium (IMDM, Gibco) with 10% FCS (Gibco) and $1\times$ GlutaMAX (Gibco). Cell density was maintained between 5×10^5mL and 2×10^6mL. RS4;11 and THP-1 cells were purchased from ATCC (www.atcc.org) and cultured in RPMI-1640 (Gibco) with 10% FBS and $1\times$ GlutaMAX. Cell density was maintained between 5×10^5mL and 1.5×10^6mL.

2.1.2 Patient samples

Patients ALL samples including infants (diagnosed at < 1 year) and children (1-18 years) were obtained from the Blood Cancer UK Childhood Leukaemia Cell Bank (now VIVO Biobank, UK; ethics approval REC: 23/EM/0130) and from Our Lady's Children's Hospital, Crumlin, Ireland (ethics approval REC: 21/LO/0195). Informed consent was secured from all participants or from those with parental responsibility.

2.2 Chromatin Immunoprecipitation Methods

2.2.1 ChIPmentation

ChIPmentation was performed as previously described (Crump, N. T. et al. 2023). 10 µL protein A-coupled magnetic beads were incubated with 1 µL primary antibody (Table A.4) for 4 h at 4 °C in PBS with 0.5% BSA and protease inhibitors. Cells were fixed by incubation with 2 mM DSG for 30 min at room temperature followed by 1% formaldehyde for 30 min at room temperature. Samples were lysed in 120 µL Lysis buffer (50 mM Tris-HCl pH 8.0, 0.5% SDS, 10 mM EDTA), 1×protease inhibitor cocktail (Merck) and sonicated to 200-300 base pairs (bp) fragments using a Covaris sonicator (ME220) before neutralisation with 1% Triton-X100 before pre-clearing with 5 µL protein A dynabeads. Antibody-coupled beads were washed, combined with pre-cleared chromatin, and incubated overnight at 4 °C. Immunoprecipitated chromatin was washed (×3) with RIPA buffer (50 mM HEPES-KOH pH 7.6, 500 mM LiCl, 1 mM EDTA, 1% NP40 and 0.7% Na deoxycholate), then washed with Tris-EDTA and 10 mM Tris-HCl pH 8.0. Chromatin was tagmented by adding 29 µL Tagmentation Buffer (10 mM Tris-HCl pH 8.0, 5 mM MgCl2, 10% DMF) and 1 µL TDE1 (Illumina) incubating for 5 min at 37 °C and stopping with 150 µL RIPA buffer. Beads were washed, and tagmented DNA amplified with NEBNext Ultra II Q5 Master Mix and indexed primers (thermal profile: 72 °C 5 min, 95 °C 5 min, (98 °C 10 s, 63 °C 30 s, 72 °C 3 min) × 12, 72 °C 5 min). Libraries were cleaned with Agencourt AMPure XP beads and sequenced on a NovaSeq X platform (Azenta) with 150 bp paired-end reads.

2.2.2 Cleavage Under Targets and Tagmentation (CUT&Tag)

CUT&Tag was performed largely as described (Kaya-Okur, H. S. et al. 2019), with the following modifications to maximise yield. Up to 5×10^4 cells per sample were pelleted at 300 × g for 10 mins at RT before resuspending in 1 ml NE1 buffer (20 mM HEPES KOH pH 7.9, 10 mM KCl, 0.5 mM spermidine, 0.1% Triton X-100, 20% glycerol, 1× Complete Protease Inhibitor [EDTA-free]) and incubating on ice for 10 mins. Nuclei were pelleted at 500 × g for 10 min at 4 °C before resuspending in 0.1% formaldehyde and incubating at RT for 2 mins, followed by quenching with 78 mM Glycine and pelleting the nuclei at 500 ×g for 10 min at 4 °C. Nuclei were resuspended in Wash Buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 1× Roche Complete Protease Inhibitor [EDTA-free]) and counted by trypan

blue staining. Concanavalin A (Con A) beads were activated by washing 2× with 1 ml Concanavalin Activation Buffer (20 mM HEPES-KOH pH 7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$). For each antibody/sample combination, 5 µl of activated ConA beads were added to 5×10^4 nuclei. Nuclei were incubated with 0.5 µl of primary antibody (antibodies listed in Table A.3) in 25 µl of antibody buffer (0.1% BSA, 2 mM EDTA in Wash Buffer) for either 1 hr at RT or overnight at 4 °C. Nuclei were then incubated with 25 µl of 1:100 secondary antibody (Guinea Pig anti-Rabbit IgG) for 30 mins at RT. Nuclei were washed with 200 µl Wash Buffer and incubated with 25 µl 1:20 pAG-Tn5 (EpiCypher) in WB-300 (300 mM NaCl in Wash Buffer) for 1hr at RT. Nuclei were washed (x2) with 200 µl of WB-300 to remove unbound pAGTn5 and then incubated at 37 °C for 1hr in 50 µl Tagmentation buffer (10 mM MgCl2 in WB300). Tagmentation was halted by washing with 50 µl TAPS Wash Buffer (10 mM TAPS pH 8.5, 0.2 mM EDTA) before lysing the nuclei with 5 µl SDS-ProK Buffer (1% SDS, 10 mM TAPS pH 8.5, Thermolabile Proteinase K) and incubating for 1 hr at 37 °C and 1hr at 58 °C. SDS was neutralised with 15 µl 6% Triton and tagmented DNA amplified with NEBNext Ultra II Q5 Master Mix and indexed primers (thermal profile: 58 °C 5 min, 72 °C 5 min, 98 °C 5 min, [98 °C 10 sec, 60 °C 10 sec, 72 °C 1 min] ×12). Libraries were cleaned with Agencourt AMPure XP beads and sequenced on NovaSeq X platform (Azenta) with 150 bp paired-end reads.

2.3 Simultaneous 5mC, 5hmC, and gDNA sequencing

2.3.1 gDNA Extraction, Fragmentation and Quantification

Cultured cell line Genomic DNA (gDNA) was extracted using the Monarch gDNA Purification Kit (NEB catalog no. T3010) according to the manufacturer's instructions. For patient and xenograft samples, 1×10^6 cells were resuspended in 500 µL digest buffer (200 mM NaCl, 10 mM Tris-HCl (pH 7.5), 2 mM EDTA and 0.2% SDS) with 10 µL Proteinase K (Thermo) and incubated at 50 °C for 2 h. Subsequently, 500 µL of a 1:1 phenol:chloroform mixture was added, the sample vortexed for 30 s, and centrifuged at maximum speed for 2 min. The upper aqueous phase was transferred to a fresh tube, and 50 µL sodium acetate, 1 mL 100% ethanol, and 1 µL GlycoBlue were added. After gentle inversion, the mixture was placed on dry ice for 10 min, then centrifuged at maximum speed at 4 °C for 10 min. The supernatant was discarded, and the pellet washed with 500 µL 70% ethanol, followed by centrifugation at maximum speed at 4 °C for 5 min. After air drying for 2 min, the pellet was resuspended in 15 µL nuclease-free water, vortexed, and incubated at 50 °C for

5 min and then overnight at 4 °C to ensure complete dissolution. Extracted gDNA was quantified using a Qubit Fluorometer (Life Technologies) and fragmented using a Covaris sonicator (ME220) at a peak power of 14, a duty factor of 20%, and 1000 cycles per burst for 100 s, resulting in an average fragment size of approximately 250 bp. The sonicated DNA was quantified using a Qubit dsDNA HS assay (Thermo Fisher) to measure gDNA concentration ($1.4\pm$ 1.29 ng/µL) and DNA D1000 Screen Tape (Agilent 2200 TapeStation system) to assess fragment size distribution (255 ± 28.4 bp).

2.3.2 Library Preparation

Library preparation for 6-letter sequencing of 5mC and 5hmC was performed using the Duet evoC Library Preparation Kit, following the manufacturer's instructions (Biomodal catalog no. 6205) (Füllgrabe, J. et al. 2023). For each sample, 34.7 ± 30.9 ng of sonicated gDNA was mixed with a spike-in control at a concentration of 0.05 ng/µL. The spike-in control consisted of methylated bacteriophage lambda DNA, a synthetic hydroxymethylated oligonucleotide (to assess the sensitivity of 5mC and 5hmC conversion), and an unmethylated pUC19 DNA plasmid vector (to assess conversion specificity). The DNA then underwent end repair and A-tailing, followed by ligation and subsequent digestion of hairpin adapters. Next, the original strand was copied to the newly synthesised strand, and forkhead adapters were ligated. Methylation at 5mC sites was enzymatically copied to the complementary strand, whereas 5hmC was glycosylated to prevent copying. Modified cytosines were protected by oxidation, and unmodified cytosines were deaminated to uracil. Finally, the resulting DNA libraries were amplified (thermal profile: 98 °C for 30 s; then 9 cycles of [98 °C for 10 s, 62 °C for 30 s, 65 °C for 1 min]; followed by 65 °C for 5 min), and quantified using a Qubit dsDNA HS assay (Thermo Fisher) (17.4 ± 8.48 ng/µL) and DNA D5000 Screen Tape (Agilent 2200 TapeStation system) to assess fragment size distribution (582 ± 56.0 bp). The amplified libraries were then pooled.

2.3.3 Methylome Enrichment and Sequencing

To increase sequencing depth while not increasing cost, equimolar pooled libraries were enriched for methylated regions using the Twist Human Methylome Panel (Twist Bioscience catalog no. 105517) using the Twist Fast hybridisation and Wash Kit (Twist Bioscience catalog no. 101278) as per manufacturer's instructions. The pooled libraries mixed with methylation pre-hybridisation solution were dried using

a SpeedVac vacuum concentrator, and then re-suspended in Fast hybridisation Mix. Immediately, hybridisation Enhancer was added, and the mixture was transferred to a preheated (95 °C) thermal cycler. The reaction was denatured at 95 °C for 5 minutes and then incubated at 60 °C for 16 hours. After hybridisation, the reaction was transferred to pre-cleared streptavidin binding beads and incubated at room temperature for 30 min, the beads were washed to remove non-specific binding. Enriched libraries were then eluted and amplified (thermal profile: 98 °C for 45 s; then 9 cycles of [98 °C for 15 s, 60 °C for 30 s, 72 °C for 30 s]; followed by a final extension at 72 °C for 1 minute). The amplified libraries were purified and quantified using a Qubit dsDNA HS assay (Thermo Fisher) (5.4 ng/µL) and D1000 Screen Tape (Agilent 2200 TapeStation system) (534 bp) prior to sequencing. High-throughput sequencing was carried out on a NovaSeq X platform (Azenta) with 150 bp paired-end reads. The fastq files were resolved using Biomodal's Duet pipeline (version 1.4.1) (Füllgrabe, J. et al. 2023). The sensitivity of 5mC and 5hmC conversion was assessed by counting the proportions of 5mC CpGs, 5hmC CpGs, and unmodified CpGs in reads mapped to the spiked-in genomes.

2.4 Whole Genome Sequencing

2.4.1 Library Preparation and Sequencing

gDNA was extracted as described above and quantified using a Qubit Fluorometer (Life Technologies). 2 ng of genomic DNA was incubated at 55°C for 15 minutes with 0.4 ul TDE1 (Illumina) before purification (Qiagen MinElute PCR purification kit). Libraries were indexed and sequenced on NovoSeq X platform (Azenta) with 150 bp paired-end reads.

3 SeqNado: Uniform and Reproducible Data Processing for ML-Ready Outputs

3.1 Introduction

Modern genomic experiments generate vast amounts of sequencing data that require extensive preprocessing before they can be analysed or integrated into ML models. Commonly used workflows often consist of ad-hoc scripts with little error handling, making them prone to crashing, data corruption, and difficulties when adding new samples. These pipelines are typically assay-specific, hard-coded, and difficult to customise or scale, especially across High-Performance Computing Cluster (HPC) and cloud environments. They can also be challenging to install or maintain, with complex dependencies and inconsistent output structures. As a result, researchers face significant barriers to consistent, high-throughput, and storage-efficient data processing, slowing both experimental progress and downstream analyses.

While modern workflow languages such as Nextflow, Snakemake, and Workflow Description Language (WDL) have improved error handling, reproducibility, and scalability, many publicly available pipelines remain assay-specific and rigidly structured. These workflows can be difficult to customise, often require substantial local storage for intermediate files, and may lack features such as seamless checkpointing or dynamic resource scaling. As a result, researchers frequently face challenges when integrating diverse datasets or adapting pipelines for new experimental designs, particularly when working across multiple assay types or preparing uniform outputs for downstream analysis and ML applications.

SeqNado was developed with two main goals, firstly to make a pipeline to process the data required for the ML tasks in this project in a uniform and reproducible

manner, secondly as a tool that could be used by others in the lab and beyond to process their data for downstream analysis with little to no prior bioinformatics experience and minimal assistance.

With goal one in mind, the pipeline is built to be efficient and robust with the ability to handle multiple samples with speed and minimal storage footprint resulting in an output directory structure that is consistent and producing the input required directly for the input for training, finetuning and inference tasks.

With goal two in mind, the pipeline was built to take the user from initialisation and configuration through experimental design to fully processed data with as few commands as possible. It was built to handle the many different assay types used in the lab. The configuration was required to be easy to understand and navigate while giving the user control over tool parameters. Configuration is done through an interactive questionnaire to flexibly accommodate the requirements of the user. Output data was kept in a consistent and logical structure and allowing the data to be visualised in UCSC sessions, prepared for GEO submission or used in downstream analysis with ease.

Given these complexities, SeqNado was developed to streamline this data processing using compute and storage efficiently to preprocess genomic datasets in minimal time using minimal resources. It was extensively tested and expanded to include the pre-processing of many assay types including ATAC-seq, ChIP-seq, CUT&Tag, RNA-seq, methylation including TAPS and Whole Genome Bisulphite Sequencing (WGBS), short-read Whole Genome Sequencing (WGS) including variant calling, Micro Capture-C (MCC), and CRISPR screens.

SeqNado has been extensively tested and optimised for use on the CCB with Slurm. It is also deployable on a laptop or on Amazon Web Services Elastic Compute Cloud (AWS EC2) with minimal setup. This ensures active instance time is spent on computation rather than on configuring third-party pipelines of uncertain maintenance. In this thesis, SeqNado allowed me to achieve consistent processing of multiple data types and their incorporation into the ML models and downstream analysis discussed in subsequent chapters.

The remainder of this chapter covers: (i) the motivation and rationale behind SeqNado's design and implementation; (ii) the workflow-language choice; (iii) SeqNado's architecture; and (iv) the user workflow. I then benchmark SeqNado against popular alternatives and conclude with a brief case study applying it to my dataset.

3.2 Pipeline framework and workflow language

3.2.1 Pipelines

A bioinformatics pipeline is an automated sequence of computational steps that converts raw sequencing data into analysis-ready outputs. Typical stages include quality control, alignment to a reference genome, duplicate removal, normalisation, and generation of summary statistics or signal files (e.g., bigWig, count matrices, peaks). Pipelines are essential for reproducibility, efficiency, and consistency across large datasets; by automating tool execution and dependency management, they enable scalable processing with minimal user intervention and error.

In practice, many groups still rely on mixtures of legacy scripts and assay-specific workflows, which can be fragile and inefficient. Even with modern workflow languages, implementations are often hard to customise and maintain. Common pain points include insufficient error handling (single-task failures abort entire runs), lack of checkpointing (adding samples forces full re-runs), hard-coded parameters, retention of all intermediates (excessive storage), and complex installations with dependency conflicts across HPC and cloud environments. These issues waste compute, risk corrupted outputs, and hinder reproducible, large-scale analyses—especially when integrating multiple assay types or preparing machine-learning-ready data.

3.2.2 Workflow languages

Modern workflow languages including Nextflow, WDL/Cromwell, and Snakemake solve much of the fragility associated with ad-hoc scripting by formalising task order, capturing software environments, and enabling robust, restartable execution at scale which are all important for reproducibility (Di Tommaso, P. et al. 2017; Voss, K. et al. 2017; Mölder, F. et al. 2021). Each of these workflow languages takes a slightly different approach. Nextflow couples a powerful execution engine with strong HPC/cloud parallelism and first-class container support; it excels for large, streaming, multi-sample workloads but can feel prescriptive when fine-tuning step-level behaviour or trimming intermediate files. WDL (executed by Cromwell) is widely adopted in consortium settings, with clear separation between workflow specification and runtime; in practice, JSON-heavy configuration and backend tuning can raise the barrier to lightweight customisation. Snakemake combines a readable Python rule syntax with native Conda/Apptainer integration and straightforward

embedding of small utilities. It is particularly well suited to lab workflows that evolve, need step-level transparency, and require easily adjustable workflows as experimental designs change.

Ultimately, we chose Snakemake because it (i) recovers cleanly from failures by saving progress and writing outputs atomically—if a rule doesn't finish, any partial files are removed; (ii) automatically tracks and deletes intermediate files as soon as they're no longer needed, keeping storage compact; (iii) defines each step as a clear, editable rule so parameters or even whole tools can be swapped at run time via the configuration, without rewriting the pipeline; and (iv) is based on Python, which was familiar and easy to extend with small helper scripts. This combination gave us the flexibility to support heterogeneous assays (currently; ATAC-seq, ChIP-seq, CUT&Tag, RNA-seq, MCC, WGS, methylation, or CRISPR) while still producing uniform, ML-ready outputs with consistent environments across HPC clusters, cloud backends, and laptops. In short, Snakemake provided the most readable and adaptable substrate for SeqNado's design goals of reproducibility and efficiency with the least friction for day-to-day research use.

3.3 SeqNado

3.3.1 SeqNado Implementation

SeqNado is organised as a set of small, well-defined rules such as trimming FASTQ files or mapping reads using bowtie2 or STAR, that can be combined for different assay types. This modularity makes it easy to add a new assay or change a tool without disturbing the rest of the pipeline. To keep results consistent across machines and keep initial installation and configuration simple and quick, all software needed for each step is bundled inside Apptainer containers (self-contained software packages).

Because the pipeline will evolve with additional assay types in particular, we built in automated tests that run every time the code is updated (via GitHub Actions). These tests use representative subset datasets to check that the three main commands seqnado-config, seqnado-design, and the main SeqNado run, successfully produce the expected outputs for each assay. The tests run in parallel by command and assay type reducing testing time.

For reproducibility, releases use clear version numbers (semantic versioning), and packages are published to Anaconda and PyPI. Container images are pinned to exact versions to prevent silent dependency changes. Each seqnado configuration file also contains the SeqNado version that was used for the run, so past runs can be re-run as samples are added with uniform outputs.

Together, these features keep SeqNado easy to extend to add new assays or tools, reliable with automatic checks on every change, and reproducible with documented versioning and containerisation.

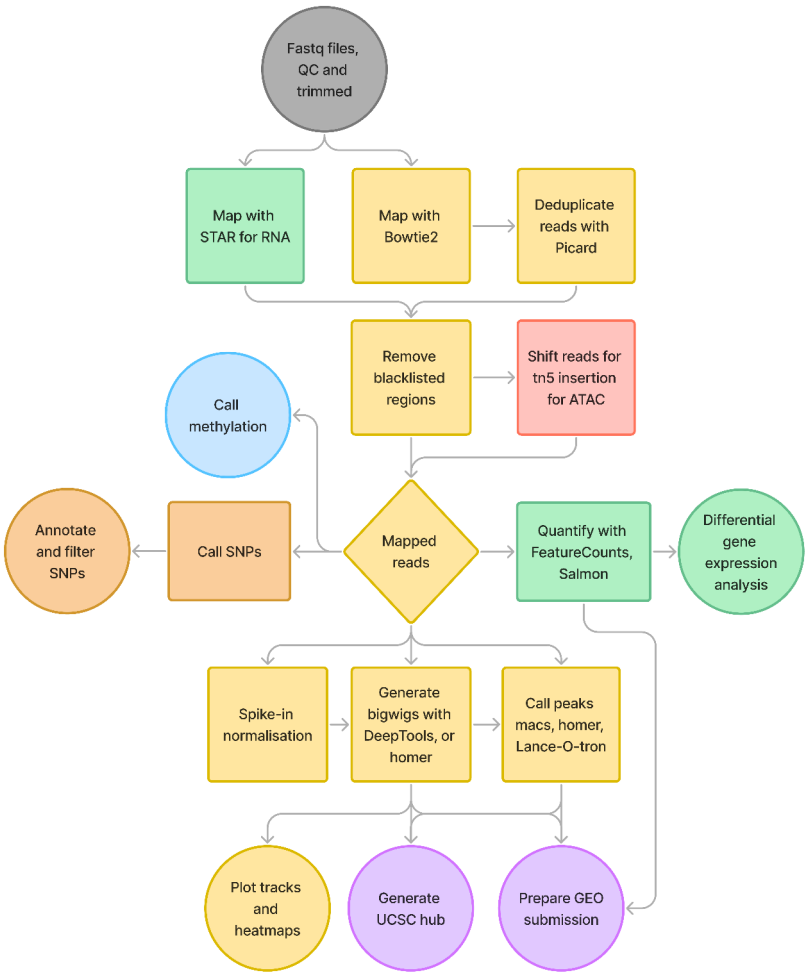


Figure 3.1: Overview of SeqNado workflow from initialisation to processed data outputs.

3.3.2 User Workflow

Installation

SeqNado was designed to be straightforward to install and run, requiring only four main commands to process raw data into ready-to-use outputs The pipeline can be installed from either PyPI or mamba, with all dependencies handled automatically through containers or Conda environments: mamba install -c bioconda seqnado or pip install seqnado

Step 1: Initialisation (one-time)

The first step is to initialise SeqNado by setting up all required reference genomes and paths using:

```
seqnado-init
```

This produces a configuration JSON file containing paths to reference genomes It only needs to be run once per environment, though the file can be edited later to include multiple reference genomes if required (Figure 3.2).

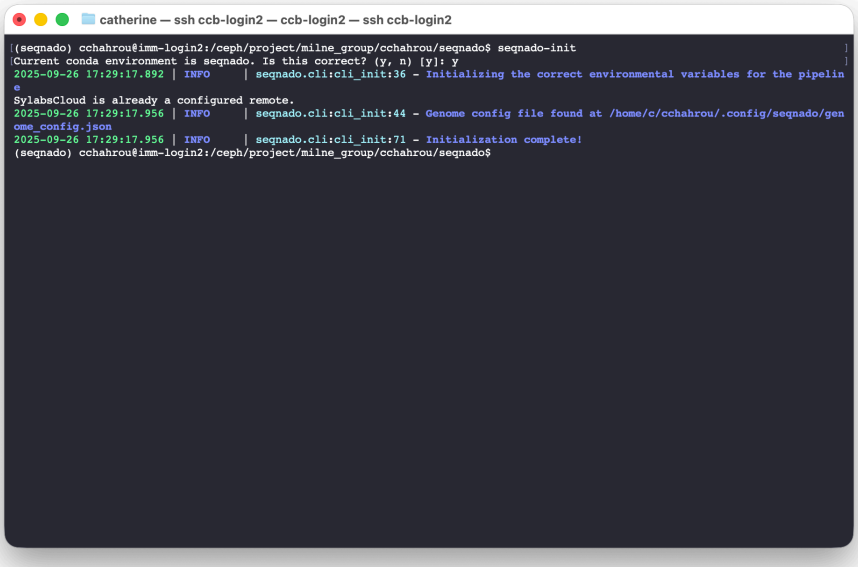


Figure 3.2: Example SeqNado initialisation.

Step 2: Configure the run

The next step is to create a configuration for a specific assay type (e.g., ATAC-seq, ChIP-seq, CUT&Tag, RNA-seq, MCC, WGS, methylation, or CRISPR) This is done with:

```
seqnado-config <assay>
```

SeqNado uses a Jinja-backed interactive questionnaire to guide the user through all available options, with sensible defaults provided for each assay type (Figure 3.3). These include settings such as whether to generate bigWigs, which peak caller to use, mapping parameters, BAM filtering thresholds, and bigWig resolution.

The output is an ISO date-stamped directory containing: A YAML configuration file, which can be edited to fine-tune individual tool parameters. A placeholder fastq/ directory into which FASTQ files are symbolically linked.

FASTQ naming convention:

For most assays:

- sample-replicate_R1.fastq.gz
- sample-replicate_R2.fastq.gz

For ChIP-seq and CUT&Tag: include the antibody or input, e.g.

- sample-replicate_MLL-N_R1.fastq.gz
- sample-replicate_input_R1.fastq.gz

This standardised naming scheme ensures that SeqNado can automatically detect sample type and experimental structure.

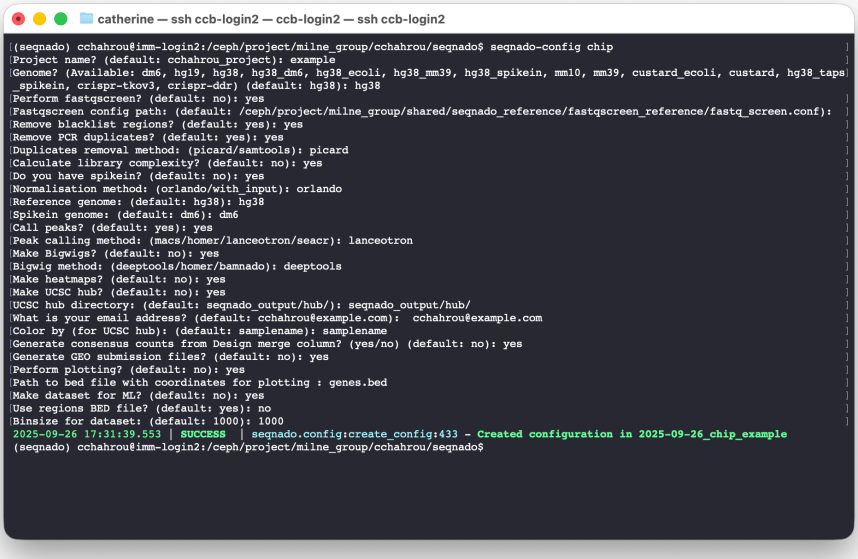


Figure 3.3: Example SeqNado configuration questionnaire for ChIP-seq.

Step 3: Generate the design file

Once the FASTQ files are linked, the following command generates a design CSV file containing all required metadata:

```
seqnado-design <assay>
```

This file includes sample names, input controls (if relevant), file paths, and information on whether data are single-end or paired-end For ATAC-seq, ChIP-seq, and CUT&Tag, an optional merge column can be added to control how outputs are combined:

- **merge=consensus**: merge all samples for a single consensus result.
- **merge=<replicate or IP>**: merge by replicate or by immunoprecipitation group.

The design file ensures that all sample relationships are clearly defined before processing begins. The design is also extendable to include additional metadata as required in the future.

Step 4: Run the pipeline

Finally, the pipeline is executed using:

```
seqnado <assay> --cores <n_cores> --preset <ls|lc|ss>
```

The `--cores` option specifies the maximum number of Central Processing Unit (CPU) cores to use across all parallel jobs.

The `--preset` option selects the execution environment: Where the options are:

- **ls** = local singularity
- **lc** = local conda
- **ss** = SLURM singularity (HPC batch mode)

Local singularity (ls) is suitable for laptops or workstations with Apptainer installed, while local conda (lc) is for systems without Apptainer but with Conda available. Both of these options run jobs in parallel on the local machine, up to the specified core limit, including on AWS EC2 instances. The SLURM singularity (ss) preset automatically submits parallel jobs per sample and step, optimising resource use on HPC systems

For deeply sequenced data, compute resources for individual steps can be increased using:

```
-s <int>
```

This option scales up memory and CPU for memory-intensive steps like alignment or peak calling, while keeping lightweight steps efficient.

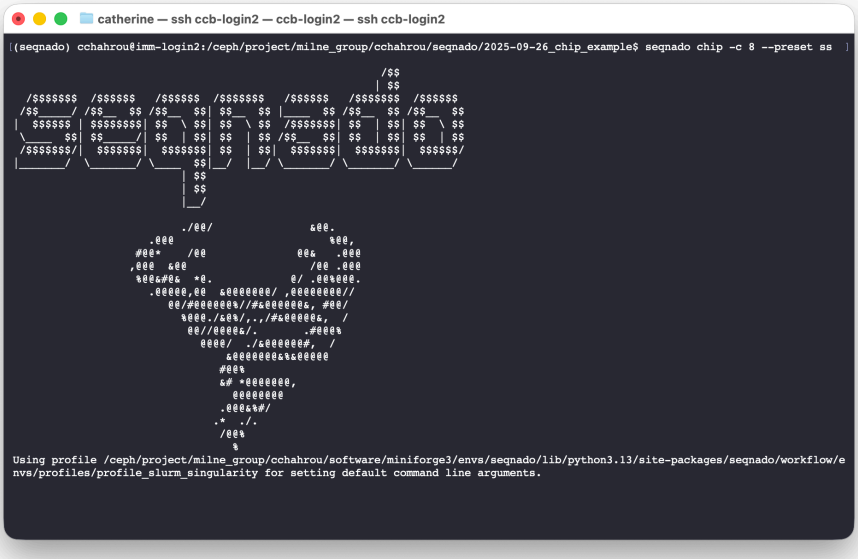


Figure 3.4: Example SeqNado run.

This triggers the entire processing workflow, building the snakemake DAG and executing it from raw FASTQ files to final outputs, with automatic error handling, checkpointing, and intermediate file management (Figure 3.4).

Outputs

SeqNado produces a uniform, organised output directory, automatically cleaning up intermediate files once they are no longer needed (Figure 3.5). This keeps storage requirements low while maintaining all final outputs and logs. Standard outputs include:

- BAM files (aligned, filtered reads)
- bigWig files (signal tracks, optionally reference-normalised)
- Peak files (e.g., LanceOtron, MACS2, SEACR, or Homer)
- Count tables (e.g., for differential analysis)
- Variant VCFs (for WGS workflows)
- Methylation calls (bedGraph or methylKit format)

- ML–ready `h5ad` files (optional, for ATAC-seq/ChIP-seq/CUT&Tag data)
- Additional outputs:
  - UCSC hubs for visualisation
  - GEO submission bundles, including `md5` checksums and a pre-filled metadata template

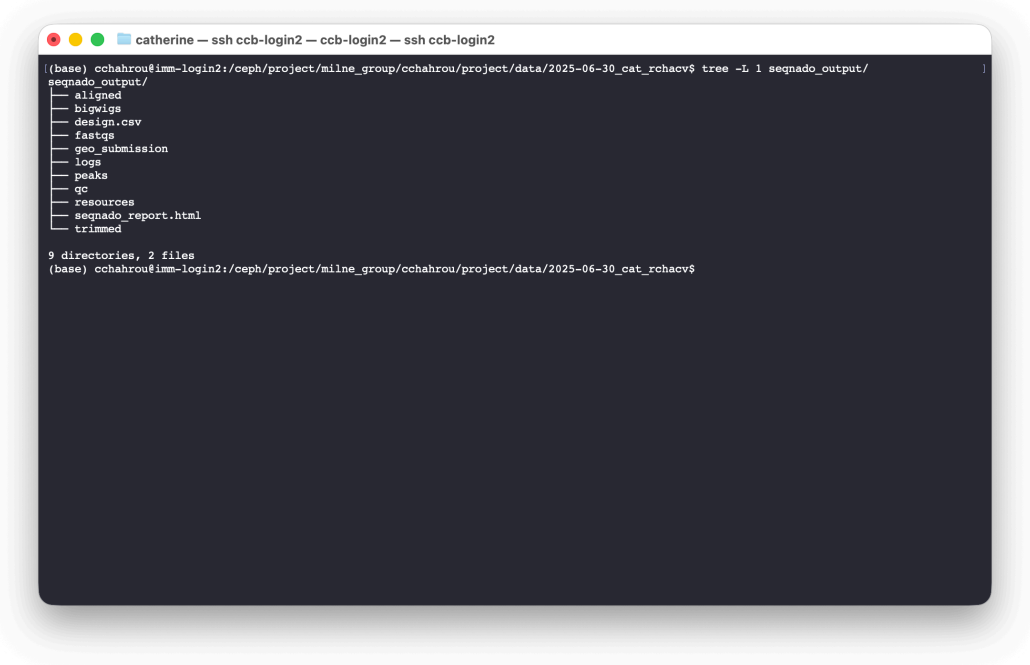


Figure 3.5: Example output file directory from SeqNado showing the uniform structure and key outputs for CUT&Tag data.

Every run also logs the SeqNado version, user ID, and run date in the output directory. Comprehensive QC reports are automatically generated using MultiQC, along with a custom seqnado-report.html summarising all key metrics, from raw FASTQ quality and contamination checks (FastQ Screen) through trimming, mapping, and library complexity (Figure 3.6).

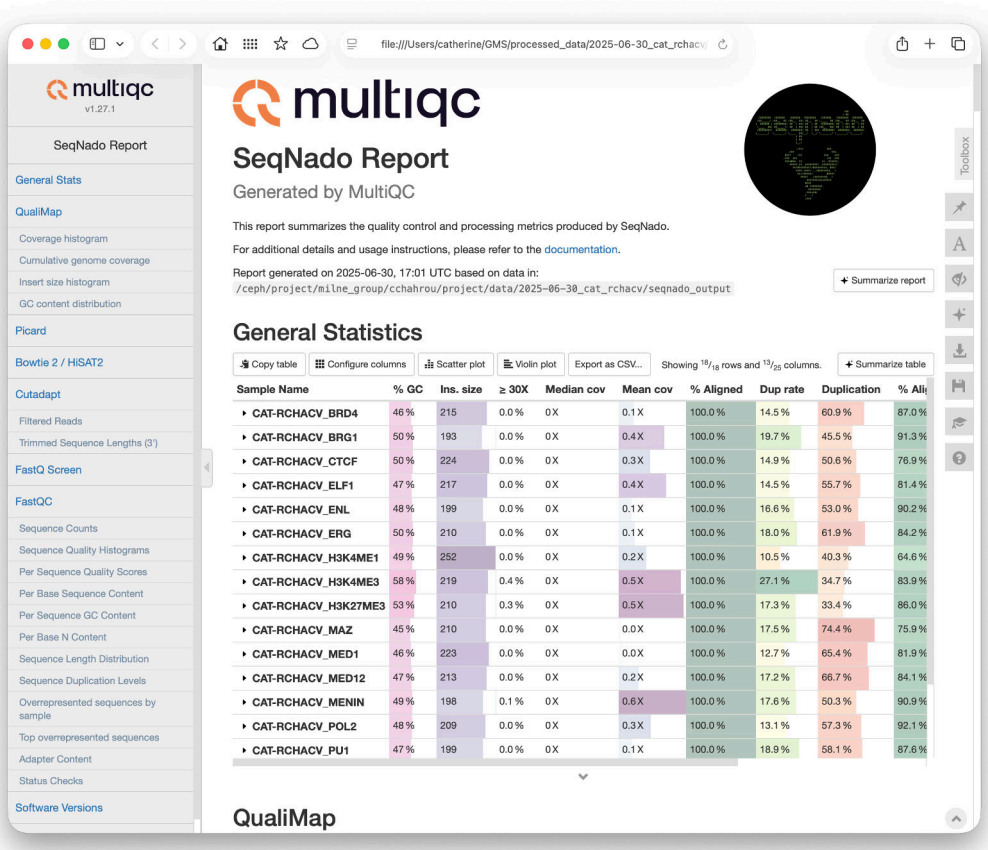


Figure 3.6: Example MultiQC report from SeqNado showing key metrics from FastQC, FastQ Screen, trimming, mapping statistics, library complexity, and QualiMap BAM QC for CUT&Tag data.

3.3.3 Summary

By standardising inputs, automating intermediate steps, and producing clean, uniform outputs, SeqNado reduces setup time and technical overhead. The entire process from installation to final outputs can be completed with four simple commands, providing a reproducible, efficient pipeline for diverse assay types.

3.3.4 Supported Assays

RNA-Seq

After QC and read trimming of the FASTQ files, the reads are mapped using STAR, they are then post-processed to sort and index, and remove blacklisted regions, and by default reads are filtered to only mapped, properly paired reads. The reads are then quantified using featureCounts with the gtf provided in the config. If a

deseq column is added to the design with “control” and any other group added in the design a pairwise differential gene expression analysis can be performed. Stranded bigWig files can also be produced; these can be added to a UCSC hub and can also be used to produce track plots if plotting coordinate files are provided. The raw data and processed counts can also be prepared for GEO submission with md5 checksums calculated ready for addition to the GEO sample metadata file.

ATAC-Seq

For ATAC-seq after QC and trimming of the FASTQ files, mapping is carried out with bowtie2, the reads are then post-processed as per RNA seq with two additional steps. Reads are de-duplicated and shifted for Tn5 cutting with +4bp for the positive strand and -5bp for the negative strand as per Buenrostro 2013. Reads are filtered to retain only mapped, properly paired reads. From these bam files, bigWigs, and consensus counts are performed as requested by the user in configuration. Peaks can either be called using bigWigs generated using the peak caller LanceOtron (Hentges, L. D. et al. 2022), or from the bam files using Homer or Macs2 as preferred. Outputs can also be prepared for UCSC hub or geo submission as above.

ChIP-seq and CUT&Tag

ChIP-seq data is processed as per ATAC-seq however without the alignment shifting. ChIP-seq and CUT&Tag data can also be exogenously reference normalised if configured to use a concatenated reference genome, this will produce reference normalised bigWig files and scaling factors. Peaks can be called using input or without if not provided.

CUT&Tag is processed as per ChIP-seq however the default peak caller is set to SEACR from the Henikoff lab (Kaya-Okur, H. S. et al. 2019). This can also be reference normalised to e-coli if the concatenated genome is provided.

Whole Genome sequencing

Short-read whole genome sequencing can also be processed. After the uniform QC, trimming and FastQ Screen, the reads are mapped using bowtie2 and bcftools is used to call variants, these are split to multi-allelic sites and can be filtered for depth or other flags as per the config. Resulting SNPs are also annotated using dbSNP or similar if an annotation vcf file is provided.

Methylation

After QC, FastQ Screen and trimming reads are mapped using bowtie2 and methylation calling carried out using methyldackel, the calls can be inverted in the case of TAPS and output can be produced as bedgraph or methylkit as specified by the user. Methylation spike-in controls are split from the bams and quantified separately.

CRISPR screen

For the convenience of users, CRISPR screen data processing has also been included. Reads are quality checked, and screened for potential contamination. Trimming of guide adaptor sequences is carried out as per the configuration file. Reads are then mapped using bowtie2 and quantified against the guides using featureCounts and a Simplified Annotation Format (SAF) file for the guides.

MCC

For processing data from MCC (Hamley, J. C. et al. 2023), reads are QC'd and trimmed. The FASTQ files are then deduplicated and flash is used to locate junctions. They are then mapped with minimap to viewpoints from a fasta containing the viewpoint sequences. CIGAR strings are used to identify soft-clipping bases which are sections of the reads not matching the viewpoints and the FASTQ headers are updated to include the viewpoint identifier, soft clip location and read orientation to viewpoint. The reads are then mapped to the interaction site in the full genome and the bam files are split on read tags to the viewpoints. The read group identifiers are then used to produce bigWig files per viewpoint. And the ligation junction position relative to the viewpoints are quantified.

3.3.5 Outputs

UCSC hub

All bigWig files produced as well as peak calling bed files can be used to programmatically produce UCSC hubs which are ready to be visualised as a UCSC session.

GEO submission

To expedite the publication of NGS data, raw and processed files are prepared for GEO submission with md5 checksums calculated and a metadata template created

to be uploaded to the gene expression omnibus (GEO) server (Supplementary Table C.1).

Plotting

Heatmap and metaplots over the GTF regions can be produced for all data where a bigWig file is produced. Plotting of tracks from bigWigs over specific genes is made possible using PlotNado.

Quality control metrics

A full multiQC report is generated for all metrics including FastQC, trimming, FastQ Screen, library complexity, mapping statistics, QualiMap BAM QC for ChIP-seq, ATAC-seq and CUT&Tag to assess coverage, fragment size distribution, and GC content. For RNA-seq Genomic origin and Gene coverage is included.

Machine Learning Integration

ML methods are being increasingly applied to genomic data. However raw sequencing data requires extensive and careful pre-processing to be made useful for use in ML tasks. Different assay types require different strategies using different quality control, mapping and downstream processing, consideration normalisation. This preprocessing is also required to be reproducible between datasets and for use by others. For ATAC-seq, ChIP-seq and CUT&Tag in addition to consensus or per Immunoprecipitation (IP) quantification, SeqNado also can output a h5ad file which quantifies mean signal from bigWig files over customisable regions, either user-defined regions of interest or uniformly-spaced genomic bins. This was implemented using `crested.import_bigwigs` (Kempynck, N. et al. 2025) This fully customisable, ML-ready output accepts bigWig files produced with the user's choice of normalisation method and stores unscaled signal to preserve information for downstream ML workflows (Figure 3.7).

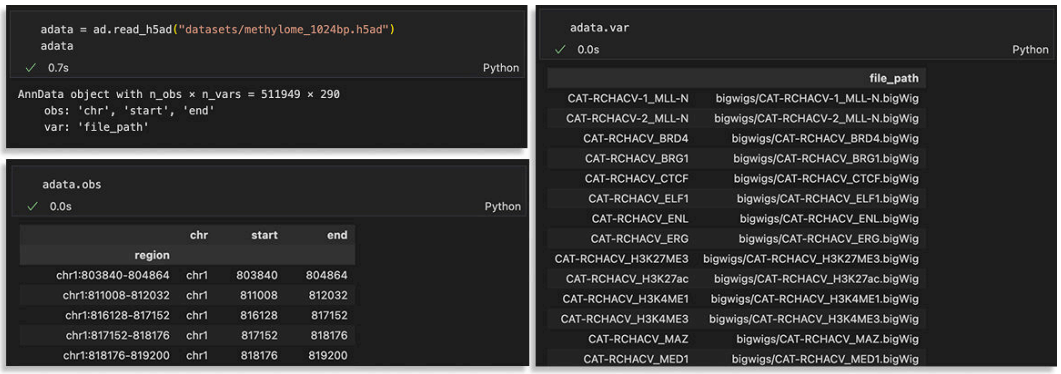


Figure 3.7: Example SeqNado ML-ready h5ad file output.

3.4 Pipelines Benchmarked

A wide range of pipelines available for processing high-throughput sequencing data are mostly assay-specific. Few are capable of handling expression, epigenomic, and genomic data within a unified framework, and none of the pipelines evaluated here natively produce ML ready outputs. Here, I selected four widely used, open-source, HPC-compatible pipelines that support a range of epigenomic and expression assays. The pipelines benchmarked against SeqNado were nf-core (v2.1.2), ENCODE-DCC (v 2.2.3), snakePipes (v3.2.0) (Table 3.1). The aim was to benchmark their usability, resource efficiency, and suitability for streamlined preprocessing across multiple assay types. To benchmark performance and usability five replicates of paired-end ATAC-seq FASTQ files (GSE117862), totalling approximately 18 GB was used. The aim was to map reads to the human genome (hg38) and produce both BigWig coverage tracks and MACS2 peak calls using minimal resources whilst monitoring computational performance, resources required and the quality of the processed output data. Reference indexing was excluded from runtime measurements. Transcription Start Site (TSS) sites were extracted from the UCSC GTF and used to calculate TSS enrichment scores for quality control.

3.4.1 nf-core

Nf-core is a community-driven framework for the development, testing, and distribution of peer-reviewed, best-practice bioinformatics pipelines built using the Nextflow workflow language (Ewels, P. A. et al. 2020). The project is widely adopted and supports standardised analysis workflows with a focus on portability and reproducibility. Each nf-core pipeline is independently developed for a specific assay and adheres to strict development guidelines. These include high-quality documentation, ver-

sion control, containerisation (Docker, Singularity/Apptainer, or Conda), continuous integration testing, and the requirement to provide test datasets and usage instructions. By building on Nextflow, nf-core pipelines benefit from native support for resource managers and container platforms, enabling platform-agnostic deployment. The pipelines address the problems of reproducibility in bioinformatics through the self-defined best-practice pipelines, and all internal tools are containerised, controlling version and dependency issues. Despite its strengths, nf-core pipelines are rigidly structured, and internal tools are often not easily configurable. Each pipeline is written for a single assay and must be configured separately. Pipelines may include tasks unnecessary for specific users' goals, with no straightforward way to disable them. While Nextflow supports symbolic linking, many nf-core pipelines rely on internal tools that require local file copies, resulting in large working directories and substantial storage overhead—often reaching terabytes in multi-sample analyses.

3.4.2 ENCODE-DCC

The ENCODE-DCC (Data Coordination Center) pipelines are the official workflows used to process data for the ENCODE consortium. They are written in WDL and executed using Caper (Cromwell Assisted Pipeline Executor), a Python-based wrapper for the Cromwell workflow engine. These pipelines follow strict reproducibility and quality control standards and are validated extensively on large public datasets. Support is provided for a wide range of assays, including ChIP-seq, ATAC-seq, RNA-seq, and Hi-C. However, configuration is complex, and pipeline parameters are often hardcoded, including limits such as a maximum of 10 replicates per assay. The workflows must be cloned from GitHub and executed locally. Although Cromwell supports multiple backends (e.g., SLURM, Amazon Web Services (AWS), Google Cloud), the ENCODE-DCC pipelines are not readily portable without backend-specific reconfiguration.

3.4.3 snakePipes

snakePipes is a Snakemake-based framework for processing epigenomics datasets, with support for assays such as ChIP-seq, ATAC-seq, RNA-seq, and WGBS. It uses YAML-driven configuration and includes built-in quality control metrics reporting(Bhardwaj, V. et al. 2019). However, snakePipes is less streamlined for non-expert users, requiring manual editing of multiple configuration files. All intermediate files are retained by default, inflating storage requirements. Additionally,

the pipeline is not fully end-to-end, users must perform separate configuration and execution to progress from read mapping to downstream analyses such as peak calling.

Category	Feature	SeqNado	nf-core	ENCODE-DCC	snakePipes
Setup	Workflow language	Snakemake	Nextflow	WDL	Snakemake
	Installation	Conda, Pip	Nextflow	GitHub	Conda
	Dependency management	Apptainer, Conda	Docker, Apptainer, Conda	Apptainer	Installation script
	Initialisation	Automatic	Manual YAML	Manual JSON	Manual YAML
	Pipeline scope	Unified multi-assay workflow	Separate pipelines per assay	Separate pipelines per assay	Multi-part assay dependant
	Assay configuration	Automated YAML	Manual YAML	Manual JSON	Manual YAML
Assays	RNA-seq	Yes	Yes	Yes	Yes
	ATAC-seq	Yes	Yes	Yes	Yes
	ChIP-seq	Yes	Yes	Yes	Yes
	CUT&Tag	Yes	Yes	Yes	–
	WGBS	Yes	Yes	Yes	Yes
	TAPS	Yes	–	–	–
	WGS	Yes	Yes	Yes	–
	CRISPR	Yes	Yes	–	–
	MCC	Yes	–	–	–
	Hi-C	–	Yes	Yes	Yes
Output	ML dataset	Yes	–	–	–
	GEO submission	Yes	–	Yes	–
	UCSC hub	Yes	IGV	–	–

Table 3.1: Comparison of pipeline setup, assay support, and key features across four bioinformatics workflows. Summary of workflow languages, configuration requirements, supported assays, and output capabilities for SeqNado, nf-core, ENCODE-DCC, and snakePipes pipelines.

3.5 Benchmarking Results

3.5.1 Computational Performance

The evaluated workflows demonstrated varying trade-offs between runtime, CPU usage, and disk input/output (Figure 3.8). nf-core achieved the fastest runtime, completing in 1 hour 39 minutes, but at the cost of very high resource usage. It consumed 40.49 CPU hours, generated 1799 GB of total I/O, and left behind a 771.85 GB output directory, 654 GB of which was a working directory that can be manually cleaned post-run. While fast, the memory and disk footprint of nf-core's Nextflow backend is considerable and may be prohibitive in constrained environments. ENCODE took 7 hours and 23 minutes hours to run using 104.5 CPU hours and resulted in 88.7 GB output, the I/O was not available from WDL. snakePipes completed in 2 hours 17 minutes using 59.76 CPU hours, with 110.89 GB total I/O and a final output of 93.69 GB. Intermediate files were not automatically purged by default. The higher CPU time suggests less efficient parallelism or more compute-intensive steps. SeqNado ran in 2 hours 16 minutes, closely matching snakePipes in wall-clock time, but required only 32.01 CPU hours, nearly half that of snakePipes. It also had the lowest disk impact, generating 139.54 GB total I/O and producing a compact 30.11 GB output directory. Intermediate files are automatically cleaned during execution, significantly reducing storage overhead.

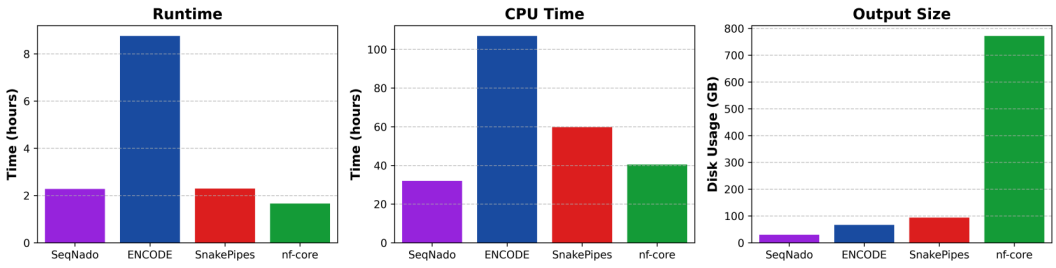


Figure 3.8: Workflow benchmarking summary (CPU time, runtime, output data size)

3.5.2 Output Quality

Given the same input data, output data quality varied substantially. Read counts varied across pipelines despite all using Bowtie2 for alignment. ENCODE and SnakePipes applied internal read filtering steps that could not be disabled, resulting in lower retained reads per sample. ENCODE retained a mean of 46 million reads (SD ± 3.6M), and SnakePipes 51.2 million (±3.5M), while nf-core and SeqNado re-

tained higher and more consistent read counts 54.4 million (±4.4M) and 54.0 million (±3.9M), respectively (Figure 3.9A).

Fraction of Reads in Peaks (FRiP) is used as a metric for signal enrichment (Landt, S. G. et al. 2012). However, in ATAC-seq, excessively high FRiP scores can be misleading—particularly when they result from overly permissive peak calling that captures background noise alongside true signal. Although MACS2 was used by all four pipelines, ENCODE called over twice as many peaks per sample compared to the others. This inflated its mean FRiP to 0.58 (±0.05), in contrast to 0.26 for nf-core, 0.21 for snakePipes, and 0.25 for SeqNado (each with ±0.02 SD) (Figure 3.9C). Specifically, ENCODE reported an average of 226,783 peaks per sample (±37,889), compared to 61,565 for nf-core, 55,518 for snakePipes, and 59,416 for SeqNado. This suggests that ENCODE's high FRiP values may reflect broad, nonspecific peak calling, artificially inflating signal enrichment. By contrast, SeqNado and nf-core produce more conservative peak counts and correspondingly lower FRiP scores, indicating greater specificity and likely more biologically plausible peak calls (Figure 3.9B).

TSS enrichment, a key measure of signal quality in ATAC-seq, reflects how well reads accumulate around transcription start sites (TSS). Although ENCODE produced the highest average enrichment (mean = 6.69, SD ± 0.11), SeqNado closely followed (mean = 6.53, SD ± 0.34) and showed slightly greater variability across replicates. Both outperformed nf-core (mean = 6.20, SD ± 0.33) and snakePipes (mean = 5.12, SD ± 0.35), which showed consistently lower enrichment. This suggests that ENCODE and SeqNado achieve better signal-to-noise around TSSs , whereas snakePipes may allow more background noise or insufficiently centre fragments during alignment or filtering (Figure 3.9D).

TSS enrichment heatmaps revealed further quality differences (Figures 3.9F, 3.9G, 3.9H, 3.9I). ENCODE displayed a sharply defined accumulation of reads at TSSs, which likely reflects its use of a 150 bp smoothing window during MACS2 signal track generation. This smoothing step centres and sharpens the signal around accessible sites, directly enhancing the apparent TSS enrichment. While this produces visually clear footprints over TSSs , it may also artificially narrow or distort broader regulatory features such as enhancers, reducing interpretability outside promoter contexts. nf-core heatmaps showed black bands across several TSS regions, likely caused by missing (NaN) values introduced during bigWig scaling or normalisation steps incompatible with deepTools. SeqNado achieved strong, consistent enrichment without such artifacts, while SnakePipes demonstrated a comparable enrichment pattern to SeqNado, albeit with a lower dynamic range and reduced

signal strength.

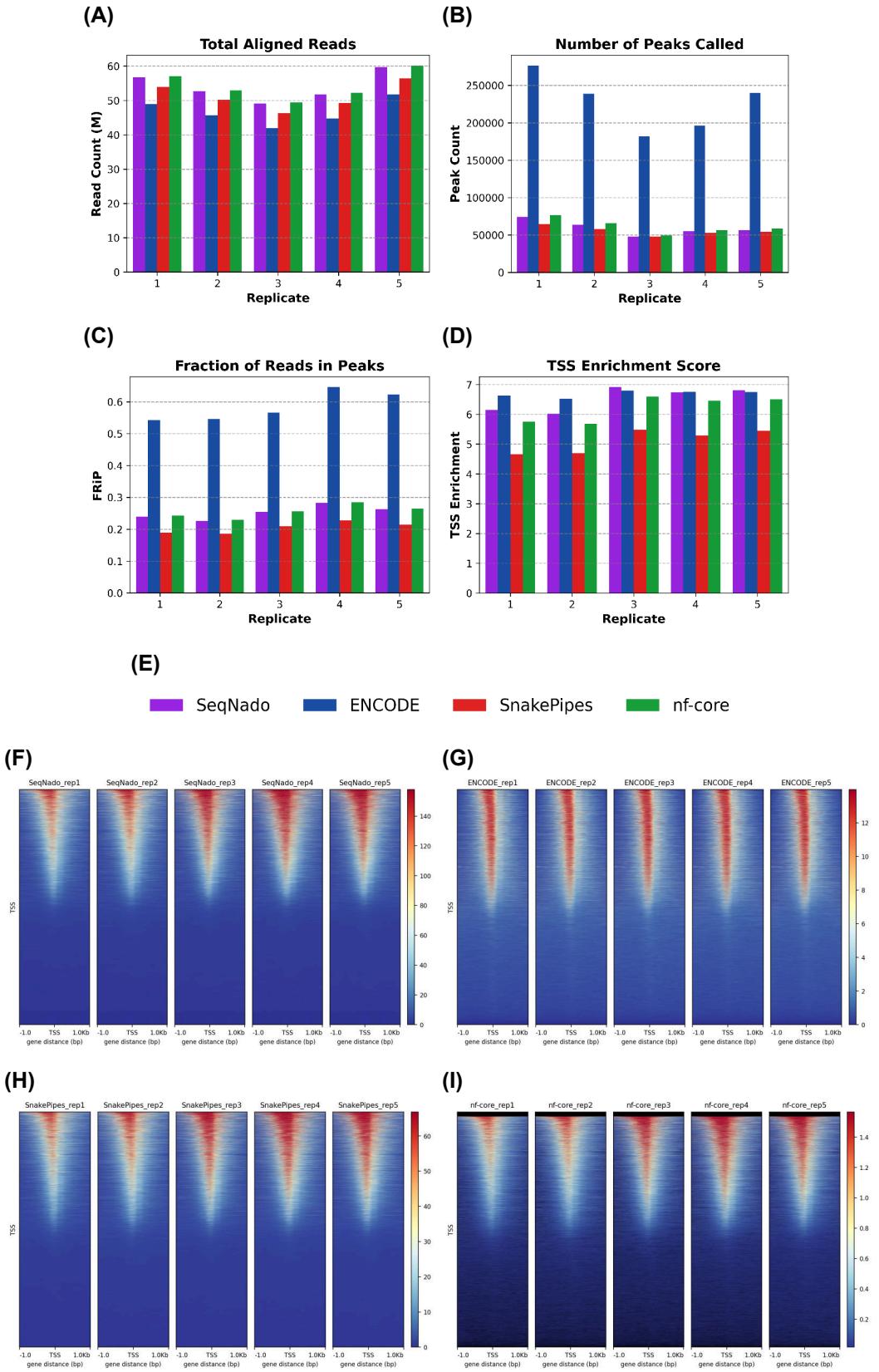


Figure 3.9: Continued next page.

Figure 3.9: (Previous page) Output data quality metrics from pipeline benchmarking of publicly available SEM cell ATAC-seq data (GSE117862) across four pipelines (SeqNado, ENCODE, snakePipes, and nf-core). **(A)** Total reads aligned, **(B)** Number of peaks called, **(C)** Fraction of reads in peaks (FRiP), **(D)** TSS enrichment scores, **(E)** Legend for **(A-D)**, **(F-I)** Heatmaps for TSS enrichment for SeqNado, ENCODE, snakePipes, and nf-core.

3.6 Case Study: High-Throughput CUT&Tag Processing

To prepare my data for all downstream ML tasks, I ran 134 samples of CUT&Tag data through SeqNado. Here, I opted to carry out all QC including FastQ Screen, normalise to the *Escherichia coli* reads from the pAG-Tn5 transposase, call peaks with LanceOtron, generate bigWigs scaled by both RPKM and reference-normalised RPKM, create a UCSC hub, and prepare the data for GEO upload. RPKM bigWigs were used to quantify signal at all promoter regions for downstream ML tasks, with reference-normalised bigWigs generated for biological analysis and visualisation. The total runtime for this was 11 hours and 9 minutes using 835 CPU hours and resumed after downtime at 2:00 a.m. as shown in Figure 3.10. Using SeqNado delivered a fully reproducible, analysis-ready dataset that underpins the ML results presented in the following chapters.

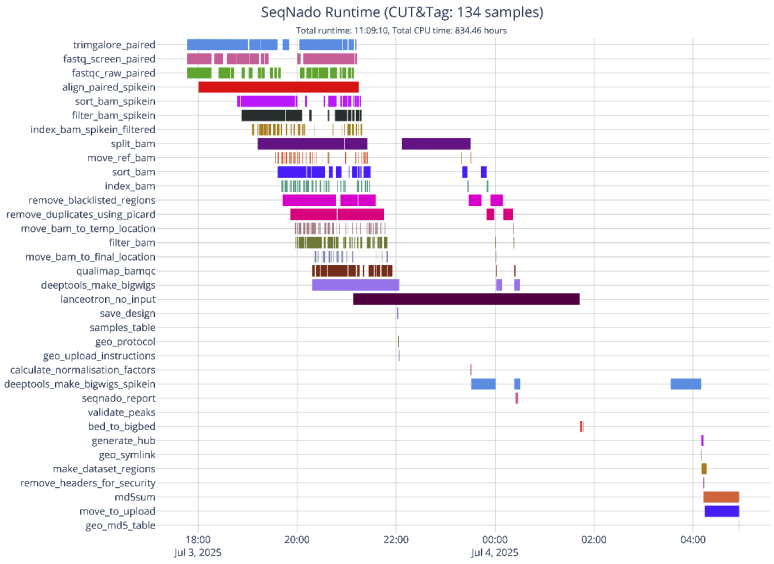


Figure 3.10: SeqNado runtime visualisation. Gantt chart visualising the execution timeline of SeqNado tasks during the processing of 134 CUT&Tag samples, highlighting parallelisation and resource allocation.

3.7 Discussion

For the ATAC-seq data benchmarked, SeqNado was the only pipeline to run end-to-end without user intervention and debugging. Nextflow was fast to run but at a considerable resource burden due to intermediate files and resulted in bigWig files which were not compatible with downstream plotting. ENCODE took the longest time while also using the most compute and resulted in highly filtered reads which skewed the downstream plotting and TSS enrichment metrics. SnakePipes ran in a reasonable time using more compute and resulted in data which was similar to SeqNado's output but with a lower overall TSS enrichment given identical data. SeqNado also has the ability to output various plots and GEO submission ready data as well as well organised and useable processed files which are ML-ready.

In addition to the ATAC-seq data using the published SEM ATAC-seq data above, used for benchmarking, I attempted to benchmark the pipelines using ChIP-seq and RNA-seq data from ENCODE for GM12878 cells. The data was downloaded and SeqNado was run on these data without issue, however in the other pipeline tools several issues were found that meant that the data could not be benchmarked using the pipelines without considerable de-bugging work done. For example, in snakePipes, the RNAseq conda environment is corrupt and for the ChIP-seq an error in logging for trim galore caused failure in the pipeline execution. For NF-core, an out of memory due to deeply sequenced RNA seq caused failure in the DupRadar task due to task level memory requirements being undefined in pipeline. In Encode, where the data was sourced, the ChIP-seq tagAlign task errors with empty array when the output file is created and is not empty. A patch fix to increase the file latency in the source code was required for the pipeline to continue. For the RNAseq an out of memory error in Kallisto quantification task for deeper RNAseq data, needing additional configuration via editing of the JSON file.

By providing a unified, containerised, and parallelised workflow, SeqNado significantly simplifies the process of transforming raw NGS reads into high-quality, uniformly processed data. This uniformity is particularly advantageous for ML applications, where standardised inputs are paramount for robust model training and performance evaluation. The pipeline's modularity, reproducibility, and ease of use position it as a powerful resource for large-scale genomic studies that aim to integrate ML strategies. Future work will include the ability to release the ML datasets beyond GEO with the pre-processed datasets hosted by hugging face for use by others for ML tasks.

SeqNado has already been used to prepare data for the following publications with more in preparation:

- Ana M. Dopico-Fernandez *et al.* (2026) 'FACT safeguards promoter topology by maintaining nucleosomes and restricting chromatin factor spreading'. *bioRxiv*, https://doi.org/10.64898/2026.02.18.706382
- Vassilena Sharlandjieva *et al.* (2026) 'Menin maintains enhancer-promoter interactions in a leukemia-specific manner'. *bioRxiv*, https://doi.org/10.64898/2026.01.16.698179
- Elisa K. Barrow Molina *et al.* (2025) 'TRANCERs: Engineering enhancers into autonomous tissue-specific expression cassettes'. *bioRxiv*, https://doi.org/10.1101/2025.10.27.684763
- Grace A. Meaker *et al.* (2025) 'Genome-wide screen identifies *Runx2* as a novel regulator of haematopoietic stem cell expansion'. *Blood*, https://doi.org/10.1182/blood.2025029115
- Valerio Ciaurro *et al.* (2025) 'Menin inhibitor DS-1594b drives differentiation and induces synergistic lethality in combination with venetoclax in acute myeloid leukemia cells with rearranged Mixed-lineage Leukemia and mutated Nucleophosmin-1'. *Haematologica*, https://doi.org/10.3324/haematol.2024.286833
- Alastair Smith *et al.* (2025) 'Enhancer heterogeneity in acute lymphoblastic leukemia drives differential gene expression in patients'. *Blood*, https://doi.org/10.1182/blood.2024028019
- Megan R. Teh *et al.* (2025) 'Iron deficiency causes aspartate-sensitive dysfunction in $CD8^+$ T cells'. *Nature Communications*, https://doi.org/10.1038/s41467-025-60204-7
- Rebecca Ling *et al.* (2024) 'The fetal specific gene *LIN28B* is essential for human fetal B-lymphopoiesis and initiation of KMT2A::AFF1 infant leukemia'. *bioRxiv*, https://doi.org/10.1101/2024.09.18.613730

4 DNA Sequence specificity in MLL binding

4.1 Introduction

This chapter investigates whether MLL binding can be predicted directly from DNA sequence, and how sequence features associated with binding differ between MLL and MLLr contexts. MLL fusion proteins (MLL-FP) arise from chromosomal translocation of the *KMT2A* gene, creating fusions which retain the N-terminal CXXC CpG-recognition and Menin/LEDGF interaction modules but lose the C-terminal SET methyltransferase domain (Figure 4.1).

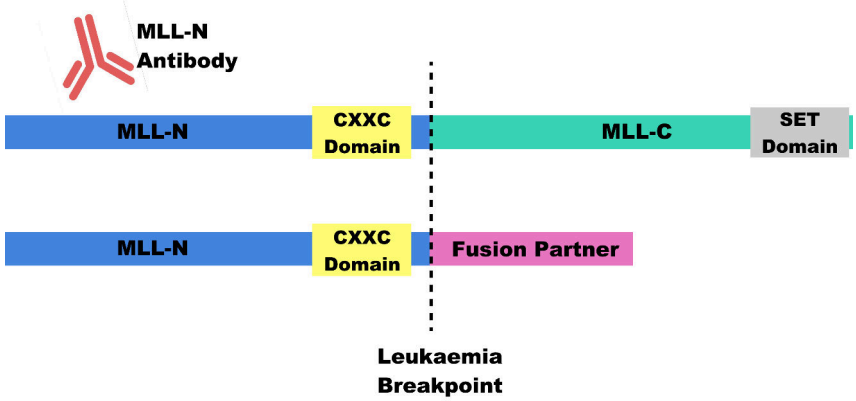


Figure 4.1: Wild-type and fusion MLL at the leukaemia breakpoint with antibody epitope. Translocation replaces MLL-C with a fusion partner while retaining MLL-N (including the CXXC domain); N-terminal antibodies therefore recognise both wild-type and fusion proteins. Adapted from (Kerry, J. et al. 2017).

Common fusion partners include AF4 (AFF1) and AF9 (MLLT3) (Meyer, C. et al. 2023). MLL binds unmethylated CpG DNA including at CpG rich promoters via its CXXC zinc-finger domain (Birke, M. 2002; Allen, M. D. et al. 2006). However, MLL only binds a subset of actively expressed genes and it remains unclear why (Milne, T. A. et al. 2005a). This is further complicated by observations that MLL-FP bind a broader and more dispersed binding patterns that extend beyond canonical CpG-

rich targets, including spreading into gene bodies (Kerry, J. et al. 2017). These contrasting profiles along with the specificity of MLL bound active genes raise the question of whether sequence determinants alone can explain the difference. Here I use ML approaches to recover predictive sequence features for both MLL and MLL-FP to explore if this can lead to insight into MLL binding logic from DNA sequence.

4.1.1 A motif for MLL

The concept of a position-weight matrix to represent base frequencies within a conserved sequence was first introduced by Stormo and colleagues, who applied a perceptron algorithm to distinguish translation initiation sites in *E. coli* (Stormo, G. D. et al. 1982). This framework was later adapted for modelling DNA binding protein motifs, and visualised through sequence logos (Schneider, T. D. and Stephens, R. 1990). In 2024, the motif database HOCOMOCO v12 update released an MLL motif (KMT2A.H12CORE.0.P.B) derived from 36 human ChIP-seq experiments (223 peak sets) (Vorontsov, I. E. et al. 2024). The motif is 23 bp in length, derived from ChIP-seq peak sets, is highly repetitive, and ∼87% GC content, consistent with CpG-island binding (Figure 4.2).

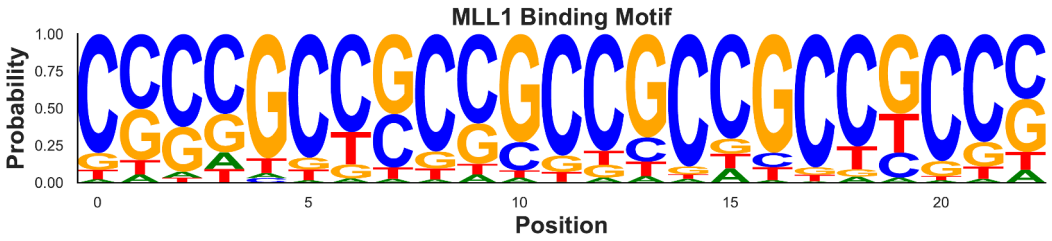


Figure 4.2: HOCOMOCO v12 MLL motif logo (KMT2A.H12CORE.0.P.B) derived from 36 human ChIP-seq experiments (223 peak sets) (Vorontsov, I. E. et al. 2024).

On inspection, the aggregated peak sets used to generate the motif span not only contains cells without MLL fusions such as RCH-ACV which is a B-cell leukaemic cell line with E2A-PBX1 fusion (Jack, I. et al. 1986), but also MLLr leukaemic cell lines including SEM, RS4;11, MV4-11, and THP-1. SEM, RS4;11 and MV4-11 all harbour MLL-AF4 fusions, while THP-1 has MLL fused to AF9 (MLL-AF9). However the cell lines were all derived from different leukaemic contexts, SEM from a childhood precursor B-ALL, RS4;11 from an adult B-ALL, MV4-11 from biphenotypic B-myelomonocytic leukaemia (AML), and THP-1 from AML (Greil, J. et al. 1994; Lange, B. et al. 1987; Stong, R. C. et al. 1985; Tsuchiya, S. et al. 1980).

In SEM cells, MLL ChIP-seq shows pronounced spreading into gene bodies a hall-

mark of MLL-AF4 fusion driven occupancy and the MLL ChIP-seq peak set from this study was included in the HOCOMOCO build (Kerry, J. et al. 2017). The collection further incorporates mouse experiments, including peak sets reported from MLL-knockout backgrounds, and cells treated with Menin inhibitor (VTP-50469). Menin inhibition is known to disrupt MLL/menin chromatin engagement and reduce MLL occupancy at target genes (Krivtsov, A. V. et al. 2019).

Given that fusion status, Menin inhibition, and knockout contexts all perturb MLL recruitment, the HOCOMOCO motif is best interpreted as a CpG-rich consensus baseline rather than a faithful readout of a specific MLL binding motif. It captures the CpG preference but does not resolve fusion-specific or treatment-dependent sequence logic.

4.1.2 DNA Sequence and Machine Learning

For two decades, transcription-factor binding was modelled with motifs represented as position weight matrices (PWMs), typically discovered from peak sets with tools like MEME Suite (Bailey, T. L. et al. 2009). PWMs capture short, local patterns but assume independence between positions and cannot model long-range context (Stormo, G. D. 2000; Bailey, T. L. and Elkan, C. 1994).

A major shift came with Convolutional Neural Network (CNN) models, which learn motif-like filters directly from raw sequence often one-hot encoded where the sequence is represented as a four-letter matrix. Early models showed that sequence alone can predict many chromatin readouts and estimate variant effects most prominently DeepSEA, Basset , and (Zhou, J. and Troyanskaya, O. G. 2015; Kelley, D. R. et al. 2016; Quang, D. and Xie, X. 2016). Later, BPNet fit base-pair resolution profiles and recovered motif grammar including spacing and orientation (Avsec, Ž. et al. 2021). Together, these studies established that DNA sequence models are predictive of chromatin features such as accessibility, transcription factor binding, and histone modifications.

With the advances to transformer models using self-attention (Vaswani, A. et al. 2017), genomics adopted natural language processing inspired approaches with one clear barrier, where in natural languages such as English, the letters are arranged into words and the broader context of sentences. DNA is clearly not represented like this, to overcome this, vocabularies are generated from DNA by tokenisation in two different ways, either by representation of set k-mer combinations of all bases which due to the frequency imbalance of tokens can result in rare word problem where some k-mers are never seen during training. To address this, DNA can

be tokenised through Byte-Pair Encoding (BPE) where a DNA vocabulary is built by repeatedly merging the most frequent adjacent bases into tokens, so common k-mers are represented as variable length tokens.

Using this vocabulary, Bidirectional Encoder Representations from Transformers (BERT) style masked language pretraining teaches the model to predict hidden tokens from context on either side (Devlin, J. et al. 2018). This pre-training allows them to learn the underlying DNA grammar, not just which short patterns (motifs) occur, but how they are arranged, their preferred spacing, order, orientation, and co-occurrence, along with broader composition cues like CpG content, repeats, and sequence patterns that correlate with nucleosome positioning. The resulting representation, learned from vast unlabelled sequence, can then be efficiently transferred by fine-tuning on specific downstream tasks.

This led to the availability for pretrained DNA Language Models (DNA-LM) including Bidirectional Encoder Representations from Transformers model for DNA-language (DNABERT) which introduced masked-language modelling over k-mer tokens (Ji, Y. et al. 2021), followed by DNABERT-2–117M (DNABERT-2) which moved to BPE encoding with learned variable-length tokens (Zhou, Z. et al. 2023). However, DNABERT models were pretrained on 135 genomes spanning mammalian, fungal, bacterial, and protozoan species. With the understanding that DNA sequence grammar could be species specific, Genome Rules Obtained Via Extracted Representations (GROVER) pretrained a BERT-like model on the human reference (Sanabria, M. et al. 2024), and the Nucleotide Transformer models scaled size and training data across species with modern positional encodings (Dalla-Torre, H. et al. 2024). These pretrained DNA-LM transfer well to new tasks and often need less labelled data than models trained from scratch.

4.1.3 Chapter Aims

MLL has sequence specificity in binding through its CXXC domain which recognises CpG rich DNA, but that it is not well understood why it only binds a subset of suitable CG rich promoters and that in a fusion context it has been observed to spread into gene bodies (Birke, M. 2002; Milne, T. A. et al. 2005a; Kerry, J. et al. 2017). These observations motivate using machine-learning on DNA sequence to infer the underlying binding grammar which motifs, spacing, and combinations distinguish MLL from MLL-fusion binding and to test how far sequence alone explains their differences. Using model interpretation methods, I then ask whether these models can recover biologically meaningful sequence rules for MLL recruitment

This chapter investigates whether MLL binding can be predicted directly from DNA sequence, and how sequence features associated with binding differ between cell contexts. To address this, the aims are:

1. Compare transformer-based DNA language models and a CNN baseline to evaluate their relative ability to generalise.
2. Use the fine-tuned multilabel classification model to determine binding preferences.
3. Using feature extraction, to explore whether predictive sequence features are shared across cell lines or are MLL fusion specific.

4.2 Methods

4.2.1 Data preparation

Regions and signal extraction

Promoter regions were defined as 1,024 bp windows centred on the TSS of RefSeq transcripts, a standard well-curated annotation with stable TSS coordinates. A hg38 RefSeq annotation file was parsed with rtracklayer, retaining protein-coding transcripts on canonical chromosomes; both NM (experimentally validated) and XM (computationally predicted) isoforms were included to maximise promoter coverage. Strand-aware coordinates were computed as $\pm$ 512 bp around the TSS and clipped to chromosome bounds using hg38.chrom.sizes. Duplicates were removed, and the resulting regions exported to BED.

MLL does not only bind to promoters but more distally (Smith, A. et al. 2025a). To capture this broader binding landscape while keeping the dataset manageable, I used regions defined by the Twist Human Methylome Panel (Twist Bioscience 2022), a commercial oligonucleotide capture panel covering approximately 123 Mb of hg38. The panel targets CpG-dense regulatory elements including CpG islands, shores, and shelves, promoter-flanking regions, open chromatin, transcription factor and CTCF binding sites, and enhancers, making it well suited to studying MLL, whose binding is enriched at CpG-rich sequence via its CXXC domain. Restricting to panel regions reduces the search space from tens of millions of genome-wide tiles to a few hundred thousand informative bins, while avoiding the accessibility bias that would arise from region selection based on chromatin accessibility peaks alone. These methylome regions, referred to throughout this thesis as methylome regions, were defined by tiling hg38 into fixed 1,024 bp bins and retaining only those overlapping Twist Methylome capture targets; scaffold and patch contigs and chrM were excluded. Non-overlapping bins were generated with PyRanges, intersected with Twist targets, sorted, labelled, and exported to BED. This resulted in a dataset enriched for CpG-rich sequence with a variable level of MLL signal, providing a balanced set of bound and unbound regions.

CUT&Tag data for MLL N-terminus (MLL-N) from six cell lines (RCH-ACV, SEM, RS4;11, MV4-11, THP-1, and OCI-AML3; Table A.1) were processed into bigWig format using the SeqNado workflow (see Chapter 3). bigWigs were quantified over promoter and methylome intervals with the Cis Regulatory Element Sequence Training, Explanation and Design (CREsted) package (import_bigwigs, target = ”mean”),

producing region × sample matrices of average Reads Per Kilobase Per Million Mapped Reads (RPKM) values. Matrices were exported to Parquet for downstream analysis for fast data extraction.

Normalisation, binarisation, and data partitioning

Per-sample RPKM signals were upper-quartile normalised to account for differences in library size across samples, and log-transformed to reduce the influence of extreme values before binarisation. Regions above the 80th percentile were labelled bound and those below the 20th percentile labelled unbound; excluding the intermediate 60% removes ambiguous loci where signal may reflect low-occupancy binding, technical noise, or sample heterogeneity, producing a cleaner binary classification task.

To avoid sequence leakage, datasets were split by chromosome: chr9 was held out for testing, chr8 for validation, and all remaining autosomes for training. Each record included the genomic interval, split label, per-sample binary labels, and the DNA sequence retrieved from hg38 using crested.fetch_sequences. Label to sample ID mappings were stored as JSON.

Sequence composition and motif enrichment

GC content was calculated per interval for both promoter and methylome datasets. Motif enrichment was performed with AME (Ma, W. et al. 2014)(v5.5.5), using each dataset as input with the other as background. A filtered version of HOCOMOCO v12, restricted to factors expressed in at least one cell line, was used as the reference database. Fisher's exact test was applied, retaining motifs with $q < 0.05$, and enrichments were summarised by transcription factor family.

4.2.2 Model architectures

Five model architectures were evaluated, four transformer models, of those two were BERT including GROVERand DNABERT-2, and two Nucleotide Transformer models, Nucleotide Transformer 500m Human Reference (NT-HUMAN), and Nucleotide Transformer v2 500m Multi Species (NT2-MULTI), which adopt the ESM-style encoder architecture but are pre-trained on DNA sequences. As a baseline CREsted deep topic model architecture CNN (Kempynck, N. et al. 2025) was also trained as a baseline. These models varied in pretraining strategy, architecture, and input encoding approaches (Table 4.1), and were selected to represent the range of

available DNA foundation models, with the CNN providing a non-pretrained baseline.

GROVER was based on the BERT architecture, was pretrained on the human reference genome using masked k-mer prediction and next-sentence prediction objectives. A BPE-based tokeniser with a fixed vocabulary of 609 tokens learned via next-k-mer prediction was used, with tokens spanning up to 12 base pairs. This allows a theoretical sequence coverage of 6,144 bp (512 tokens × 12 bp), though practical coverage was often lower due to token segmentation. GROVER comprises approximately 87 million trainable parameters, consistent with its configuration as a 12-layer BERT model with a hidden size of 768 and 12 attention heads per layer. The majority of parameters are concentrated within the transformer blocks, each of which includes a multi-head self-attention mechanism and a feedforward network. For each layer, the attention sublayer includes three linear projections for the query, key, and value matrices (each of size 768 × 768), as well as an output projection of the same size, contributing roughly 2.36 million parameters per layer. The feedforward network further contributes approximately 4.72 million parameters per layer, based on two linear transformations between the hidden size (768) and the intermediate size (3072). Across all 12 layers, these components together account for approximately 85 million parameters. The remaining parameters derive from the token embedding matrix (609 tokens × 768 dimensions), the positional embeddings (512 × 768), and LayerNorm layers, summing to an additional $\sim$0.9 million parameters. Thus, the model's total parameter count is dominated by its core transformer architecture, with relatively minor contributions from the embedding and normalisation components.

DNABERT-2, also based on the BERT architecture and pretrained on multi-species genomic sequences using a Byte Pair Encoding (BPE) tokeniser trained via SentencePiece tokenisation. A vocabulary of approximately 4,096 variable-length tokens is learned and replacing fixed k-mers with a data-driven representation. This model is optimised for generic sequence-based prediction tasks and used ALiBi positional encoding to support extended input without learned embeddings. DNABERT-2 contains approximately 117 million parameters, following a standard BERT architecture with 12 transformer layers, each comprising 768 hidden units and 12 self-attention heads. As in GROVER, the bulk of parameters are concentrated in the attention and feedforward sublayers within each transformer block. Each layer's attention mechanism includes projections for the query, key, value, and output matrices, each of size 768 × 768, contributing around 2.36 million parameters per layer. The position-wise feedforward network, with intermediate dimensionality of 3072, adds

approximately 4.72 million parameters per layer. Altogether, these account for approximately 85 million parameters across the 12 layers. The increased parameter count in DNABERT-2 relative to GROVER is largely due to its larger vocabulary (4,096 tokens vs. 609 in GROVER), which increases the size of the embedding matrix to over 3.1 million parameters (4096 × 768). Additional parameters arise from positional embeddings (typically 512 × 768), LayerNorm layers, and classification or pretraining heads. These collectively contribute the remaining $\sim$29 million parameters, resulting in a total of approximately 117 million. This expanded capacity is intended to support greater generalisation across multi-species genomic inputs and varied downstream tasks.

NT-HUMAN was implemented as an ESM-style encoder, pretrained on a vocabulary of non-overlapping 6-mers and singleton nucleotides (Adenine (A), Cytosine (C), Guanine (G), Thymine (T), Any Nucleotide (N)) derived from the human genome. A maximum of 1,000 input tokens was supported, corresponding to 6,000 bp of sequence. NT-HUMAN contains approximately 480 million parameters, consistent with its 24-layer ESM-style architecture using a hidden size of 1280 and 20 attention heads per layer. Each transformer layer includes self-attention and feedforward sublayers, with attention projections contributing approximately 6.55 million parameters per comprising three attention projections (query, key, value) and one output projection, each of size 1280 × 1280, and the feedforward network adding a further $\sim$13.1 million parameters per layer (intermediate size 5120). Together, these account for roughly 472 million parameters across all 24 layers. The remaining $\sim$8 million parameters include the embedding matrix (4101 tokens × 1280 dimensions), positional embeddings (1000 × 1280), LayerNorm layers, and any classification or output heads.

NT2-MULTI, which used the same vocabulary as NT-HUMAN, was pretrained on a multi-species corpus and incorporated rotary positional embeddings, allowing for an extended input length of 2,048 tokens (12,288 bp). This design was intended to improve generalisation across divergent genomic regions. NT2-MULTI extends this architecture to 29 layers with a hidden size of 1024 and 16 attention heads per layer, resulting in $\sim$494 million total parameters. Each layer's attention module contributes $\sim$4.2 million parameters (4 × 1024 × 1024), while the feedforward network adds $\sim$8.4 million (1024 to 4096 and back). Across 29 layers, this yields $\sim$364 million parameters, with the remaining $\sim$130 million split between the embedding layer (vocab size 4101), rotary positional encoding support, and task-specific heads. The model's increased depth and feedforward capacity are designed to capture more complex patterns, particularly when learning from diverse species.

The CNN (Kempynck, N. et al. 2025) received one-hot encoded inputs of shape 1,024 × 4, representing 1,024 base pair sequences across four nucleotide channels. The architecture consisted of five convolutional blocks with increasing filter depths and progressive max pooling, followed by two fully connected layers. Residual connections were applied in the final two convolutional blocks. Batch normalisation, GELU activation, and dropout (rate = 0.1) were used throughout. Mixed-precision training was enabled via the `mixed_float16` policy in TensorFlow to accelerate training. The final model contained 12.7 million trainable parameters.

To ensure fair and truncation-free benchmarking across all models, an input sequence length of 1,024 base pairs was selected. This length was chosen to remain within the 512-token limit imposed by DNABERT-2 and GROVER, below the 1,000-token cap of NT-HUMAN, and well within the extended capacity of NT2-MULTI. Transformer-based models were fine-tuned using Hugging Face's transformers library. tokenisation was performed using the associated Autotokeniser, and sequences were padded to a uniform maximum length of 512 tokens to minimise RAM usage during training. For the CNN model, fixed-length one-hot encoded inputs with dimensions 1,024 base pairs × 4 nucleotide channels were used in place of token embeddings, enabling direct representation of raw DNA sequences without tokenisation.

Table 4.1: Summary of model architectures evaluated for MLL binding prediction.

Model	NT-HUMAN	NT2-MULTI	GROVER	DNABERT-2	CNN
Architecture	ESM	ESM	BERT	BERT	CNN
Layers	24	29	12	12	5 Conv + 2 Dense
Hidden Size	1280	1024	768	768	1024
Number of Heads	20	16	12	12	N/A
Activation	GELU	SiLU	GELU	GELU	GELU
Positional Embeddings	Absolute	Rotary	Absolute	Absolute	N/A
Vocabulary Size	4101	4101	609	4096	N/A
Vocabulary Type	k-mer (6) + singleton A,C,G,T,N	k-mer (6) + singleton A,C,G,T,N	BPE (next-k-mer prediction)	BPE (SentencePiece)	One-hot (A,C,G,T)
Max Tokens	1000	2048	512	512	N/A
Parameters (M)	480	494	87	117	12.7
Pretraining Data	Human genome	Multi-species genomes	Human genome	Multi-species genomes	N/A

4.2.3 Fine-tuning and evaluation

Convolutional neural network baseline

A CNN baseline was implemented using the CREsted DeepTopic architecture. Input sequences of 1,024 bp were one-hot encoded across four nucleotide channels. The network comprised five convolutional blocks with increasing filter depths and max-pooling, residual connections in the final two blocks, and batch normalisation, ReLU activations, and dropout (rate = 0.1) throughout, followed by two fully connected layers. Training was performed in TensorFlow 2.18 with mixed-precision enabled on one AWS g5 A10G instance. Optimisation used Adam (learning rate = 1×10^{-3}, weight decay = 0.01, warmup ratio = 0.1). A per-device batch size of 8 was used with gradient accumulation to an effective batch size of 32. Models were trained for up to five epochs with early stopping (patience = 3). Progress was logged with TensorBoard.

Transformer models and LoRA fine-tuning

Four pretrained DNA language models were benchmarked: GROVER, DNABERT-2, NT-HUMAN, and NT2-MULTI. Models were instantiated with Hugging Face's AutoModelForSequenceClassification using problem_type = "multi_label_classification", with output dimensionality equal to the number of assays. Sequences were tokenised with model-specific fast tokenisers and padded to 512 tokens.

Optimisation used AdamW (learning rate = 2×10^{-5}, weight decay = 0.01, warmup ratio = 0.1). Per-device batch size was 8 with gradient accumulation of 4, yielding an effective batch size of 32. Models were trained for up to five epochs with early stopping patience of 3 evaluations. Mixed-precision (fp16 or bfloat16) was enabled where supported.

To reduce trainable parameters, LoRA was applied to self-attention layers using the Parameter-Efficient Fine-Tuning (PEFT) library. LoRA introduced trainable low-rank matrices (rank = 8, α = 16, dropout = 0.1) alongside frozen attention weights. Only LoRA-injected parameters and classification heads were updated. This reduced the proportion of trainable parameters to 0.36-0.89% of the full model while retaining competitive performance (Table 4.2). LoRA adapters and merged checkpoints were saved for inference.

Table 4.2: Comparison of model parameter counts and reduction achieved via LoRA fine-tuning.

Model	Target Modules	LoRA Params	Classifier Params	Trainable Params	Trainable (%)
GROVER	['query', 'value']	294,912	15,380	310,292	0.36
DNABERT-2	['Wqkv', 'wo']	663,552	15,380	678,932	0.58
NT-HUMAN	['query', 'value']	983,040	3,304,980	4,288,020	0.89
NT2-MULTI	['query', 'value']	950,272	2,119,700	3,069,972	0.62

Evaluation and threshold calibration

During training, model performance was assessed using threshold-independent metrics that summarise prediction quality across all possible decision thresholds. The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) measures the trade-off between the true-positive rate and false-positive rate, reflecting overall discriminative ability. The Precision-Recall (PR) AUC emphasises performance on the positive class by plotting precision versus recall, which is particularly informative under class imbalance. The Average Precision (AP) represents the area under the PR curve, equivalent to the weighted mean of precision values achieved at each recall level. The Harmonic Mean of Precision and Recall (F1) score, while sometimes reported at a fixed threshold, can also be summarised across thresholds as the harmonic mean of precision and recall, capturing the balance between sensitivity and specificity.

To obtain discrete predictions, per-label thresholds were optimised on validation logits to maximise F1 and frozen for test evaluation. Reported metrics included both threshold-free (AUC, AP) and thresholded (precision, recall, F1). Confusion matrices and prevalence baselines were plotted.

Embedding analysis

Embeddings from the final hidden states of fine-tuned transformers were extracted for chr9 test sequences. Token embeddings were pooled (mean), standardised, reduced with Principal Component Analysis (PCA), and projected with Uniform Manifold Approximation and Projection (UMAP) (n_neighbors = 30, min_dist = 0.1, cosine). Points were coloured by binding probabilities, feature type (Promoters vs Non Promoters), or log1p(RPKM), and faceted by cell line and replicate.

4.2.4 Attribution and motif discovery

Attribution analyses were performed to identify sequence features underlying model predictions. Layer Integrated Gradients were applied with Captum on word level embeddings using a di-nucleotide shuffled sequences as baseline with 50 integration steps. Token-level scores were calculated for true positive sequences and were then mapped back to base-pair resolution using tokeniser offsets. The resulting base-pair resolution attribution outputs, consisting of one-hot encoded sequences paired with attribution scores, were analysed with TF-MoDISco-lite (Av Shrikumar et al. 2022)(v2.4.0). For each label and polarity, up to 20,000 high-scoring subsequences (seqlets) were clustered into de novo motifs, which were then compared against the HOCOMOCO v12 database using Tomtom. Base-resolution attribution outputs, consisting of one-hot encoded sequences paired with attribution scores, were analysed using TF-MoDISco-lite (v2.4.0). For each label, the input comprised attribution score matrices and the corresponding one-hot encoded sequences. TF-MoDISco was run with a sliding window of 1,000 bp, flank size of 21 bp, and parameters tuned for clustering efficiency (`--trim-to-window 25`, `--flank-size 8`, `--gaps-allowed 10`, `--jitter 5`, `--lwmc 5`, maximum 4,000 seqlets per metacluster). Separate analyses were performed for each cell-type and replicate. Seqlets were clustered into de novo motifs. Motifs were then matched against the HOCOMOCO v12 database using Tomtom (Gupta, S. et al. 2007)for annotation, and motif TF classes.

Genomic annotation of predictions

Prediction outcomes (True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN)) were exported as BED files and annotated with ChIPseeker. Annotation distributions were visualised as stacked bar plots with ggplot2.

Integration with gene expression

RNA-seq expression values obtained from the Human Protein Atlas (Uhlén, M. et al. 2005). Transcript-level Transcripts Per Million (TPM) values for the five haematopoietic cell lines matched to predicted promoter binding. Only motifs corresponding to TFs expressed above 1 TPM in at least one line were retained in enrichment analysis.

4.2.5 Software versions and environments

All analyses were performed in Python 3.10.12 and R 4.3.1. See Appendix for software and package versions (Tables B.3, B.4).

4.3 Results

4.3.1 Dataset Design and Predictive Task Formulation

This section outlines the dataset design and model formulation used to test whether DNA sequence features alone can predict MLL occupancy across both MLL and MLL-FP contexts. The design emphasised biological signal, minimised technical artefacts, and enforced stringent generalisation. The datasets comprise MLL-N CUT&Tag from six cell lines (Table A.1): RCH-ACV (E2A-PBX1; B-ALL; non-MLLr), SEM, RS4;11 and MV4-11 (each harbouring MLL-AF4), THP-1 (MLL-AF9; AML), and OCI-AML3 (AML without MLL fusions). OCI-AML3 serves as a second non-fusion control in a myeloid background; it carries a NPM1 exon-12 type-A mutation (NPM1c+) and a heterozygous DNMT3A mutation (Quentmeier, H. et al. 2005).

Including both lymphoid and myeloid MLLr lines, together with two non-fusion controls, allows assessment of which sequence features are conserved across contexts and which are specific to particular cellular or fusion backgrounds.

Assay choice and data labelling

Both CUT&Tag and ChIP-seq datasets were available for MLL. CUT&Tag was chosen to define binding labels because it provides higher spatial resolution and lower background than conventional ChIP-seq, producing a clearer separation between bound and unbound regions (Kaya-Okur, H. S. et al. 2019). In CUT&Tag, a protein-specific antibody recruits a protein Tn5 transposase fusion that tagmentates DNA *in situ*, generating fragments precisely at binding sites and substantially reducing off-target signal compared with ChIP-seq. For example, CUT&Tag profiling of CTCF achieves $\sim$80 bp footprints, whereas ChIP-seq typically yields broader 150-300 bp enrichments, illustrating the gain in spatial precision (Kaya-Okur, H. S. et al. 2019). This punctate distribution yields more bimodal enrichment distributions suitable for confident binary labelling. Especially at TSS, where CUT&Tag provides clear peaks directly over the TSS where ChIP-seq showed promoter occlusion at these sites (Figure 4.3).

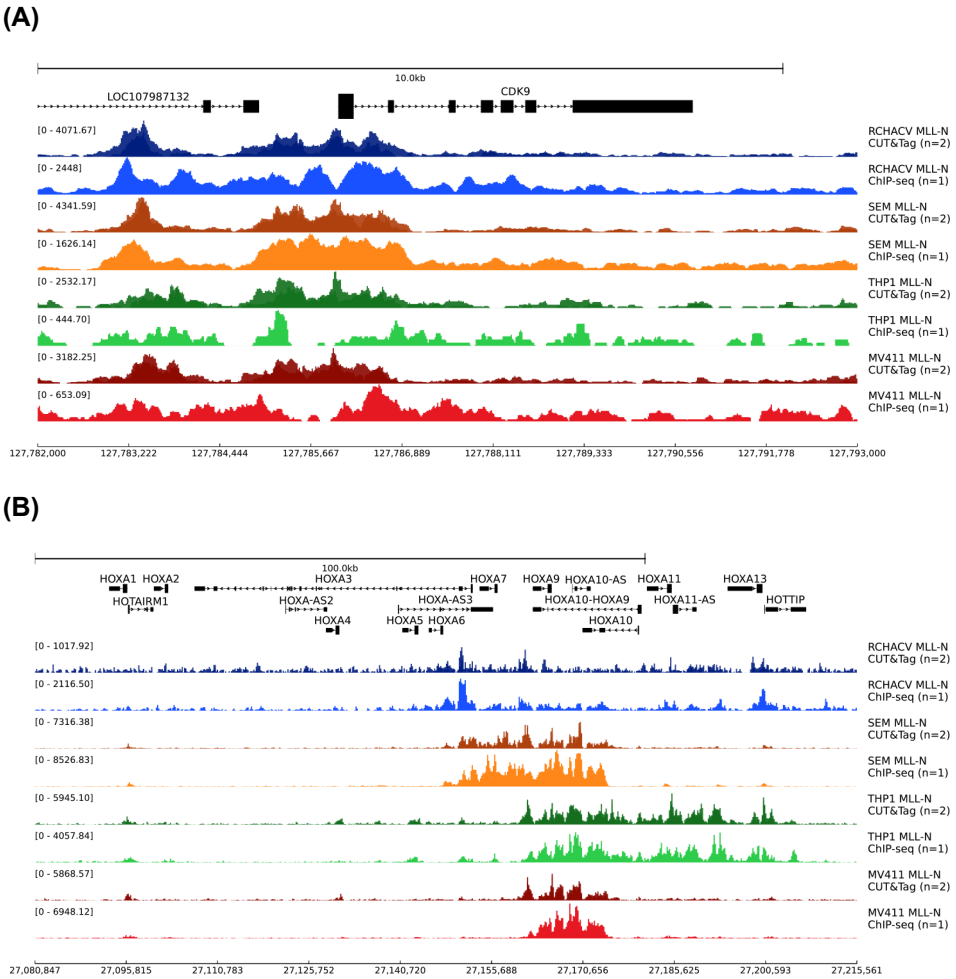


Figure 4.3: Comparison of MLL-N CUT&Tag (darker tracks) and ChIP-seq (lighter tracks) at the **(A)** *CDK9* locus (chr9:127,782,000-127,793,000) and **(B)** *HOX* gene cluster (cchr7:27,080,847-27,215,561) in four cell lines. The CUT&Tag data show sharp, punctate peaks with low background with clear signal especially at the TSS.

Region choice

Modelling was carried using two main datasets generated from the CUT&Tag data, these were promoter regions and methylome regions. Promoter windows served as a smaller dataset useful for benchmarking. These regions were derived from NCBI RefSeq gene annotation (Pruitt, K. D. et al. 2014) with promoter regions defined from transcription start sites $\pm$ 512 bp.

MLL preferentially binds at unmethylated, CpG-rich promoters and gene bodies. Regions from the Twist Human Methylome panel (Twist Bioscience 2022) were therefore used as the dataset regions. This choice offered several advantages including enrichment for CpG-dense loci where MLL binding is biologically plausible, reducing the search space from tens of millions of genome-wide tiles to a few

hundred thousand informative regions, and reduced bias compared with ATAC-seq peaks, which reflect chromatin accessibility and would exclude bound sites in less accessible CpG islands.

To ensure generalisability, I used 1,024 bp tiled regions over chromosome 9 as an additional test dataset, excluding blacklisted regions, which are highly repetitive and poorly mappable (Amemiya, H. M. et al. 2019). All analyses used the human reference genome hg38/GRCh38, genomic coordinates were defined on hg38 and DNA sequences were retrieved from the hg38 fasta (Schneider, V. A. et al. 2017).

Windowing and normalisation

All sequences were represented as fixed 1,024 bp windows, a length chosen to balance biological interpretability with computational constraints. Biologically, this captures the typical CpG structure and short-range motif combinations relevant to MLL recruitment, and the extra 24 bp beyond 1 kb is negligible. Computationally, using a power-of-two length is slightly more efficient for CNN models avoiding internal padding and aligning with Graphics Processing Unit (GPU) memory blocks and remains fully compatible with transformer token limits. Window lengths of 512 bp and 2,048 bp were also tested: 512 bp under-captured regulatory context and reduced performance, whereas 2,048 bp caused token truncation in DNABERT and GROVER and increased training computation without benefit (Figure 4.4).

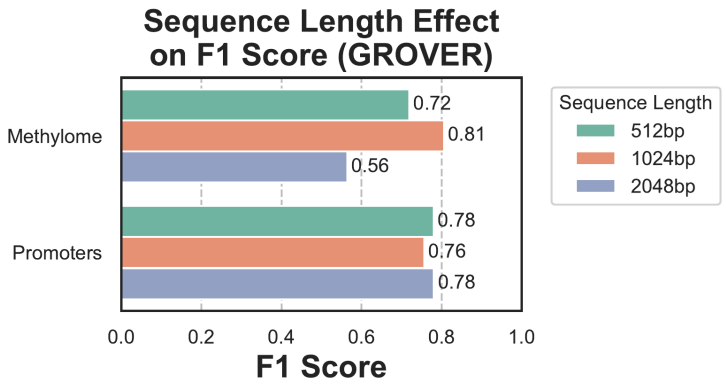


Figure 4.4: Effect of input sequence length on model performance. GROVER was trained on promoter and methylome regions with input lengths of 512, 1,024, and 2,048 base pairs as a multi-label task to predict MLL binding. Performance was evaluated on a held-out test set (chr9) using the F1 score. The 1,024 bp input length provided the best balance between capturing relevant sequence context and computational efficiency, avoiding token truncation issues.

All MLL-N CUT&Tag samples were prepared from a standardised input of 5×10^4 nuclei (Chapter 2) and yielded 25.7–41.9 million read pairs with 94–96% alignment

rates, above the sequencing depth at which CUT&Tag signal reaches saturation ($\sim$3 million reads (Kaya-Okur, H. S. et al. 2019)). The same MLL-N antibody was used across all cell lines. Signal bigWigs were generated with RPKM normalisation (bin size 1 bp), correcting for library depth at the track level. Per-sample signals were then upper-quartile scaled and log-transformed to account for residual efficiency differences between experiments, and enrichment thresholds were applied within each sample's own signal distribution (Section 4.3.1), making the labelling inherently robust to any between-sample sensitivity differences without requiring read subsampling.

Binarisation and label validation

Regions above the 80th percentile of normalised signal were labelled bound and those below the 20th percentile unbound, excluding the middle fraction. These thresholds balanced positives and negatives classes while avoiding ambiguous intermediate values.

Regression versus classification

Initial attempts to predict continuous CUT&Tag signal intensities using regression heads resulted in unstable training and poor generalisation. This likely reflects the heavy-tailed, zero-inflated distribution of CUT&Tag signal, as well as the contribution of non-sequence factors such as chromatin context and copy number variation. Reformulating the task as binary classification provided more stable and discriminative fine-tuning by converting noisy continuous measurements into discrete binding states. This approach more directly addresses the biological question of identifying sequence features associated with MLL occupancy, rather than modelling absolute binding strength, which is substantially influenced by chromatin accessibility, nucleosome positioning, co-factor availability, and protein dynamics. Given that occupancy varies across cellular contexts, the task was implemented as a multi-label classification problem, assigning one binary label per each replicate of each cell type. This design captures context-dependent recruitment while enabling direct comparison of shared and cell-type-specific sequence determinants.

Data partitioning and leakage control

Sequence leakage occurs when sequences in the test set are highly similar to those in the training set, causing artificially inflated performance because the model is

effectively evaluated on data it has already seen in a closely related form. In genomics, this can arise from multiple sources: direct proximity (neighbouring genomic regions share sequence context), paralogous gene families (genes with shared evolutionary origin have near-identical promoter sequences), and repetitive elements (transposons and satellite repeats recur genome-wide). Chromosome-based splitting effectively addresses local proximity and limits repeat-based leakage, as held-out chromosomes are largely distinct in sequence context from training chromosomes. For gene families with paralogous members distributed across multiple chromosomes, however, some leakage may remain as related sequences can appear in both training and test sets. Chromosome 8 was held out for validation and chromosome 9 for testing, both chosen for their representative GC content and gene density. This leave-one-chromosome-out design is the established best practice for controlling sequence leakage in genomic ML (Whalen, S. et al. 2022).

Class balance and negative sampling

MLL binding is sparse relative to candidate regions, creating substantial class imbalance. To balance training data, regions were selected from the upper and lower quartiles of normalised CUT&Tag signal with the top quartile representing bound regions and the bottom quartile, unbound. This quantile-based balancing maintains a biologically meaningful contrast while preventing trivial majority-class predictions and keeping metrics comparable across datasets (Whalen, S. et al. 2022). Genome-wide negatives were not used, as they are dominated by CpG-poor intergenic regions and would cause the model to learn GC content rather than true motif features.

Together, these design choices ensured that any predictive power observed reflects intrinsic DNA-sequence features relevant to MLL recruitment rather than artefacts of assay design, chromatin accessibility, or data leakage.

4.3.2 Dataset Distribution and labelling

Three separate datasets were used to train and evaluate the model. Promoter regions were derived from NCBI Refseq gene annotations and collapsed by transcript to give 19,860 individual promoter regions. Methylome regions were derived from the Twist Methylome panel. These 515,400 regions were annotated to genomic feature which 21.8% were located at promoters. In addition to using the hold out chromosomes for evaluation (chr8) and testing (chr9), the entirety of chr9 was tiled to 1024 bp and blacklisted regions were excluded. 7% of these 108,820 regions

were located at promoters. This tiled chromosome was used to assess the models' ability to generalise to previously unseen sequences with more diverse sequence composition than the training sequences (Figure 4.5).

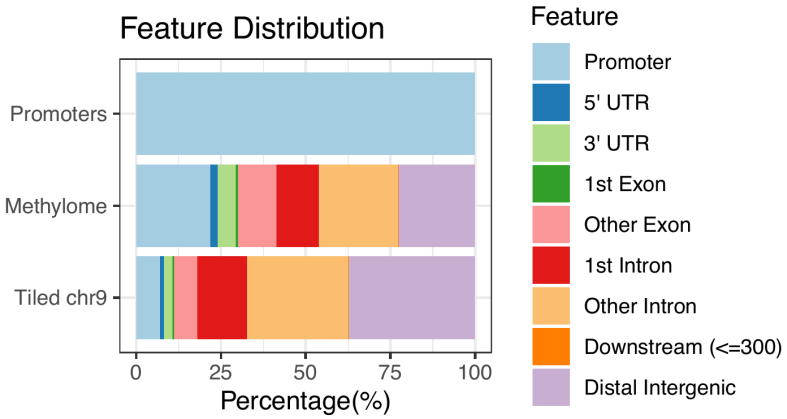


Figure 4.5: Promoter regions were defined as 1,024 bp windows centred on the transcription start site (TSS) of RefSeq transcripts. Methylome Panel regions were defined by extending Twist methylome regions into fixed 1,024 bp windows. Tiled chromosome 9 regions were defined by tiling chr9 into fixed 1,024 bp windows, excluding blacklisted regions. Annotations were performed with ChIPseeker, and the distribution of genomic features is shown as stacked bar plots.

For each dataset, MLL-N signal was extracted from the corresponding CUT&Tag bigWig files (RPKM, binsize 1 bp) over the methylome and promoter regions (Figure 4.6A, 4.6B). Signal was transformed as log(1+RPKM). To mitigate scaling differences across samples, upper-quartile (75th percentile) size-factor normalisation was applied per sample, and then quantile-based binarisation was performed on the normalised log values with regions in the upper 20% labelled as bound, and the lower 20% unbound, and the middle 60% were excluded from training to avoid ambiguity (Figure 4.6C, 4.6D). This produced class-balanced datasets, so models were trained using standard Binary Cross-Entropy (BCE) loss which measures the deviation between a model's predicted probability and the true binary label (Figure 4.6E, 4.6F). For the methylome dataset this yielded $\sim$25,000 bound and $\sim$20,000 unbound regions per label (total unique regions after ambiguous regions excluded, 45,090), split by chromosome hold out into 40,727 training regions, 2,949 evaluation regions (chr8), and 1,414 test regions (chr9). For the promoter dataset, there were $\sim$1,616 bound and $\sim$1,414 unbound regions per label (total unique regions 3,030), split into 2,751 training regions, 178 evaluation regions (chr8), and 101 test regions (chr9) (Table 4.3).

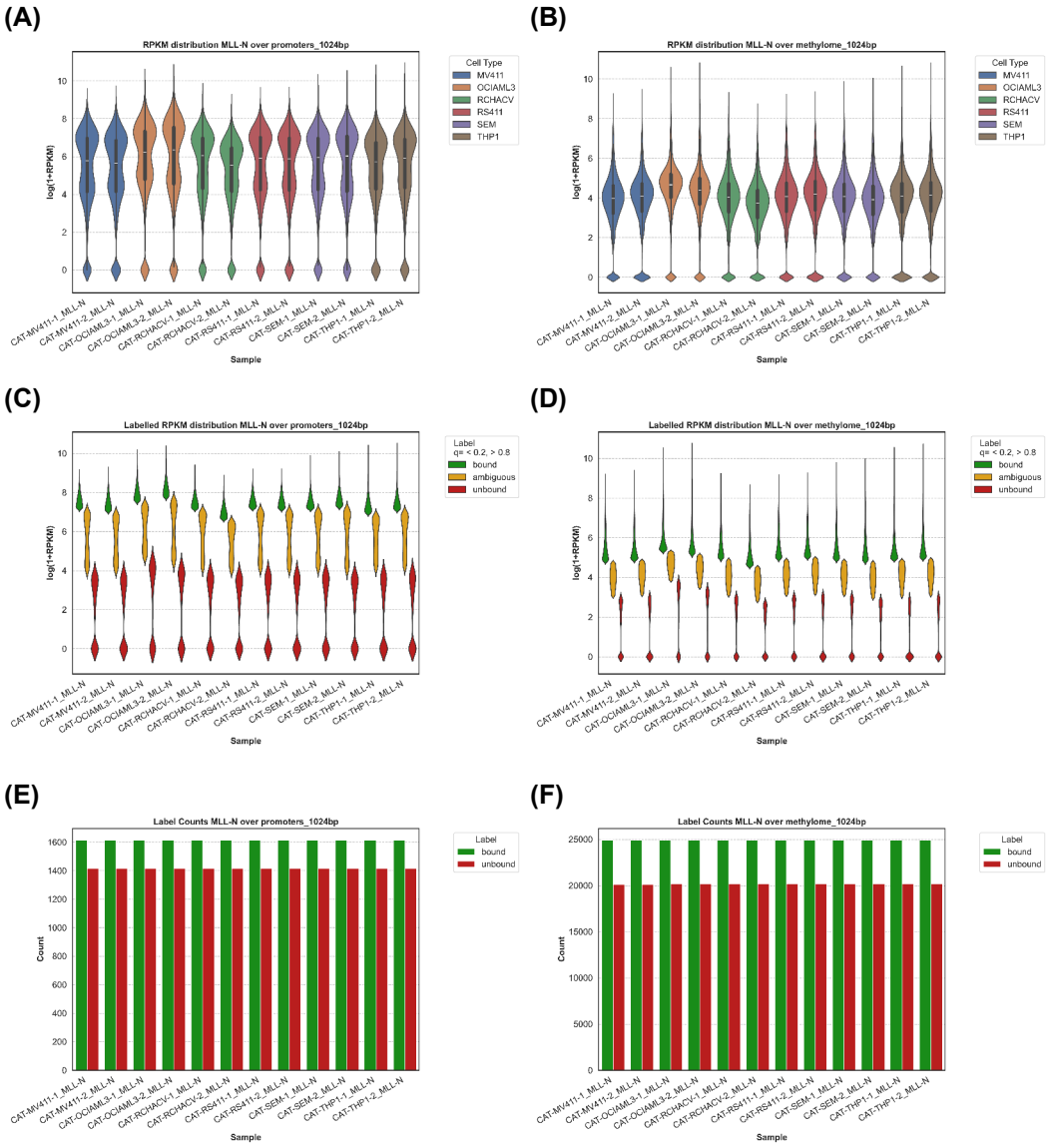


Figure 4.6: (A-B) Violin plots showing the distribution of log(1+RPKM) signal for MLL-N CUT&Tag over promoter regions and methylome regions **(B)**. Each colour corresponds to a different cell line. **(C-D)** Violin plots showing the distribution of log(1+RPKM) signal after upper-quartile normalisation and quantile-based binarisation into bound (green), unbound (red), and ambiguous (yellow) classes for promoter regions **(C)** and methylome regions **(D)**. **(E-F)** Bar plots showing the number of bound and unbound labels per sample after binarisation bound (green), unbound (red) for promoter regions **(E)** and methylome regions **(F)**.

Table 4.3: Summary of dataset sizes and chromosome-based splits for training, evaluation, and testing.

Dataset	Training	Validation (chr8)	Test (chr9)
Methylome	40,727	2,949	1,414
Promoters	2,751	178	101

4.3.3 Model Benchmarking

To determine the model best suited for predicting MLL-N binding from DNA sequence, four DNA-LM and one CNN were benchmarked. As some DNA-LM are pretrained on multiple species, two tasks of varying complexity were used for evaluation. The simpler task involved predicting binding over promoter regions, while the more complex task used methylome regions, both defined as 1024 bp windows. This design allowed assessment of whether models pretrained specifically on human DNA outperformed multispecies-pretrained models when applied to less conserved genomic regions. To ensure comparability across architectures, all models were fine-tuned under the same conditions (maximum 5 epochs, fixed batch size, and early stopping based on validation loss). As all transformer models were fine-tuned using LoRA, only a small set of adapter parameters were updated rather than full model weights, meaning convergence was reached well within the 5-epoch maximum.

Model performance after fine-tuning for MLL-N binding (multilabel classification)

On the promoter test dataset (Figure 4.7A Top right), performance was uniformly high across transformer models. NT2-MULTI (ROC AUC = 0.991) and DNABERT-2 (0.991) were marginally higher than GROVER (0.985) and NT-HUMAN (0.966), while CNN was lower at 0.922. Promoters are GC-rich and motif-dense, providing highly stereotyped sequence features that most architectures could readily capture, leading to near-perfect performance.

In contrast, the methylome test dataset (Figure 4.7A Top left) presented a more complex prediction task due to its more heterogeneous GC composition and broader motif diversity. Here, GROVER achieved the highest ROC AUC (0.859), outperforming DNABERT-2 (0.853), NT2-MULTI (0.846), CNN (0.825), and NT-HUMAN (0.819). This indicates that GROVER's human-specific pretraining and model archi-

tecture were better suited to handling the variability of methylome regions. Although differences on the methylome test set were modest, they reflect converged models, with all runs stopping well within the 5-epoch maximum.

The advantage of GROVER was most apparent in generalisation to chr9 tiled regions (Figure 4.7A Bottom). For promoters, performance declined across all models, with CNN (0.936) highest, but GROVER (0.750) remained substantially stronger than DNABERT-2 (0.464), which dropped dramatically when used on tiled Chr9 regions, indicating a failure to generalise beyond the training distribution. For methylome regions, GROVER achieved the highest ROC AUC (0.949), outperforming CNN (0.933), NT-HUMAN (0.928), and NT2-MULTI (0.928), while DNABERT-2 again dropped sharply (0.406).

Taken together, these results show that while multiple models can solve the relatively simple promoter classification task, GROVER performed best on methylome prediction, which draws on more heterogeneous DNA sequences, and achieved the highest AUC when predicting over tiled chr9 methylome regions. The higher AUC on tiled chr9 regions for the methylome-trained model compared to the promoter-trained model reflects the greater sequence diversity of methylome training data, which generalises more readily to full-chromosome tiling than models trained on the more conserved sequence context of promoter regions. GROVER's strong performance on the methylome task and its stability when predicting over tiled chr9 regions make it the most suitable model for MLL-N binding prediction. Per-label ROC curves further highlighted differences in generalisation on chr9 tiled methylome regions. Although it was the second best performer on the methylome dataset with decent per label ROC (AUC range 0.83–0.86; Figure 4.7B), DNABERT-2 showed substantial variability between labels (AUC range 0.11–0.67; Figure 4.7C), with several cell lines performing close to random. This illustrates that while DNABERT-2 performed well on both the promoter and methylome test sequences with similar composition to the training set it was unable to generalise to more heterogeneous sequence contexts. This may be due to its multispecies pretraining, which could have biased it towards conserved promoter features at the expense of more variable regions. GROVER, by contrast, maintained both high accuracy and stability across all samples, underscoring its robustness on the more complex prediction task, and was therefore selected as the pretrained model for downstream multi-label classification per cell line under the same training protocol.

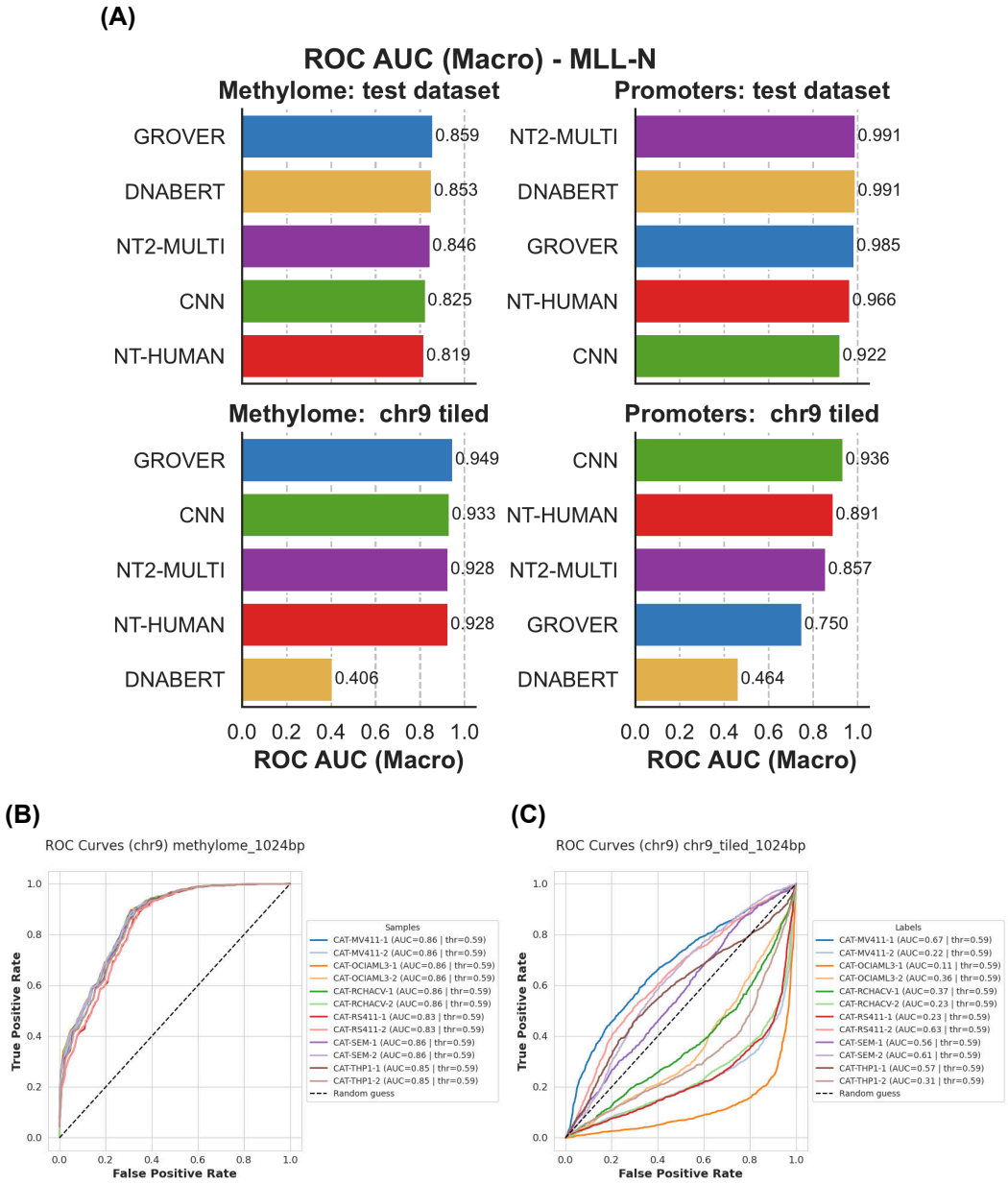


Figure 4.7: (A) Model performance by Macro ROC AUC for both methylome and promoter datasets with evaluation over chr9 of the training set and tiled chr9 regions. **(B)** DNABERT2 per label ROC curves over the methylome chr9 regions (AUC range 0.83–0.86). **(C)** DNABERT2 per label ROC curves over the tiled chr9 regions (AUC range 0.11–0.67)

4.3.4 MLL Binding Across Cell Lines

MLL-N Predictions per label for methylome generalised to chr9 tiled 1024bp regions

GROVER, which outperformed the other DNA language models, was selected for downstream prediction of MLL-N binding across the chromosome 9 tiled dataset. Training loss curves for GROVER (Figure 4.8) show that validation loss stabilised within the first few training epochs while training loss continued to decline, indicating onset of overfitting. However, early stopping and selection of the best-performing checkpoint ensured the final model was not adversely affected.

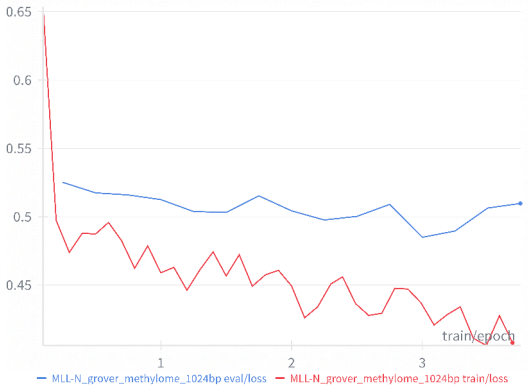
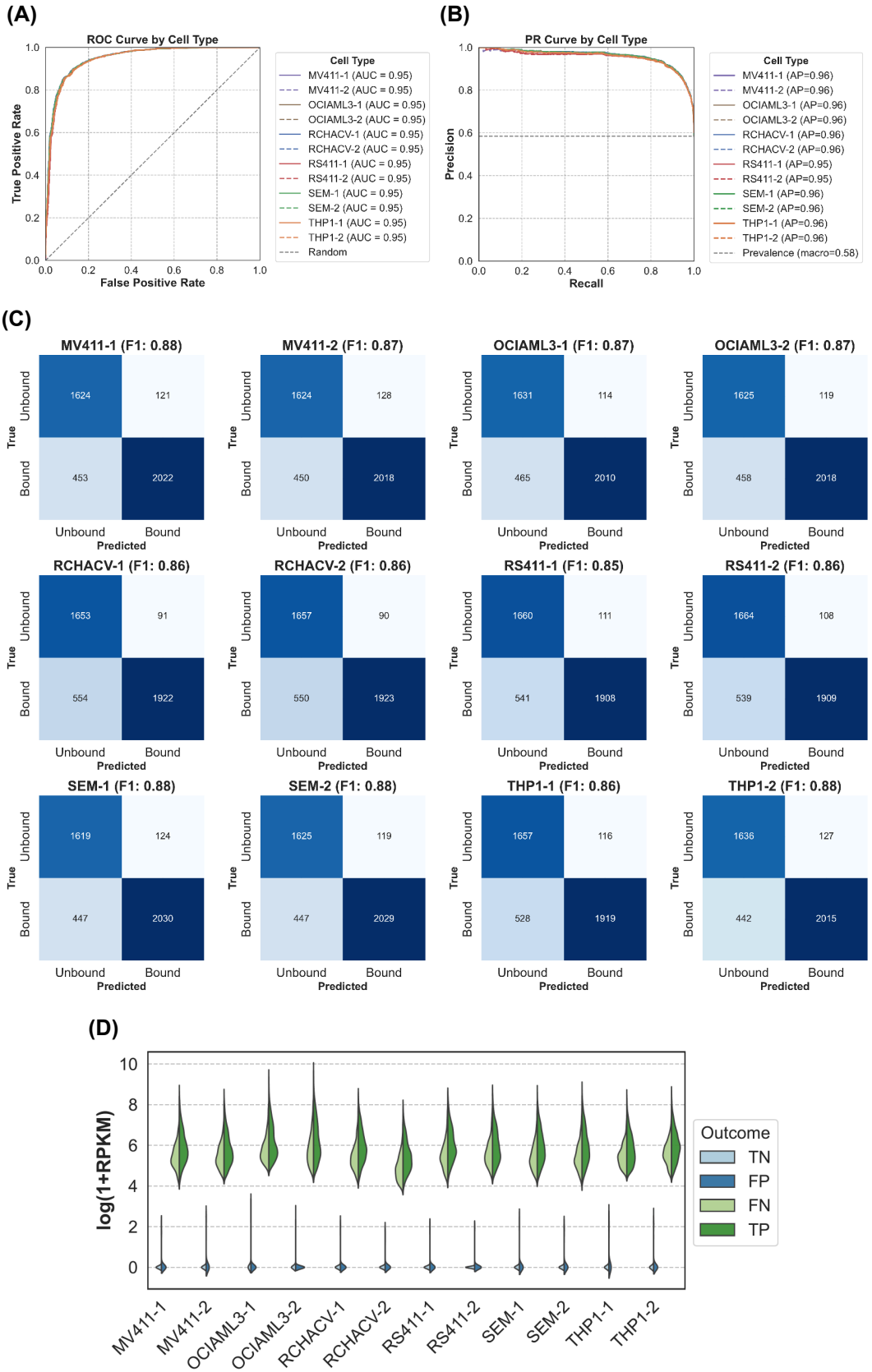


Figure 4.8: GROVER training (Red) and validation (Blue) loss curves for MLL-N binding prediction on methylome regions. Validation loss stabilised within the first few training steps, confirming that the 5-epoch training was sufficient to capture generalisation performance.

The model trained on the methylome regions achieved a high overall performance with a mean ROC AUC of 0.95 across all labels (Figure 4.9A), PR AUC ranging from 0.95–0.96 (Figure 4.9B), and per-label F1 scores ranging from 0.85 to 0.88 (Figure 4.9C). RPKM distributions by prediction outcome showed that false negatives retained bound-level signal and false positives had near-zero signal across all cell lines (Figure 4.9D). Further examination of predictions revealed systematic patterns in misclassifications (Figures 4.9E, 4.9F). False negatives were enriched in intronic and distal intergenic regions with lower GC content (50%) than true positives (60%), suggesting the model failed to capture MLL-N occupancy at regions with lower CpG density outside canonical promoter contexts. False positives showed GC content comparable to true positives but had reduced promoter enrichment relative to true positives, with more intronic and distal intergenic regions, indicating that CpG-rich sequence features alone are insufficient to predict occupancy at distal sites. Consistent with biological expectation, GROVER correctly identified MLL-N

binding predominantly at promoter regions, and true negatives were predominately distal intergenic regions.

As the model was fine-tuned as a multi-label classifier, it was possible to examine cell line-specific binding patterns. Overall, the model predicted similar binding patterns across cell lines, with a high degree of overlap in predicted bound regions especially at targets of MLL such as *CDK9* (Figure 4.10A) and *GNAQ* (Figure 4.10B). This reflects the shared underlying DNA sequence features learned by the model. However, the model did not capture regions where MLL binding spread into the gene body as it does at *GNAQ* for the cell lines wit h MLL-AF4 (MV4;11, RS4;11 and SEM cells) (Figure 4.10B).



Continued on next page.

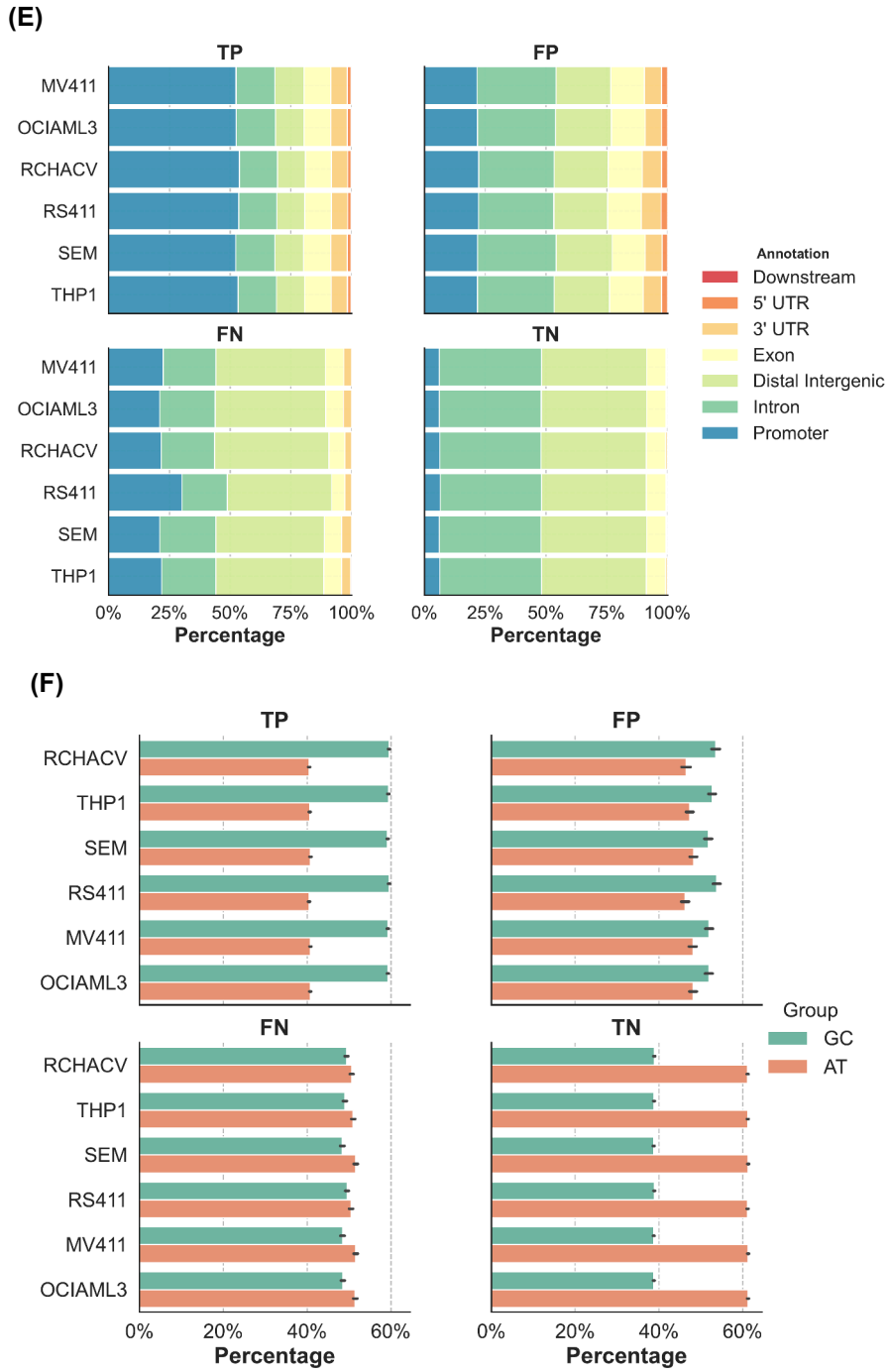


Figure 4.9: (Previous page) GROVER predictions of MLL-N binding across chromosome 9 tiled 1,024 bp regions. **(A)** ROC (mean ROC AUC ≈ 0.95). **(B)** PR (macro PR AUC ≈ 0.95–0.96). **(C)** Per-label confusion matrices (F1 ≈ 0.85–0.88). **(D)** log(1+ RPKM) by outcome. **(E)** By genomic annotation. **(F)** By GC/AT bias.

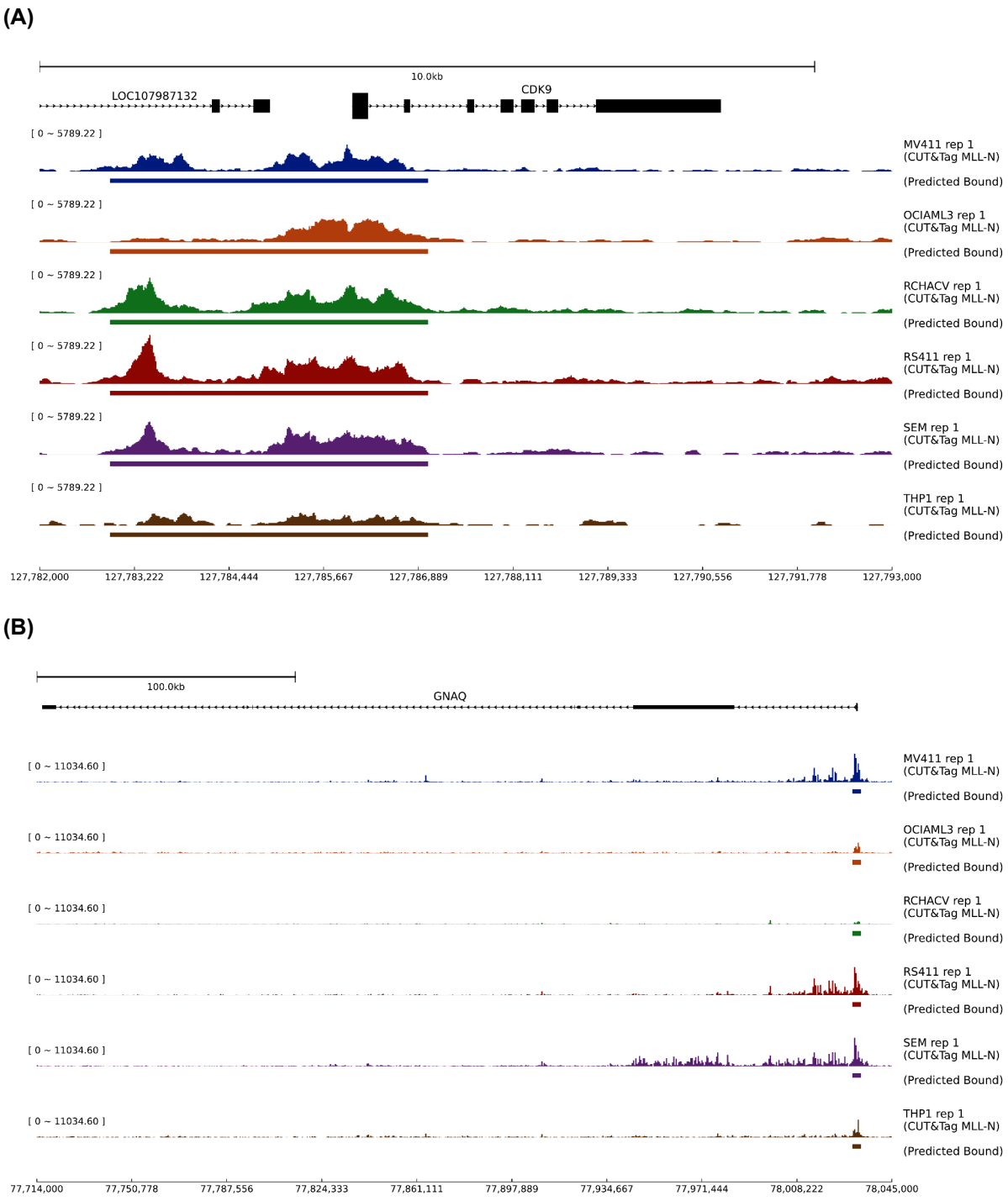


Figure 4.10: Examples of GROVER predictions of MLL-N binding across chromosome 9 tiled 1,024 bp regions. **(A)** *GNAQ* (chr9:77,714,000-78,045,000) **(B)** *CDK9* (chr9:127,782,000-127,793,000) Each panel shows the RPKM signal for each cell line replicate from CUT&Tag (top), the predicted binding shown underneath.

Model Embeddings

To assess whether the fine-tuned GROVER model captured meaningful structure sequence embeddings were extracted from the fine-tuned GROVER model and projected into two dimensions with UMAP. Genomic annotation (promoter vs non-promoter Figure 4.11A) and log(1+RPKM) CUT&Tag (Figure 4.11B) signal were overlaid on the projection, revealing a clear separation in which promoter regions formed high-signal clusters while non-promoter regions occupied lower-signal areas. Predicted binding probabilities (sigmoid-transformed logits Figure 4.11C) were then mapped onto the same embedding; higher probabilities aligned with promoter/high-signal clusters and lower probabilities with non-promoter/low-signal regions. Collectively, these patterns indicate that the learned latent space captured biologically meaningful binding structure rather than superficial artefacts.

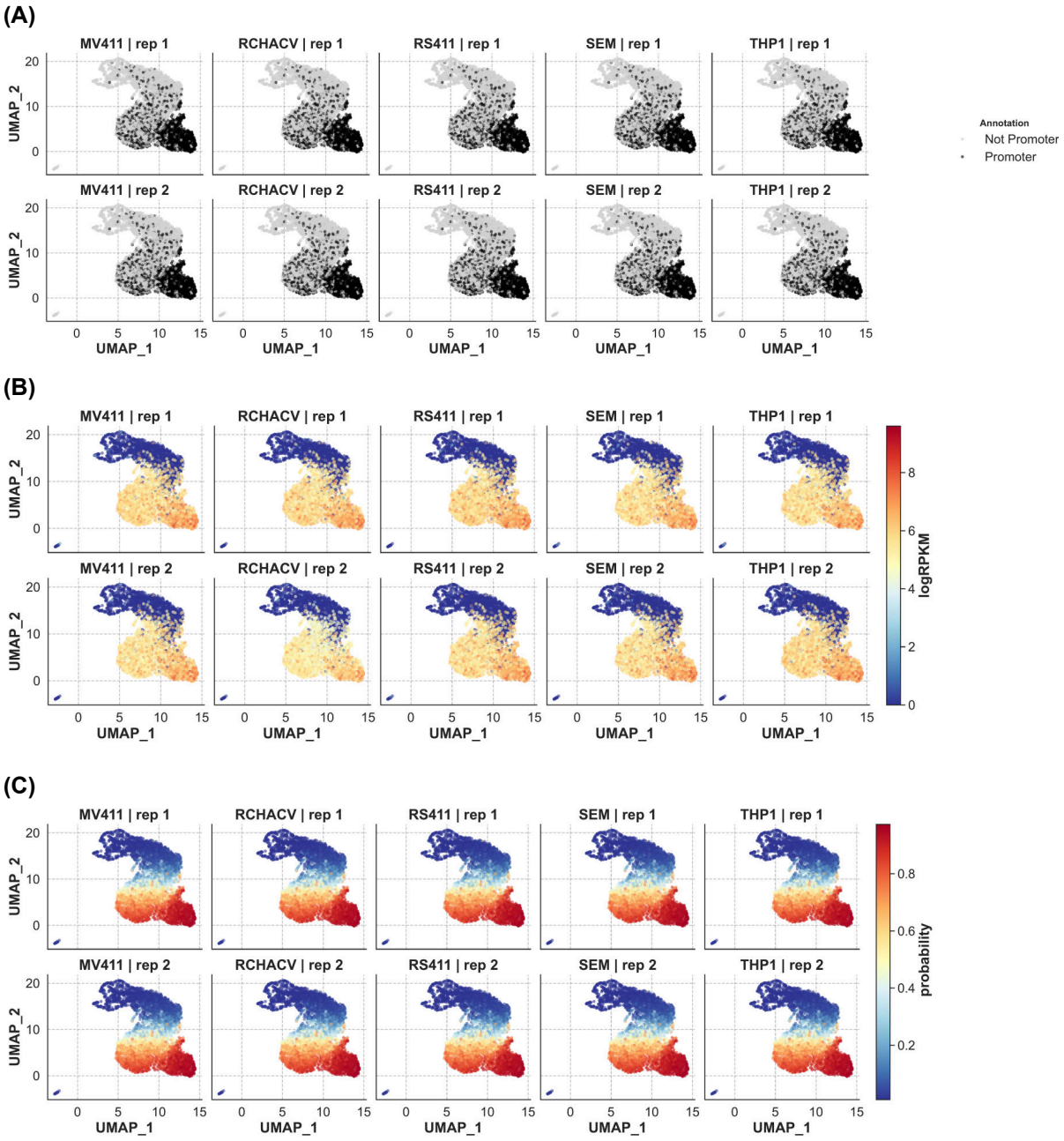


Figure 4.11: UMAP projections of sequence-level embeddings extracted from the fine-tuned GROVER model on chromosome 9 test regions, faceted by cell line (columns) and replicate (rows). **(A)** Points coloured by genomic annotation, with promoters highlighted in black and non-promoter regions in grey, illustrating the preferential alignment of MLL-N binding with promoter sequences. **(B)** Points coloured by quantitative CUT&Tag signal log(1+ RPKM), revealing continuous gradients of binding intensity that align with the embedding topology, consistent with the model capturing biologically meaningful variation in occupancy. **(C)** Points coloured by predicted binding probability (sigmoid-transformed logits), showing clear separation between bound and unbound regions.

Feature Extraction

To interpret sequence features learned by the classifiers, I focused on true positive tiled chr9 regions so that only correctly predicted binding sites contributed to the interpretation. I looked at two complementary levels: token-level attributions, which summarise the contribution of k-mer tokens (1–12 bp) across many loci, and nucleotide-level attributions (used later for motif discovery), which give base-resolution maps that can be aggregated into position-weight motifs. This split provides stable summaries at the token level which are too short for motif comparison while still capturing base-level detail for motif extraction.

Attributions were computed with Layer Integrated Gradients (LIG) at the embedding layer. LIG integrates gradients along a path from baseline to input, avoids saturation, and attributes directly to the learned DNA tokens. Sequence-matched di-nucleotide shuffled baselines controlled for composition and length, and 50 integration steps gave a good balance between accuracy and runtime.

Validation with RUNX1

To validate that the token attribution method (Layer Integrated Gradients) was able to identify known motifs, I finetuned a single label binary classification GROVER model using RUNX1 CUT&Tag data from a primary patient sample with MLL-AF4 (patient-22620, Table A.3). From a single CUT&Tag experiment, and fine-tuned for 3 epochs, the model achieved high performance on the held-out test set (chr9) with an F1 score of 0.82 and ROC AUC of 0.93 (Figure 4.12A-B).

Using Layer Integrated Gradients, the top 30 by mean absolute token attribution score were extracted from the test set and aligned to the known RUNX1 core motif (TGTGGT) using the Smith-Waterman (SW) algorithm (Smith, T. and Waterman, M. 1981). This approach identified several high-similarity tokens, including TGGGGC and TGTGGC within the top 14 of these 30 tokens, which closely match the RUNX1 motif (Figure 4.12C-D). Many of the other top tokens also contained partial matches to the RUNX1 motif, and were less than 6 bp long, indicating that they could be part of a longer motif instance. This validated that the feature attribution method was able to identify known motifs. However, unlike MLL motif identified by HOCOMOCO, RUNX1 motif is relatively short and can be captured within a single token by the BPE tokeniser used by GROVER.

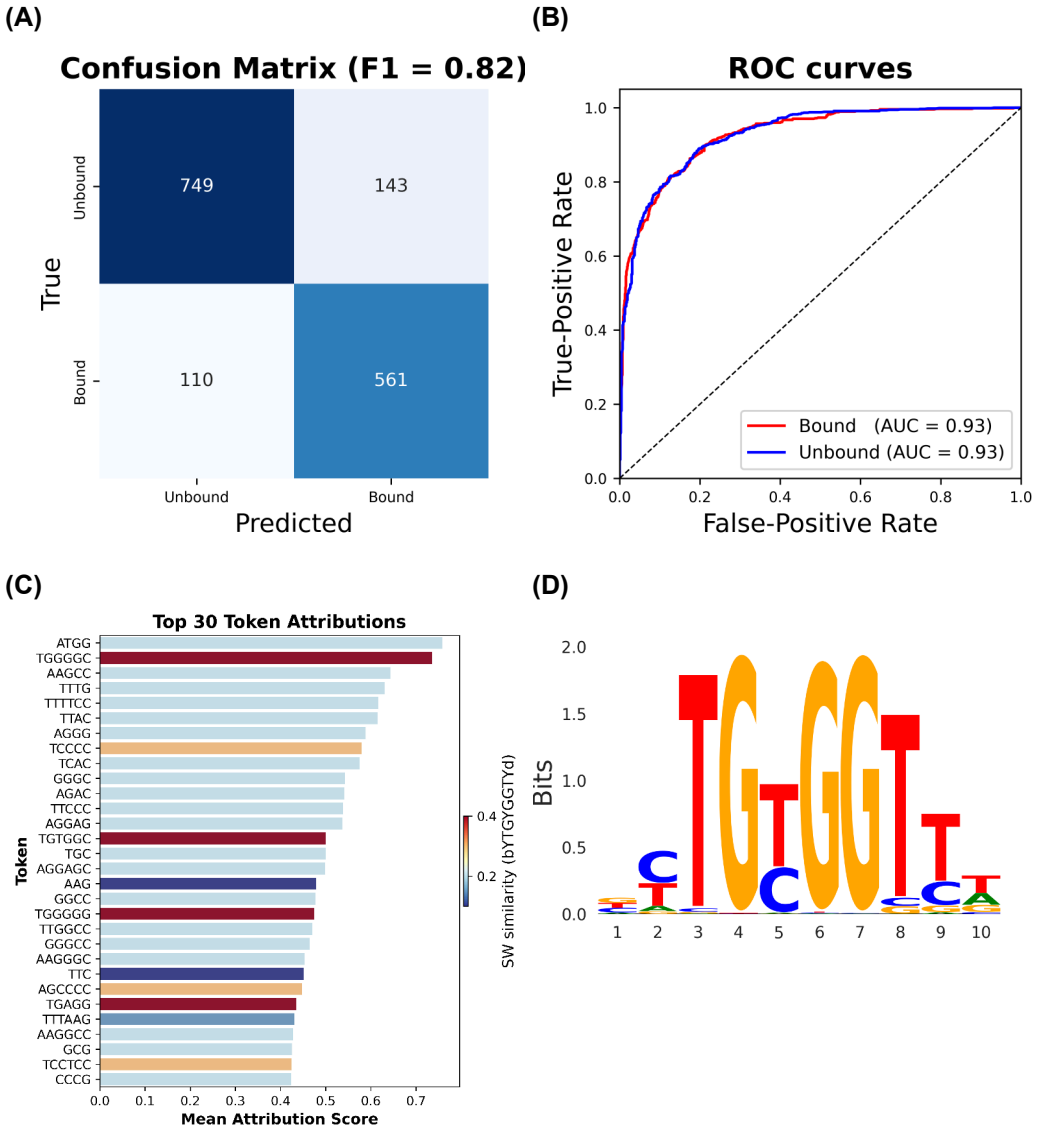


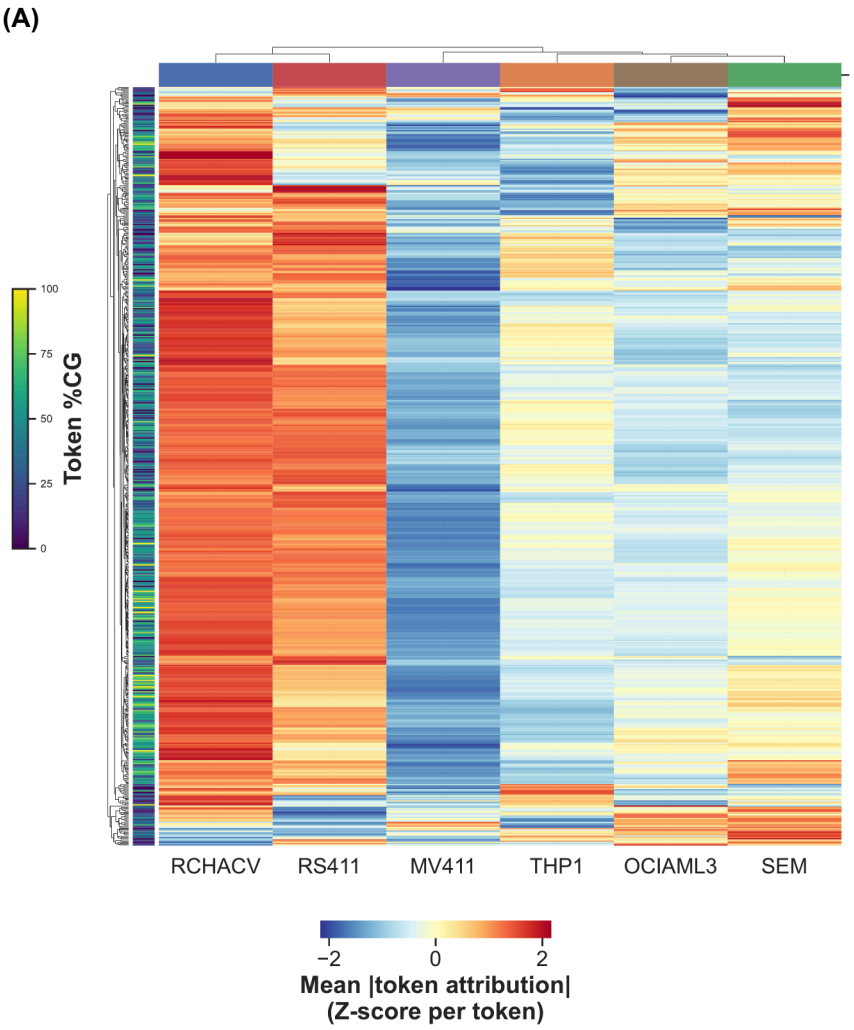
Figure 4.12: **(A)** Confusion matrix for RUNX1 binding prediction on promoter regions (F1 = 0.82). **(B)** ROC curve for RUNX1 binding prediction on promoter regions (AUC = 0.93). **(C)** Top 30 token attributions for RUNX1 binding predictions, coloured by Smith-Waterman (SW) similarity scores to the RUNX1 motif. **(D)** RUNX1 motif from HOCOMOCO database v12 from (Vorontsov, I. E. et al. 2024).

Token level attribution

For each cell line, token attributions were averaged across TP sequences and summarised as log2 fold-change over the other cell types (Log2 Fold Change (log2FC) One-Versus-Rest (OVR)). Tokens with low support (n < 100) were excluded, and values were Z-scored across lines for visualisation. This analysis revealed clear variability in token preferences (Figure 4.13A): RCH-ACV (MLL) and RS4;11 separated from MV4-11, OCI-AML3 and SEM, while THP-1 showed a distinct profile.

Enrichment of CG-rich tokens was observed particularly in tokens which had high magnitude attribution in RCH-ACV cells consistent with known MLL binding patterns through its CXXC domain.

RCH-ACV (MLL) showed the strongest positive correlation between token importance and GC content, consistent with CpG-island binding, whereas fusion-positive lines showed weaker or negative correlations (Figure 4.13B). When token frequency was taken into account, the most common GC-rich tokens were impactful across all lines, indicating that while CpG tokens particularly drove predictions in MLL, they also contributed in the MLL fusion contexts (Figure 4.13C).



Continued on next page.

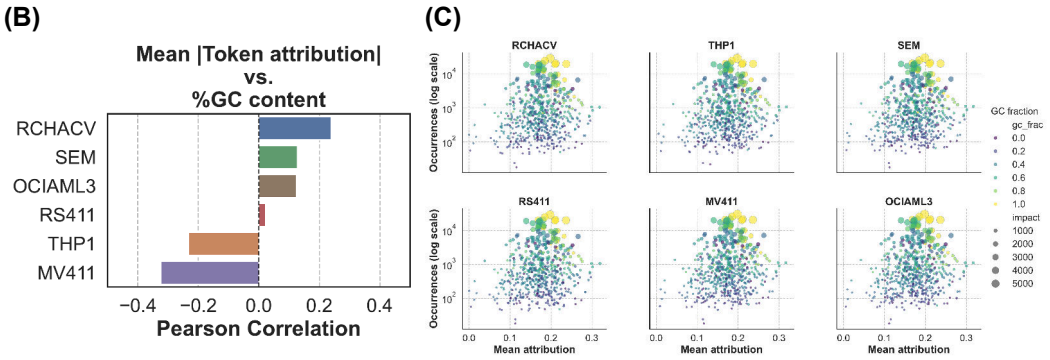


Figure 4.13: (Previous page) Token-level attributions were computed with Layer Integrated Gradients (50 steps, paired baselines) at the embedding layer of fine-tuned GROVER models, using only true-positive sequences. **(A)** Heatmap of mean absolute attribution scores (Z-scored per token), showing distinct clustering of tokens between RCH-ACV MLL and fusion-positive cell lines. The left bar indicates token %GC content. **(B)** Pearson correlation between token importance and %GC per cell line. RCH-ACV (MLL) showed the strongest positive correlation, consistent with CpG-island binding, whereas fusion lines showed weaker or negative correlations. **(C)** Bubble plots of token frequency vs mean attribution, coloured by GC fraction and sized by impact score, illustrating that while CpG-rich tokens were particularly emphasised in MLL, they also contributed across fusion contexts.

Sequence-level attributions and motif discovery

To further interpret the sequence features underlying model predictions, I extended the analysis from token-level attributions to base-resolution maps suitable for motif discovery. These sequence level attributions were clustered into seqlets with TF-MoDISco-lite. Using tomtom, motifs were annotated against HOCOMOCO v12, and replicate results were collapsed to generate cell-type consensus profiles and presented as a bubble plot (Figure 4.14).

For positive attributions, motifs of the C2H2 zinc finger class dominated across all cell lines. In particular, the VEZF1 motif (enriched for G/C nucleotides) was consistently recovered in positively attributed seqlets in every cell type. The (MLL) motif itself was also identified across all lines. Notably, seqlets matching the MLL motif contributed both positive and negative attributions, indicating that the model sometimes leveraged this feature to support binding predictions, while in other contexts the same motif was treated as evidence against binding.

A striking contrast was observed between positive and negative attribution patterns. Whereas positive attributions converged on a relatively consistent set of CpG-associated motifs across cell types, negative attributions were markedly more diverse. Each cell line displayed enrichment for a distinct set of negatively weighted motifs spanning multiple classes, including bHLH (e.g. LYL1, USF2, MYC), RUNX,

and homeodomain factors.

Together, these results suggest that what distinguishes MLL predictions between cell lines is not the positive sequence features, as the CpG-associated motifs are largely shared, but which alternative motifs are actively suppressed. In contrast, each cell line shows negative weighting of a distinct set of non-CpG motifs spanning bHLH factors (LYL1, USF2, MYC), RUNX family motifs, and homeodomain factors. These motifs are enriched at negatively attributed seqlets in a cell-line-specific manner, suggesting their presence is associated with reduced MLL binding predictions in each context. The model thus achieves cell- and fusion-context specific predictions not by learning unique positive sequence codes, but by down-weighting the alternative motifs that are active in each lineage.

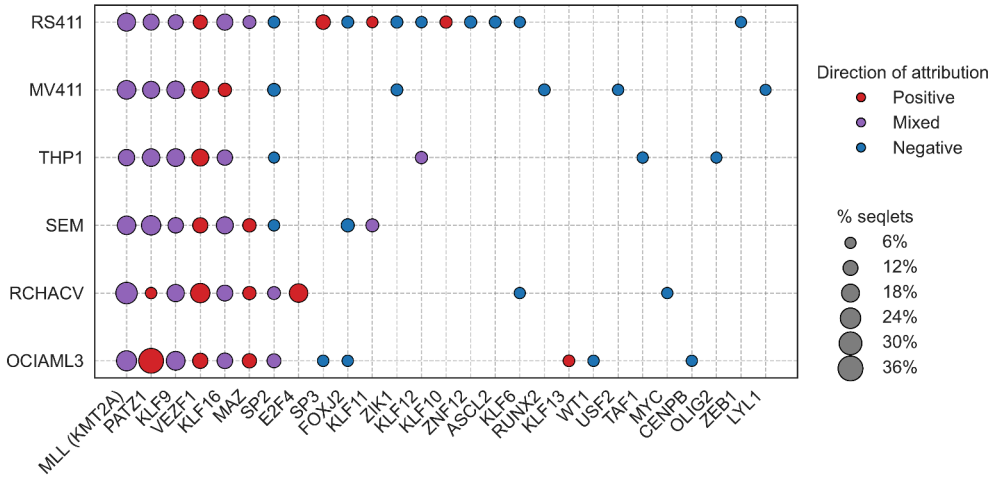


Figure 4.14: Motif enrichment from nucleotide-level attributions across cell types. Seqlets derived from base-resolution attributions of fine-tuned GROVER models were clustered with TF-MoDISco-lite and matched to HOCOMOCO v12 motifs including the previously identified MLL (KMT2A) motif from HOCOMOCO v12. Replicates were collapsed to generate cell-type consensus profiles. Bubble size indicates the proportion of seqlets assigned to each motif (% of total), and colour denotes attribution direction as positive (red), mixed (purple), and negative (blue) groups.

4.3.5 Sequence composition

Promoter sequences exhibited higher GC content (median ∼60% vs ∼45% in methylome windows) and strong enrichment of many transcription factor families, consistent with a dense regulatory grammar centred on CpG islands (Figure 4.15C). This pattern mirrors the enrichment of CXXC zinc fingers and Krüppel-like factors observed at unmethylated CpG-associated MLL peaks, reinforcing the view that MLL binding is tightly linked to CpG-rich promoters. In contrast, methylome windows are compositionally more heterogeneous and show lower per-kb motif en-

richment, reducing signal-to-noise and making the task more challenging for predictive models. Motif enrichment in promoter sequences (relative to methylome) was dominated by GC-box binders (SP/KLF family, MAZ, PATZ1), CpG-island-associated CXXC proteins (KMT2A/KMT2B), E2F/TFDP, NRF1, and basic helix-loop-helix or basic leucine zipper transcription factors (USF1/2, CREM, JUN/AP-1) (Figure 4.15A, 4.15B). Effect sizes reached log2 fold enrichment ≈ 3.3 (median ≈ 1.22), with adjusted P-values at the multiple-testing floor. By contrast, enrichments in the methylome set were weaker and more diffuse across families such as bHLH/bZIP, HMG-domain, homeobox, STAT/Rel-homology region, and nuclear receptors, consistent with lower motif density outside CpG islands.

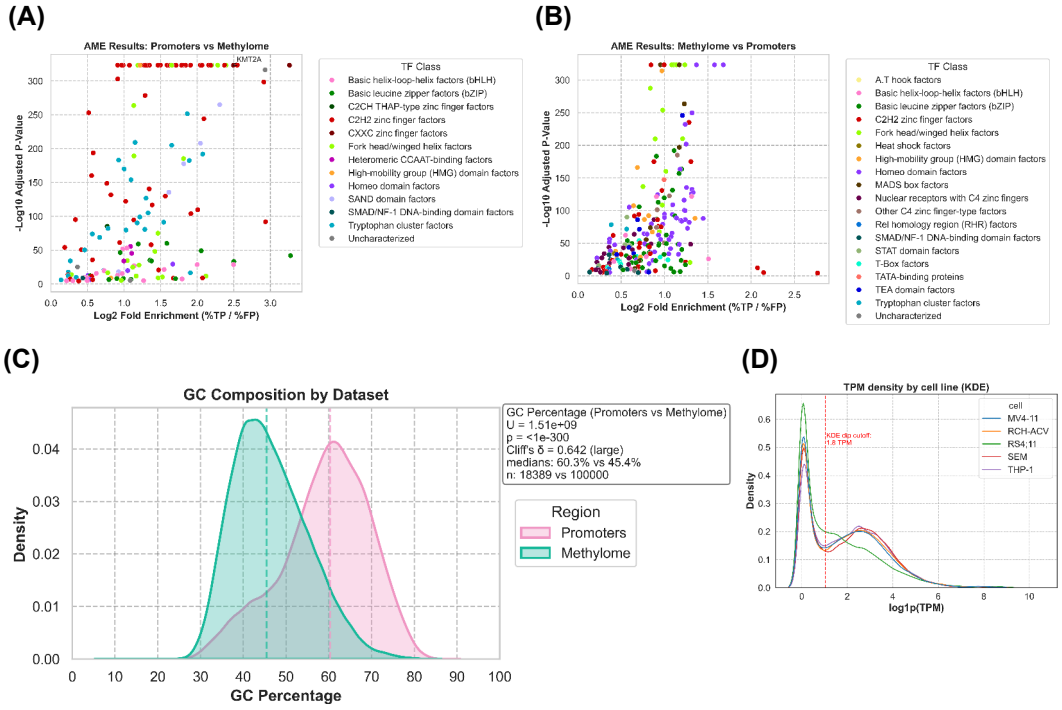


Figure 4.15: **(A)** Motif enrichment in promoter regions relative to methylome, grouped by transcription factor family. Enrichment is dominated by GC-box binders (SP/KLF family, MAZ, PATZ1), CpG-island-associated CXXC proteins (KMT2A/KMT2B), and additional promoter-associated factors. **(B)** Motif enrichment in methylome regions relative to promoters, showing broader and weaker enrichments across diverse TF families (bHLH/bZIP, HMG, homeobox, STAT/RHR, nuclear receptors). **(C)** GC content distributions for methylome versus promoter regions (%). Promoter sequences are markedly GC-rich compared with the more heterogeneous methylome windows. **(D)** Filtering of enriched motifs by expression thresholds. Only motifs corresponding to transcription factors expressed above a defined TPM cutoff in the relevant cell line panel were retained for downstream analyses, ensuring that reported enrichments reflect factors with the potential to be active in the experimental system.

4.3.6 Motif enrichment

To assess whether MLL-N binding differed according to CpG context, motif enrichment was performed separately for MLL-N peaks which were unmethylated (5mC data from Chapter 6) (MLL-N ∩ uCpG) compared with all MLL-N peaks to determine the motifs from HOCOMOCO v12 enriched in unmethylated MLL-N peaks relative to all MLL-N peaks. Separately, motifs enriched in all MLL-N peaks relative to those overlapping unmethylated CpG MLL-N were also identified. This was carried out in two cell lines, SEM with MLL-AF4 and RCH-ACV (Figure 4.16). In RCH-ACV, unmethylated MLL-N peaks were dominated by CXXC zinc finger motifs (including MLL itself), Krüppel-like factors (KLF family), and forkhead/winged-helix factors, consistent with direct recognition of CpG-rich regions. By contrast, peaks with methylation were enriched for a broader spectrum of motifs, including nuclear factor I C (NFIC), hypermethylated in cancer 1 (HIC1), LYL1 (basic helix-loop-helix family), and basic leucine zipper (bZIP) transcription factors, which do not represent canonical MLL binding determinants. SEM cells showed a similar split: CXXC zinc fingers and KLF motifs enriched at unmethylated MLL-N peaks, but comparatively weaker than in RCH-ACV, alongside stronger enrichment of diverse zinc finger proteins and basic helix-loop-helix motifs at methylated MLL-N peaks. These results suggest that while MLL binding is tightly linked to CpG-island recognition, MLL-AF4 binding extends into CpG-poor regions where non-canonical transcription factor motifs may provide alternative recruitment pathways.

While these enrichment analyses highlight clear differences in motif usage between unmethylated and methyalted MLL binding, they do not establish whether such sequence features are sufficient to predict binding genome-wide, nor how they differ between MLL and fusion contexts.

The enrichment of CXXC zinc finger motifs (KMT2A) and CpG-associated C2H2 factors (KLFs, SP family, MAZ, PATZ1, VEZF1) at unmethylated MLL-N peaks is consistent with direct recognition of CpG-rich regions by the CXXC domain of MLL. In contrast, the broader range of motifs enriched at methylated MLL-N peaks (NFIC, HIC1, LYL1, bZIP factors) suggests that alternative recruitment mechanisms may operate outside of unmethylated GC-rich contexts. This occurs in both RCH-ACV (MLL) and SEM (MLL-AF4) cells, but the smaller effect size observed in SEM for motifs such as KLFs and SP family members indicates a reduced reliance on Unmethylated Cytosine-Phosphate-Guanine (uCpG) recognition in the fusion context.

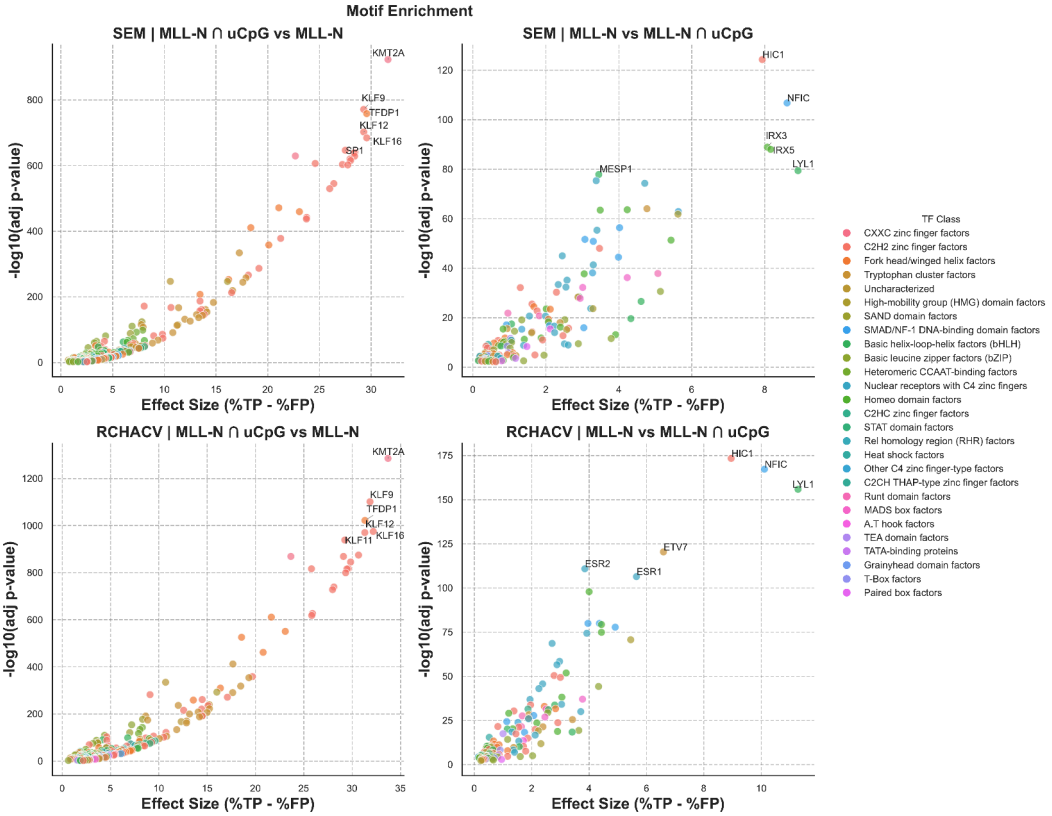


Figure 4.16: Motif enrichment was calculated using AME (MEME Suite) for MLL-N peaks overlapping unmethylated regions (MLL-N ∩ uCpG) compared with all MLL-N peaks, and vice versa. Scatter plots show effect size (% true positives − % false positives) on the x-axis and statistical significance (-log10 adjusted p-value) on the y-axis. Each point corresponds to a motif, coloured by transcription factor class. Upper left: RCH-ACV: motifs enriched in all MLL-N peaks relative to unmethylated MLL-N peaks. Upper right: RCH-ACV: motifs enriched in MLL-N vs. uCpG relative to all MLL-N peaks. Lower left: SEM: motifs enriched in all MLL-N peaks relative to MLL-N vs. uCpG. Lower right: SEM: motifs enriched in MLL-N vs. uCpG relative to all MLL-N peaks.

4.4 Discussion

4.4.1 MLL binding is shaped by CpG-island context

Motif enrichment analyses confirmed that MLL (RCH-ACV) peaks overlapping unmethylated CpG islands were dominated by canonical CXXC zinc fingers (KMT2A) and CpG-associated C2H2 factors (KLFs, SP family, MAZ, PATZ1, VEZF1). Outside CpG islands, a broader range of motifs was enriched (NFIC, HIC1, bHLH family), indicating potential alternative recruitment mechanisms. SEM (MLL-AF4) showed weaker enrichment of CpG determinants and stronger representation of diverse zinc fingers outside uCpGs, consistent with more dispersed binding. These results highlighted CpG-island recognition as a shared determinant of MLL binding, with non-canonical motifs distinguishing fusion contexts.

4.4.2 Dataset composition influences predictive challenge

Promoter datasets were GC-rich (≈ 60%) and motif dense, while methylome datasets were compositionally heterogeneous (≈ 45% GC) with weaker motif enrichment. Promoter enrichments recapitulated CpG-associated motifs, whereas methylome windows were more diffuse. This compositional contrast framed promoters as a relatively simple prediction task, and methylome windows as a more challenging, heterogeneous task.

4.4.3 DNA language models recover MLL binding with high accuracy

All models performed well on promoters, but GROVER achieved the highest robustness on the more complex methylome task (ROC AUC ≈ 0.86) and generalisation to tiled chr9 windows (AUC ≈ 0.95). MLL fusion cell lines (SEM, RS4;11, MV4-11) remained predictable but showed subtle differences in per-label F1 scores, reflecting distinct sequence determinants. Embedding projections showed that CUT&Tag signal intensity and promoter regions aligned with predicted binding probabilities aligning.

4.4.4 Token-level attributions reveal GC-driven features

Layer Integrated Gradients (LIG) highlighted CpG/GC-rich tokens as the most impactful across all lines. RCH-ACV (MLL) showed the strongest positive correlation

between token attribution and GC content, consistent with direct CXXC -mediated CpG recognition. MLL fusion cell lines showed weaker (SEM, RS4;11) or even negative correlations (THP1, MV4-11), suggesting reduced reliance on CpG content. When weighted by frequency, GC-rich tokens remained impactful across all contexts.

4.4.5 Seqlet-level motifs highlight conserved CpG factors and diverse suppressors

Motif clustering of sequence-level attributions (TF-MoDISco + Tomtom) revealed consistent recovery of CpG-associated motifs (KLF9/10/11/12/13/16, SP2/3, MAZ, PATZ1, VEZF1) across all cell types, confirming that the models captured canonical CpG-driven determinants. The MLL motif itself was recovered across lines with both positive and negative polarity, suggesting that the same sequence element could either support or oppose binding predictions which could be due to the motif itself being defined with just ChIP-seq data and rated a quality score of just B by Hocomoco v12. In contrast, negative attribution motifs were highly diverse and cell-line specific, spanning bHLH (LYL1, USF2), RUNX, homeodomain, and other families. This indicates that cell-line specificity arises less from the presence of canonical CpG determinants, and more from the selective suppression of alternative motifs, which down-weight competing features and sharpen lineage-specific predictions.

4.4.6 Limitations and future directions

While this study provides a comprehensive benchmark of DNA language models for MLL-N binding prediction, several limitations should be acknowledged.

The training datasets were derived from a limited number of cell lines, which may not capture the full diversity of MLL-N binding contexts. Future work could expand to additional cell types and primary samples and also include sample specific genome sequences to capture full sequence variability. The models were trained on 1,024 bp windows, which may not fully capture long-range regulatory interactions influencing MLL-N binding.

While the model was able predict binding accurately for these cell lines, the interpretation of the features learned by the model is limited to sequence features and future work could investigate the positional effects of these features. In-silico mutagenesis or other perturbation-based attribution methods could complement the

gradient-based approaches used here, providing orthogonal validation of key sequence features.

Finally, experimental validation of predicted binding sites and motifs would strengthen the biological relevance of the findings. This could include targeted mutagenesis of predicted motifs followed by ChIP-qPCR to confirm their functional role in MLL-N recruitment.

4.4.7 Conclusions

Although all tested DNA language models performed well on the relatively simple promoter prediction task, GROVER which was trained exclusively on human DNA sequences demonstrated superior robustness and accuracy on the more complex methylome task and better generalisation to tiled chr9 regions. Layer Integrated Gradients highlighted GC-rich tokens as the most impactful features, particularly in MLL contexts. Motif clustering of sequence-level attributions revealed consistent recovery of CpG-associated motifs across all cell types, while negative attribution motifs were highly diverse and cell-line specific. These findings suggest that MLL-N binding is shaped by a conserved core of CpG-driven determinants, modulated by the selective suppression of alternative motifs to achieve cell-line specificity.

However, the model did not perform well in predicting MLL-N binding specifically in the context of MLL-AF4 fusion proteins at sites where they spread into gene bodies. This indicates that additional factors such as interactions with other proteins, histone marks, and DNA methylation may influence binding in these contexts which will be explored in subsequent chapters.

5 MLL binding in the context of cooperating factors

5.1 Introduction

5.1.1 Co-factor binding and chromatin context

Precise regulation of gene expression relies on the coordinated activity of many regulatory proteins. While TFs bind specific DNA motifs to determine recruiting transcription machinery, they rarely act alone. Instead, they also recruit cofactors such as chromatin remodellers, histone-modifying enzymes, and large multi-protein complexes that collectively shape the chromatin environment and fine-tune transcriptional output (Spitz, F. 2012; Lambert, S. A. et al. 2018).

These interactions are dynamic and highly context-dependent. Pioneer TFs can open compacted chromatin, remodellers such as SWI/SNF (SWItch/Sucrose Non-Fermentable) family reposition nucleosomes to expose regulatory DNA (Clapier, C. R. and Cairns, B. R. 2009), and histone acetyltransferases like p300 and CBP deposit activating marks such as H3K27ac to reinforce accessibility and gene activation (Tie, F. et al. 2009). Feedback loops then stabilise either active or repressive states, creating a responsive and adaptable regulatory landscape (Zaret, K. S. and Carroll, J. S. 2011).

At the genome-wide level, combinations of TFs and cofactors form regulatory modules. Clusters of these proteins converge at promoters and enhancers to drive the expression of target genes. High-throughput profiling such as ChIP-seq and CUT&Tag reveal dense co-binding at shared loci, forming enhancer clusters that expand through cooperativity and act as hubs for transcriptional control (Kaya-Okur, H. S. et al. 2019; Blayney, J. W. et al. 2023). These structures are essential for maintaining cell identity and are frequently hijacked in cancer to drive abnormal gene expression programmes such as the formation of so-called superenhancers

(Hnisz, D. et al. 2013).

Beyond static co-binding, recent single-molecule and live-cell work has shown that transcription factor cooperativity is dynamic, with clusters of transient, overlapping interactions modulating occupancy and transcriptional output over time. Studies from the Lenstra group exemplify this, revealing how binding kinetics and local crowding shape cooperative behaviour at promoters and enhancers (Pomp, W. et al. 2024). This dynamic view complements ChIP-seq and CUT&Tag maps by explaining how apparent co-binding can emerge from time-averaged interactions rather than fixed assemblies.

Recent techniques such as PADIT-seq have expanded this view by mapping not only high-affinity TF binding sites but also networks of overlapping lower-affinity motifs surrounding them (Khetan, S. et al. 2025). These flanking sites can modulate competition between related factors and increase overall occupancy, suggesting that the apparent co-binding detected by ChIP-seq may reflect a continuum of binding affinities rather than discrete, isolated sites.

Understanding how these complex interactions form and change is critical for deciphering transcriptional control and how it becomes dysregulated in disease. This is especially relevant in leukaemias driven by MLL rearrangements, where recruitment of specific cofactors reprograms chromatin and misdirects the transcriptional machinery toward leukaemogenic targets. Although some of the general principles of how specific combinations of transcription factors interact to regulate binding and transcriptional output have begun to be revealed (Pomp, W. et al. 2024; Khetan, S. et al. 2025), how chromatin-associated proteins fit into these rules remains less well understood, despite their central role in shaping the regulatory landscape. MLLr leukaemias represent a powerful model for addressing this gap, as they exemplify an aggressive disease driven by the interplay between TFs and chromatin proteins, where aberrant cofactor recruitment fundamentally reprograms gene regulation.

5.1.2 Histone marks as the regulatory landscape

Histone modifications act both as a readout of transcriptional activity and as signals for factor recruitment. Characteristic combinations distinguish promoters, enhancers, and actively transcribed regions, integrating with TF binding to shape regulatory outcomes. At promoters, H3K4me3 deposited by MLL marks transcription start sites and helps maintain an open chromatin state (Bernstein, B. E. et al. 2005), whereas H3K27ac catalysed by p300/CBP highlights active promoters and enhancers (Tie, F. et al. 2009). In parallel, DOT1L-mediated H3K79me2 tracks

transcriptional elongation and sustained enhancer activity (Godfrey, L. et al. 2019). These modifications operate combinatorially H3K27ac broadly marks active regulatory elements including both promoters and distal enhancers, while H3K4me3 specifically distinguishes active promoters, together producing distinct recruitment landscapes that bias which cofactors are engaged. Mapping these signatures is therefore essential for understanding where and how proteins like MLL are positioned across the genome.

5.1.3 MLL biology and transcriptional dysregulation

The Mixed Lineage leukaemia 1 (MLL, KMT2A) protein is a histone methyltransferase critical for developmental gene regulation and haematopoiesis. Under normal conditions, MLL deposits H3K4me3 at transcription start sites to maintain an open chromatin state and support lineage-specific gene programmes (Ruthenburg, A. J. et al. 2011; Shilatifard, A. 2012).

MLL is recruited to CpG-rich promoters via its CXXC domain and stabilised through interactions with Menin, LEDGF, and the PAF1 complex (See Chapter 1, Section 1.1.1), motivating their inclusion as predictive features in this chapter (Milne, T. A. et al. 2010; Yokoyama, A. and Cleary, M. L. 2008).

In MLLr leukaemias, chromosomal translocations fuse the N-terminal portion of MLL (containing DNA-binding and Menin/LEDGF interaction domains) to diverse partner proteins (Meyer, C. et al. 2023). This replaces the normal C-terminal SET domain, abolishing MLL's intrinsic methyltransferase activity and converting it into a potent recruiter of elongation machinery (Krivtsov, A. V. and Armstrong, S. A. 2007). Through its fusion partners, MLL engages elongation factors and coactivator complexes, misdirecting transcription to leukaemogenic loci.

In MLLr leukaemia, both MLL and MLL-FP can co-occupy target loci. Tagged-allele ChIP-qPCR experiments show that MLL remains bound and in some cases increases at promoters even when fusion proteins are present, rather than being displaced (Milne, T. A. et al. 2005b). Knockdown studies further demonstrate that reducing MLL-AF4 levels leads to a partial loss of binding, with residual promoter-proximal MLL-N signal persisting from the protein (Kerry, J. et al. 2017; Smith, A. et al. 2025a). This co-occupancy means that, in fusion-positive cell lines, promoter-linked signals from MLL are layered on top of enhancer and elongation associated signals driven by the fusion complex, which must be considered when interpreting correlation structures and model attributions. Tagged-allele ChIP-seq for the fusion protein could have improved labelling specificity for MLLr ML models, as MLL-N

conflates fusion with WT MLL signal; however, such data were unavailable in the cell lines used here.

5.1.4 Previous Machine Learning Approaches to Predicting TF Binding

The complexity of transcriptional regulation has motivated the development of computational algorithms to predict TF binding directly from genomic and epigenomic data. Early approaches relied on simple motif scanning, in which position weight matrices (PWMs) were used to identify potential TF recognition sites within the genome (Stormo, G. D. 2000). While useful for mapping consensus motifs, these methods are limited by their inability to account for chromatin context or cooperative binding effects, and they tend to produce large numbers of false positives when applied to genomic sequences alone given that not all motifs are bound by their corresponding protein and this is a dynamic process (Wasserman, W. W. and Sandelin, A. 2004).

With the advent of higher throughput sequencing techniques and resulting availability of data, more complex probabilistic approaches were developed to improve TF binding prediction. Models such as CENTIPEDE used DNase-seq footprinting to infer TF occupancy, exploiting the characteristic depletion of cleavage signal directly beneath a bound protein, combined with DNA sequence motif scores to classify sites as bound or unbound (Pique-Regi, R. et al. 2011).

More recently, deep learning methods have revolutionised this field by enabling the direct modelling of complex, non-linear interactions between sequence and chromatin features. Convolutional neural networks (CNNs) such as DeepBind and DeepSEA were among the first to demonstrate that models trained on raw genomic sequence could learn motifs de novo and capture higher-order dependencies (Alipanahi, B. et al. 2015; Zhou, J. and Troyanskaya, O. G. 2015). Building on this, BPNet introduced a base-resolution framework for predicting strand-specific TF binding profiles from DNA sequence alone, allowing the model to learn motif syntax and cooperative interactions between factors (Avsec, Ž. et al. 2019). More recently, transformer-based architectures such as Enformer have shown that cooperative binding patterns and long-range regulatory interactions can be learned directly from sequence, providing deeper insights into the combinatorial rules that govern TF recruitment (Avsec, Ž. et al. 2019). These deep learning approaches have also proven valuable for predicting how non-coding genetic variants alter TF binding and regulatory activity.

Despite these advances, relatively few studies have explicitly modelled the binding of a specific factor from the patterns of co-binding TFs and chromatin-associated proteins. Most approaches treat TFs independently or predict aggregate occupancy. Biologically, however, TFs act within multi-protein complexes whose recruitment depends on specific combinations of interacting factors. Recent sequence or accessibility based models predict individual TF binding with high accuracy, but their interpretability is often limited to motif extraction and does not resolve factor-factor interactions. These considerations motivate the approach taken in this chapter. By modelling MLL binding as a function of co-binding proteins and histone modifications, I aim to move beyond sequence based predictors and directly capture the cooperative interactions that specify MLL recruitment. This strategy enables accurate prediction of MLL occupancy while yielding biologically interpretable features that illuminate the mechanisms underlying leukaemic transcriptional programmes.

5.1.5 Chapter Aims

Building on evidence that transcriptional regulation emerges from cooperative interactions among sequence-specific factors, chromatin state, and elongation machinery, this chapter tests whether MLL recruitment can be predicted from the local constellation of co-binding proteins and histone marks across distinct genomic contexts. With this goal in mind I pose the following questions.

1. Do local co-factor and histone patterns predict MLL-N binding at CpG-island promoters and diverse methylome regions and do the rules differ between them?

2. Which cofactors and chromatin features are the strongest drivers, and do their contributions align with established MLL biology?

3. Do the model explanations resolve coherent, mechanistic co-factor interactions that clarify how leukaemic programmes are enforced at promoters and distal elements?

5.2 Methods

5.2.1 Datasets and Preprocessing

CUT&Tag and ChIP-seq data were processed using SeqNado as described in chapter 3. The bigwigs scaled using RPKM and binsize 1 were imported over either the promoter or methylome regions as per chapter 4, as mean RPKM over the regions.

Normalisation was applied per track in four light steps. First, values were scaled by the non-zero 75th percentile to put typical signal on a common scale. Next, extreme highs were capped at the 99.5th percentile to limit outliers. A log1p transform was then applied to compress remaining skew. Finally, values were min-maxed scaled to 0–1 for regression modelling. The data were split by holding out chr9 regions for test, chr8 for evaluation and the remaining chromosomes used as the training set (Table 5.1).

Dataset	Training	Validation (chr8)	Test (chr9)
Methylome	472,377	18,539	18,539
Promoter	18,591	704	777

Table 5.1: Summary of dataset sizes and chromosome-based splits for training, evaluation, and testing.

5.2.2 Modelling Approaches

GANDALF

GANDALF is a neural architecture designed for tabular data that uses Gated Feature Learning Units (GFLUs) to perform adaptive feature selection and non-linear interaction modelling across inputs (Joseph, M. and Raj, H. 2022). Each GFLU applies a stage-specific learnable mask for soft feature selection and a gating mechanism to progressively refine the learned representation. Its shared backbone supports joint multi-output regression under a unified optimisation objective, making it well suited to settings where targets are expected to be correlated. This is appropriate here as all four MLL-N targets (two replicates per cell line) are predicted jointly, chromatin co-factor interactions are expected to be complex and non-linear, and the number of co-factors differs between cell lines. The shared backbone additionally stabilises SHAP-based feature importance estimates, by pooling signal

across all four targets during training, the model identifies co-factors that are consistently informative, yielding more robust importance rankings than independent per-target models would provide. Ten GFLU stages were stacked, followed by a linear prediction head. Feature masks were initialised with low sparsity (initial feature sparsity = 0.05) and refined during training, allowing the model to adaptively select informative features across the 87 pooled co-factor inputs. Regularisation was applied through embedding and GFLU dropout (each 0.01). Outputs were produced by a linear head with one unit per target, and scaling to [0,1] was enforced via the library's target range setting for every label.

Hyperparameter sweep

A Bayesian hyperparameter sweep was conducted in Weights & Biases to maximise the validation Coefficient of Determination (R^2), representing the proportion of variance explained by the model. Ten trials were executed per sweep invocation. Search spaces were defined as follows: learning rate sampled log-uniformly from 1×10^{-4} to 1×10^{-3}; weight decay sampled log-uniformly from 5×10^{-4} to 5×10^{-3}; gradient-norm clipping sampled uniformly from 0.5 to 3.0; embedding dropout sampled uniformly from 0.00 to 0.02; GFLU dropout sampled uniformly from 0.00 to 0.05; initial feature sparsity sampled uniformly from 0.03 to 0.08; and the number of GFLU stages chosen from at even integers between 10 and 20 inclusive. For each trial, a GANDALF model with a linear regression head was instantiated with target range between 0 and 1 for all outputs.

Training during the sweep was performed with a batch size of 2048, a maximum of 12 epochs, and early stopping on validation R^2 with a patience of 3 epochs. The default Adam optimiser provided by the framework was used, global gradient-norm clipping was applied to increase training stability.

Model selection and final training

After the sweep, a configuration was selected and trained to convergence with the following settings: learning rate 0.02, weight decay 0.001, global gradient-norm clip 1.0, batch size 2048, maximum 50 epochs, and early stopping on validation R^2 with a patience of 5 epochs, restoring the best checkpoint. Computation was executed on Apple Metal Performance Shaders (MPS). Randomness was controlled with a fixed seed (42).

Evaluation and logging

During training, validation R^2 and Mean Squared Error (MSE) were recorded each epoch and configurations and training logs were captured by Weights & Biases. After training, held-out test performance was assessed by predictions generated on the test set and compared with ground truth to compute per-target R^2 and MSE, with macro R^2 and macro MSE obtained by uniform averaging across targets.

XGBoost

To benchmark GANDALF, a gradient-boosting baseline was trained using XGBoost in a multi-output configuration, with an XGBRegressor wrapped by MultiOutputRegressor to produce one regressor per target. Features and targets were separated from the pre-processed dataset. All regions on chromosome 9 were held out as a test set, and the remaining chromosomes were used for training and model selection. Within the training pool, rows were shuffled with a fixed seed (42) to ensure reproducibility. Chromosome labels derived from genomic coordinates were used to construct non-overlapping folds via GroupKFold (three folds), ensuring that entire chromosomes were contained within a single fold and preventing leakage.

Hyperparameters were tuned with a randomised search over 20 sampled configurations. The search space covered tree depth, learning rate, number of estimators, subsampling, column subsampling, minimum child weight, L2 and L1 regularisation, and the histogram-based tree method. Evaluation during the search used uniform-average macro R^2 across outputs as the scoring metric. Parallel execution was enabled, and a fixed random seed (42) was applied. The best configuration identified by cross-validation was extracted and a final multi-output model was refit on the full training set (all non-chr9 regions).

Performance was assessed on the held-out chr9 set. Predictions were generated for each target, and macro R^2 and macro mean squared error were computed as uniform averages across outputs, alongside per-target R^2 and MSE.

5.2.3 Feature Extraction

With the model in evaluation mode, a lightweight prediction wrapper was defined to accept NumPy arrays, route it through the GANDALF's continuous features input, and return the raw logits without additional activation. The SHAP background distribution was approximated from the training features by k-means clustering (k = 50)

the resulting centroids were used as the background dataset to represent typical regions of the feature space while keeping the explainer tractable. Kernel SHAP was then applied with the custom predictor function and the k-means background, with explicit feature and output names supplied to preserve column alignment. Because the model is multi-target, the explainer produced one attribution matrix per output. For each output, per-feature global importance was computed as the mean absolute attribution across the sampled test instances. All attributions were computed in the model's input space, i.e., on the pre-processed continuous features exactly as seen by the network during training and inference. This procedure relies on an approximate independence assumption under the chosen background distribution and is computationally demanding. Accordingly, a reduced test subset and a k-means background were used to keep the analysis tractable. To manage runtime, up to 1,000 rows were randomly sampled from the test split using a fixed seed (42) to form the evaluation subset.

5.2.4 Software versions and environments

All analyses were performed in Python 3.13.7. See Appendix for software and package versions (Table B.5).

5.3 Results

5.3.1 MLL Correlates with Distinct Co-factor Networks Across Genomic Contexts

MLL Correlation in RCH-ACV Cells

RCH-ACV cells are a B-ALL line carrying an E2A-PBX1 fusion but no MLL rearrangement (Jack, I. et al. 1986), and thus serve as a model of MLL binding in leukaemia, providing a contrast to SEM cells, which harbour MLL-AF4. I generated a panel of co-factor CUT&Tag datasets in RCH-ACV to complement MLL-N and H3K27ac CUT&Tag prepared by Alastair Smith (Table A.3), with additional ChIP-seq tracks from the Milne lab archive (Table A.4). Pairwise Pearson correlations were computed across CpG-island-rich promoter windows ($\pm$512 bp around annotated TSSs) and Twist methylome panel regions ($\pm$512 bp re-centred to midpoints; see Section 4.2.1 for panel description). The scaled CUT&Tag and ChIP-seq datasets were combined into a single matrix and visualised as clustered correlation heatmaps (Figure 5.1).

At promoters, MLL-N clustered tightly with Menin, the SWI/SNF subunit BRG1, and transcription associated factors such as RNA polymerase II (Figure 5.1A). These strong correlations reflect MLL's established role in promoter tethering through Menin and in maintaining transcription initiation at active genes. Histone marks characteristic of promoters, including H3K4me3 and H3K27ac, were also highly correlated with MLL-N, consistent with MLL occupancy occurring within active promoter-associated chromatin signatures.

In contrast, the methylome regions exhibited a more heterogeneous correlation structure (Figure 5.1B). While MLL-N continued to correlate with Menin and several transcription associated factors, additional associations became prominent. H3K27ac remained strongly correlated, reflecting its role in active regions, but distal regions also showed higher correlations with elongation related features, including DOT1L, as well as elongation associated factors such as BRG1 and BRD4 which is part of the Transcription Elongation Factor b (TEFb) complex. In the methylome regions we also see correlation with E2A, PBX1 and RUNX1 which in RCH-ACV where E2A-PBX1 is the driver fusion, known to activate RUNX1 indicating an association between MLL and active regions outside of the promoters (Pi, W.-C. et al. 2020). These patterns suggest that, outside promoters, MLL recruitment reflects

a balance between promoter tethering and cooperative interactions with enhancer activity and elongation machinery.

Overall, these correlation patterns indicate that MLL participates in distinct, context-dependent co-factor networks. At promoters, MLL is embedded within a transcription initiation module defined by Menin, and H3K4me3, whereas at distal elements captured by the methylome assay, MLL recruitment appears to be shaped by a broader array of factors linked to active regions and transcriptional elongation. These observations provided the motivation for subsequent modelling analyses aimed at systematically predicting MLL occupancy and identifying the most influential co-factors in each genomic context.

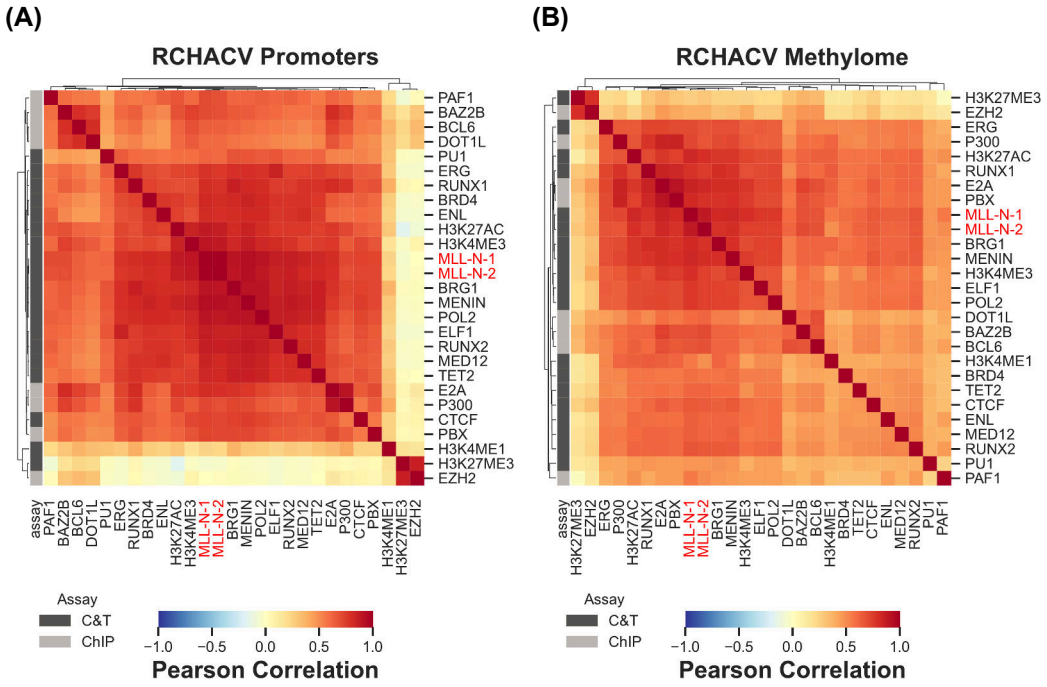


Figure 5.1: MLL-N correlations with co-factors differ between promoters and methylome regions in RCH-ACV cells. **(A)** Pearson correlation heatmap of MLL-N, co-binding proteins, and histone modifications across CpG island-rich promoter regions. MLL-N replicates (red) cluster closely with Menin, BRG1, RNA polymerase II, and active promoter-associated histone marks H3K4me3 and H3K27ac. **(B)** Equivalent correlation heatmap across methylome-captured regions, representing a more heterogeneous genomic landscape. While MLL-N remains correlated with Menin and transcriptional cofactors, additional associations emerge with RCH-ACV fusion protein E2A::PBX1 and RUNX1, and H3K27ac-marked active regions. Together, these patterns suggest context-specific MLL recruitment, with promoter-centric interactions shifting toward enhancer and elongation networks at distal elements.

MLL Correlation in SEM Cells

To compare MLL and MLL-fusion contexts, I next examined MLL-N relationships in SEM cells. Correlations were computed using the same region sets and preprocessing described above for a wider range of co-factors as more data was available (promoters and Twist methylome panel) and visualised as hierarchically clustered heatmaps (Figure 5.2).

At promoters, the correlation structure was more diffuse than in RCH-ACV, indicating a broader and more heterogeneous network of associations (Figure 5.2A). MLL-N remained positively correlated with active promoter-associated cofactors such as , H3K27ac, RNA polymerase II, GCN5 which is a HAT and part of the SAGA complex. Distinct blocks emerged, including a repressive module defined by EZH2, H3K27me3 and surprisingly LEDGF, which showed weaker or negative correlation with MLL-N. Although LEDGF is a known component of the MLL tethering complex (Yokoyama, A. and Cleary, M. L. 2008; El Ashkar, S. et al. 2017), the LEDGF ChIP-seq contains substantial background signal (as do AF9, CDK6 and SATB1). Nevertheless they were retained for the ML tasks to avoid omitted-variable bias and to allow the model to explicitly down-weight uninformative features. Under strong regularisation (sparse gating, dropout, weight decay) and chromosome-aware validation, their inclusion did not degrade performance; instead, it enabled the model to absorb any residual signal while correctly assigning low importance where appropriate. In addition, the scaling pipeline assumes a more zero-biased distribution; combined with higher background, this compressed LEDGF variance and diminished pairwise correlations. Overall, this separation suggests that SEM promoter regions comprise both actively transcribed targets bound by MLL and regions dominated by repressive chromatin states.

Across the methylome-captured regions, MLL-N displayed a more distinct pattern of associations with a more select range of cofactors and histone marks (Figure 5.2B). Strong positive correlations with Mediator subunits MED1 and MED12, the elongation factor BRD4, and nucleome remodellers such as the FACT subunit SSRP1 and TET2 which modifies DNA methylation were observed. These results indicate that, in the wider genomic landscape, MLL-N binding reflects not only promoter tethering but also enhancer-driven activity and transcriptional elongation.

Comparing datasets and scaling

The RCH-ACV and SEM datasets differed in both scale and feature diversity, reflecting differences in data availability. The RCH-ACV analysis incorporated 25

co-factors and histone marks from CUT&Tag that I produced for this chapter (Table A.3), while the SEM dataset included 62 features, representing a substantially broader view of the chromatin landscape. Despite this increase in the number of cofactors and histone marks, the scaling procedure preserved the global correlation structure and maintained the integrity of technical replicates, with the two MLL-N replicates clustering tightly together in both datasets. This indicates that the normalisation process faithfully captured biological relationships even with a much larger feature set. The broader coverage and robust scaling in the SEM data provided a particularly rich foundation for downstream modelling, enabling us to test whether the determinants of MLL binding were consistent across fusion contexts.

(A)

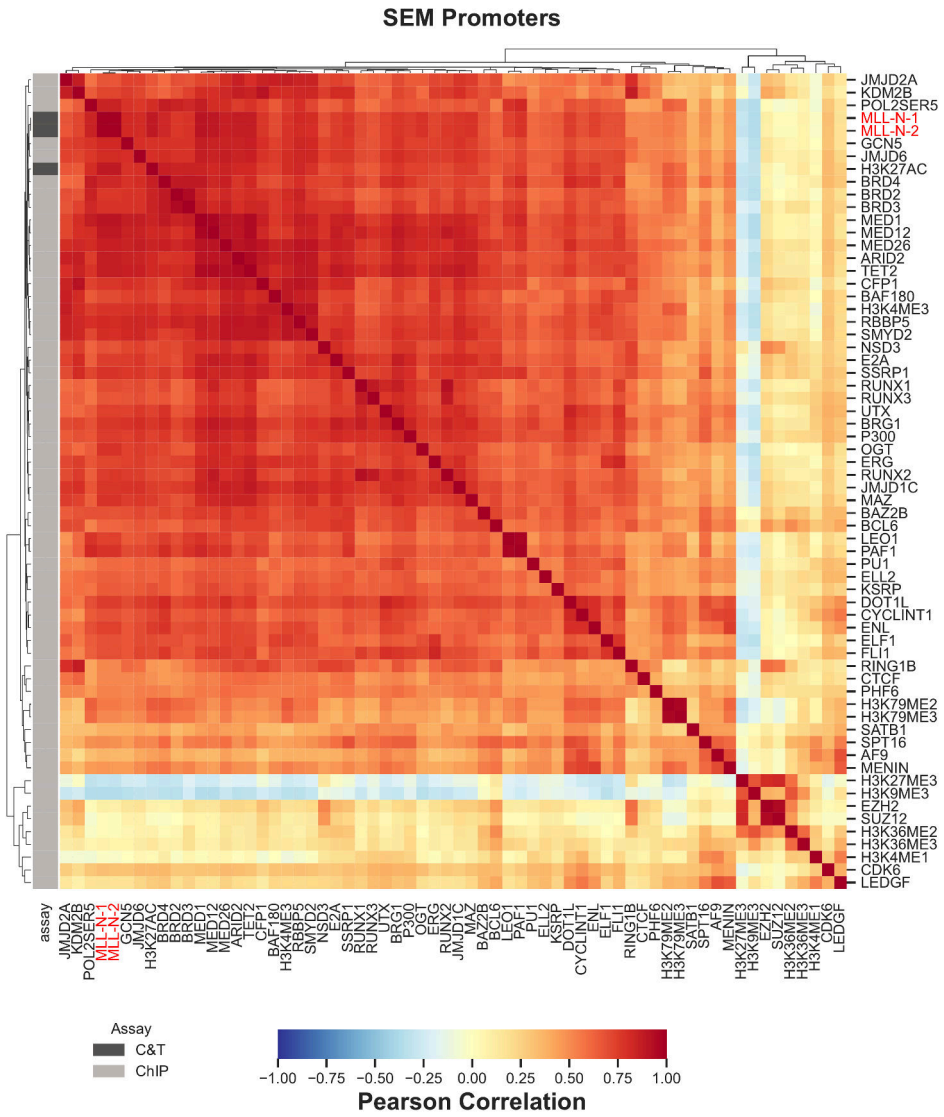


Figure 5.2: Continued next page.

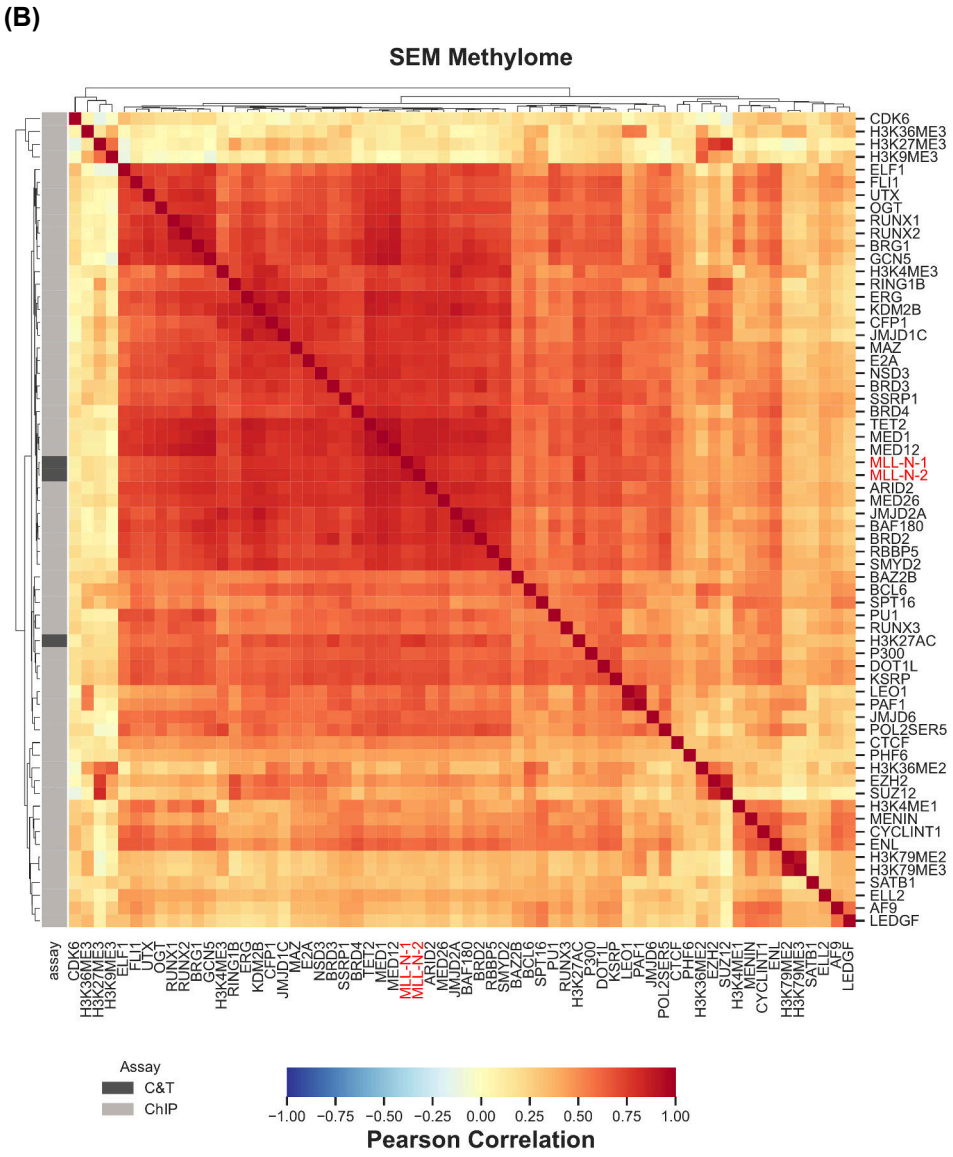


Figure 5.2: (Previous page) Correlation structure of MLL-N and co-binding factors across promoter and methylome regions in SEM cells. **(A)** Pearson correlation heatmap of MLL-N, co-binding proteins, and histone modifications across CpG island-rich promoter regions. MLL-N replicates (red) are moderately correlated with promoter-associated factors such as Menin, BRG1, Mediator components, and RNA polymerase II, but the overall clustering is more diffuse than in RCH-ACV cells, indicating a broader diversity of co-factor associations. Distinct clusters emerge, with a subset of factors including EZH2, LEDGF, and H3K79me3 forming a repressive module with weaker correlation to MLL-N. **(B)** Equivalent correlation heatmap for methylome-captured regions. MLL-N displays a more uniform pattern of associations across a wider range of cofactors and histone marks, reflecting the heterogeneity of distal genomic elements. While Menin and H3K27ac remain correlated, enhancer-associated features such as BRD4 and elongation-related factors like PAF1 and MED12 also show increased connectivity. These patterns suggest that, in SEM cells, MLL recruitment is influenced by a complex interplay between promoter tethering, enhancer activity, and chromatin-modifying complexes.

5.3.2 Model benchmarking

XGBoost was included as a strong gradient-boosted tree baseline owing to its competitive performance on tabular data, efficient training, and exact SHAP-based interpretability. Nevertheless, a neural approach (GANDALF) was preferred for the primary analyses because multi-target regression can share representations across outputs, target ranges can be constrained directly, and high-order, smooth interactions can be captured via gated feature learning. Both models were trained on identical splits with the same normalisation and evaluated with R^2 as the primary metric and mean squared error (MSE) as a secondary measure. To ensure a fair comparison, chromosome-aware splitting was used in both cases (chr8 for evaluation, chr9 held out for testing), identical feature matrices were supplied per context, and a fixed seed (42) was applied throughout.

Across promoter regions, GANDALF matched or modestly exceeded XGBoost, yielding higher macro R^2 and lower MSE (Figure 5.3). The gain was subtle but present over methylome-captured regions, where the broader and more heterogeneous feature space benefited from GANDALF's ability to model non-linear, higher-order interactions; here, improvements in macro R^2 were consistent across replicates and targets and were accompanied by uniformly lower MSE. These trends held despite the difference in feature numbers between cell types (25 factors in RCH-ACV versus 62 in SEM), indicating that the neural model's capacity and regularisation were sufficient to avoid overfitting while exploiting additional information where available.

Taken together, the benchmarking supports the use of GANDALF as the primary model for subsequent analyses. In the sections that follow, I therefore focus on model explanations from the GANDALF model to identify the co-factors and chromatin features that most strongly drive MLL-N predictions in each genomic context.

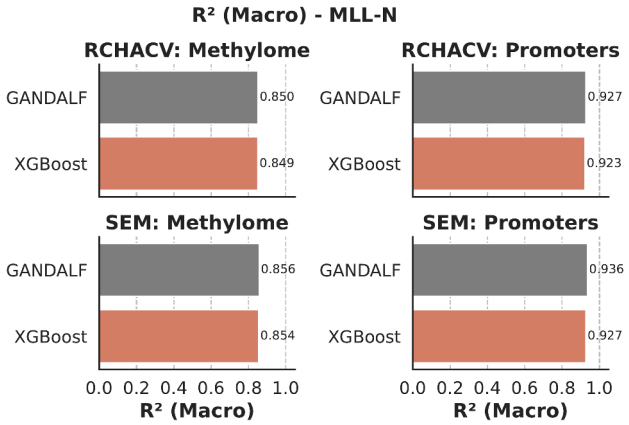


Figure 5.3: Model Benchmarking: Macro R^2 scores for GANDALF and XGBoost across different genomic contexts.

5.3.3 Co-factors which recruit MLL at promoters

Model evaluation

For promoters, a single GANDALF model was trained on the full feature panel pooled across both cell lines (i.e., all RCH-ACV- and SEM-derived tracks retained with their source labels). To assess whether the model learned the underlying MLL signal rather than dataset-specific artefacts, embeddings from the penultimate layer were extracted for the evaluation/test promoters, reduced with PCA, and visualised with UMAP. When points were coloured by the mean scaled MLL-N RPKM across targets, the manifold arranged promoters along a smooth continuum with a clear gradient from low to high signal (Figure 5.4A), indicating that the latent space encodes quantitative variation in MLL binding rather than discrete clusters. Notably, samples did not segregate by source cell line or replicate, and nearest-neighbour structure was driven by MLL intensity and co-factor composition, consistent with the model learning shared recruitment rules across datasets. Together, these observations show that the joint promoter model captured the MLL signal in a cell-line-agnostic fashion, providing a suitable basis for subsequent feature attribution and mechanistic interpretation.

I evaluated the final promoter model on the held-out test set, generating predictions separately for each replicate in each cell line. Scatter plots of predicted versus observed mean scaled MLL-N signal showed tight agreement along the identity line, with R^2 ranging from 0.93 to 0.95 and MSE between 0.003 and 0.004 (Figure 5.4B).

Performance was consistent across cell lines and replicates, indicating that the model generalises to unseen promoters and captures quantitative variation in MLL

binding rather than overfitting to a specific dataset.

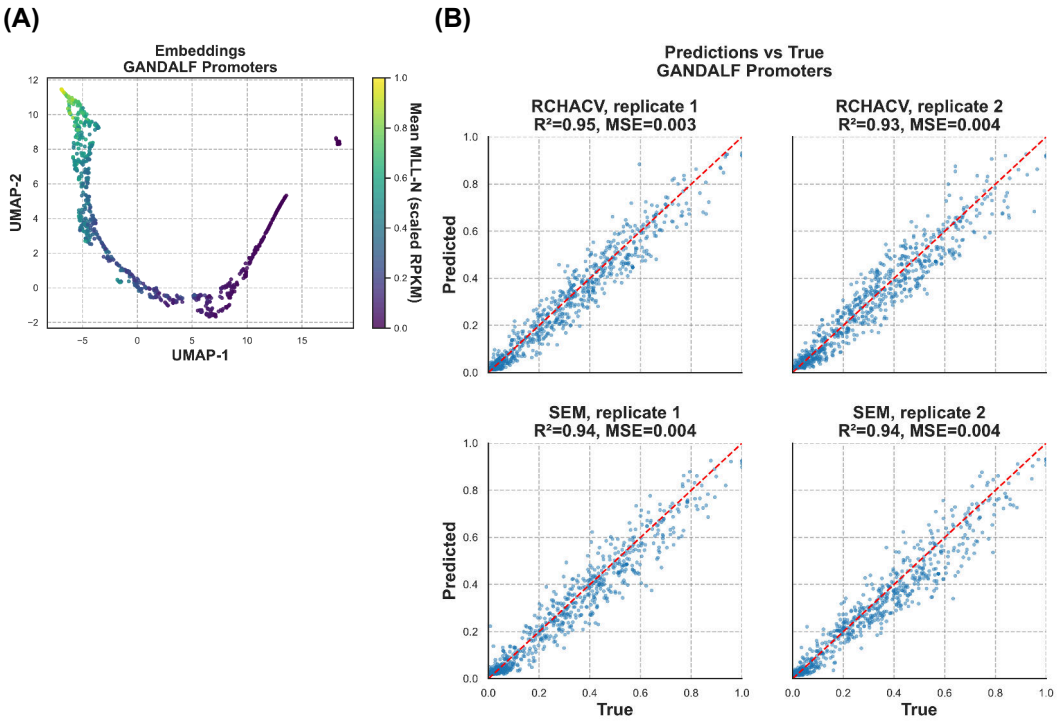


Figure 5.4: Promoter GANDALF prediction performance. **(A)** UMAP of GANDALF promoter embeddings coloured by mean scaled MLL-N signal. The penultimate-layer embeddings from the final promoter GANDALF model were reduced with PCA and visualised using UMAP. Each point represents a promoter, coloured by its mean scaled MLL-N RPKM across replicates. The promoters are arranged along a smooth continuum from low (purple) to high (yellow) MLL-N signal, indicating that the model has learned a quantitative representation of MLL-N binding. The absence of discrete clusters suggests that MLL-N occupancy varies gradually across the promoter landscape, with the embedding capturing shared recruitment rules rather than cell line-specific patterns. **(B)** Predicted versus observed MLL-N signal for the promoter model. Predictions from the final GANDALF promoter model plotted against ground truth for each replicate in RCH-ACV (top row) and SEM (bottom row). The red dashed line denotes y = x. Across panels, the model achieves R^2 = 0.93–0.95 with MSE = 0.003–0.004, demonstrating accurate generalisation to held-out promoters in both cell lines.

Feature Importance at Promoters

To identify the factors most predictive of MLL-N binding at promoters, I applied a Kernel SHAP explainer to the final promoter model and computed mean absolute SHAP values per feature across the 777 held-out promoter regions, separately for each replicate in each cell line. To preserve provenance, features were labelled by chromatin immunoprecipitation (ChIP) target and source cell line (e.g., MENIN_RCHACV, H3K27ac_SEM). Hierarchical clustering of the mean absolute SHAP matrix revealed three coherent groups (Figure 5.5A). A top cluster comprised high-importance features shared across both contexts, including MENIN_RCHACV,

POL2SER5_SEM, BRG1_RCHACV, ARID2_SEM, and TET2_SEM. A second cluster contained SEM-sourced features that were preferentially predictive in SEM (notably H3K27ac_SEM, MED1_SEM, KDM2B_SEM, MED12_SEM). A third cluster contained RCH-sourced features that were more predictive in RCH-ACV (including; H3K27ac_RCHACV, H3K4me3_RCHACV, E2A_RCHACV, ELF1_RCHACV, RUNX2_RCHACV, POL2_RCHACV, PAF1_RCHACV).

The low SHAP ranking of both H3K4me3_SEM and H3K4me3_RCHACV for SEM predictions across both the promoter and methylome models (Figure 5.8A) reflects the loss of the SET domain in MLL-AF4. Unlike WT MLL, the fusion protein cannot deposit H3K4me3 at its targets, and this mark is not important in these models for predicting MLL occupancy in SEM regardless of its source. In contrast, H3K4me3_RCHACV was highly ranked in RCH-ACV predictions, consistent with WT MLL actively depositing the mark at bound sites.

For the top three predictors within each cell line, scaled RPKM values in the test set showed strong positive associations with mean scaled MLL-N signal (Spearman's $\rho > 0.82$ in all cases; all tests significant Figure 5.5B), confirming that features ranked highly by SHAP also exhibit monotonic relationships with observed binding. Biologically, the highest-ranking promoter features mark active transcription: Menin (direct MLL interactor), H3K27ac (active promoter/enhancer acetylation), RNAP II (initiation/pausing), and BRG1 (SWI/SNF remodeller). Together, these results indicate that the model learned MLL-N occupancy as a signature of active promoters in both SEM and RCH-ACV, rather than as a cell-line-specific artefact.

To conclude the promoter analysis, the SHAP feature extraction confirmed and extended the patterns observed in the initial correlation heatmaps. Features that were strongly correlated with MLL-N at promoters such as Menin, Ser5-phosphorylated RNA Pol II, BRG1, and H3K27ac were also those ranked most predictive by the model, consistent with their roles in active transcription and with the observed co-occurrence of MLL-N at promoter-proximal chromatin. The separation of SEM and RCH-ACV specific feature clusters mirrored the context-dependent modules seen in the correlation matrices, highlighting how MLL engages with distinct co-factor networks in MLL versus MLL-FP driven settings. This concordance between unsupervised correlation patterns and model-derived importance scores shows that GANDALF learned biologically meaningful relationships rather than artefacts, establishing a robust foundation for the next stage of analysis focused on the more heterogeneous methylome regions.

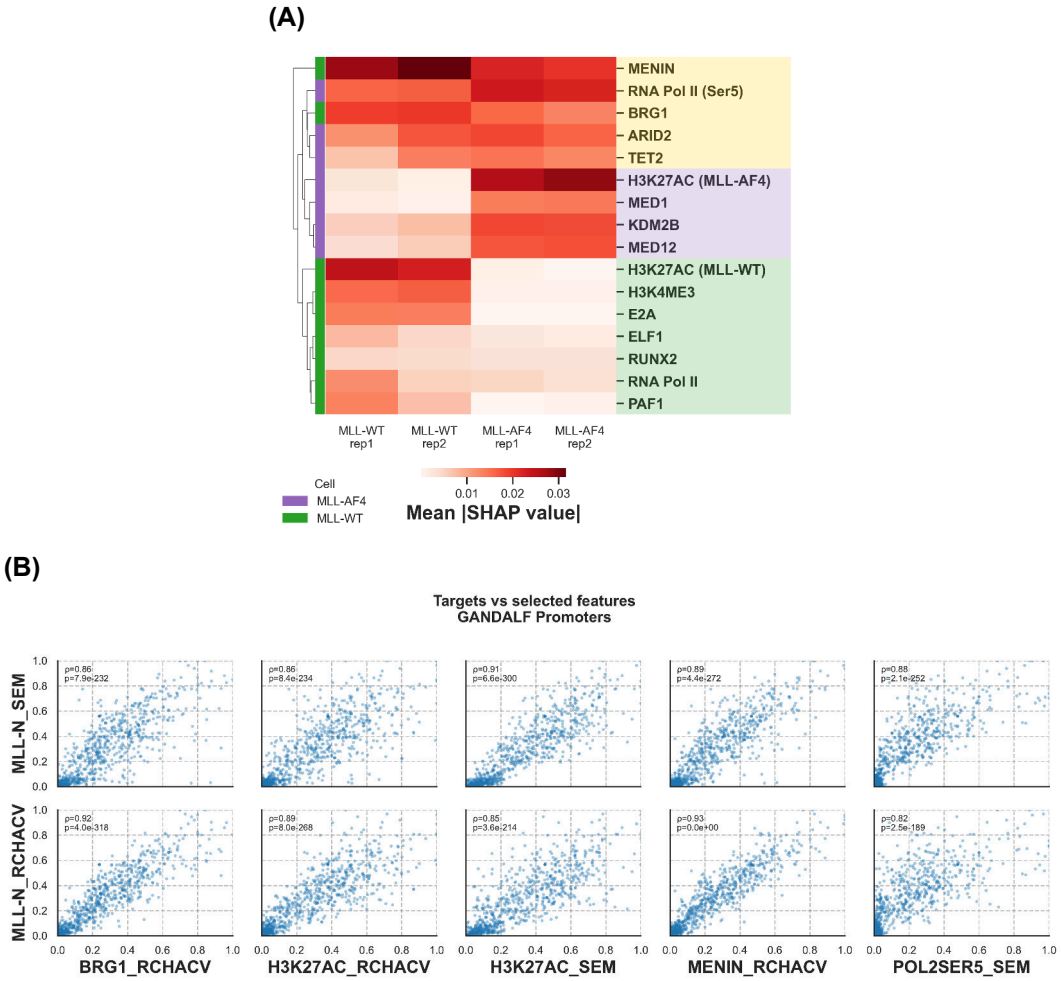


Figure 5.5: SHAP feature importance and associations for the promoter GANDALF model. **(A)** Heatmap of mean absolute SHAP values for the four MLL-N outputs across 777 held-out promoter regions. Features are colour-coded by source cell line (MLL-AF4 in purple, MLL-WT in green). Three distinct clusters emerge: a shared high-importance group spanning both cell lines (yellow highlight) comprising Menin, RNA Pol II (Ser5), BRG1, ARID2 and TET2; a MLL-AF4-specific group (purple highlight) comprising H3K27ac (MLL-AF4), MED1, KDM2B and MED12; and a MLL-WT-specific group (green highlight) comprising H3K27ac (MLL-WT), H3K4me3, E2A, ELF1, RUNX2, RNA Pol II and PAF1. **(B)** Scatter plots showing the relationship between scaled RPKM values of the top three features for each cell line and mean scaled MLL-N signal across the test set. All features show strong positive monotonic associations with MLL-N binding (Spearman's ρ > 0.82, all $p < 1^{-200}$), confirming that the most predictive features identified by SHAP also display direct biological correlation with MLL occupancy. Together, these plots demonstrate that GANDALF has learned biologically meaningful recruitment signals at promoters.

5.3.4 Co-factors which recruit MLL at methylome regions

Having established that GANDALF accurately predicted MLL occupancy at promoters and identified biologically meaningful features, I next tested whether this ap-

proach extended to more heterogeneous methylome-captured regions. These regions encompass a mix of promoters, enhancers, and distal elements, presenting a more complex modelling challenge. To visualise the learned representations, I extracted embeddings from the model's penultimate layer and reduced them using PCA followed by UMAP projection (Figure 5.6A). As with the promoter model, the methylome embeddings showed a clear gradient of MLL-N signal, indicating that the model captured quantitative variation in MLL occupancy. However, the structure was less smooth than in the promoter analysis, reflecting the greater number and diversity of genomic regions represented in the methylome dataset.

The model's predictive performance was evaluated on a held-out test set covering chromosome 9 (Figure 5.6B). Predictions showed strong agreement with observed MLL-N signal across both cell lines, with R^2 ranging from 0.84 to 0.88 and MSE between 0.004 and 0.005. While these values were slightly lower than those achieved for promoters, they demonstrate that the model generalised well despite the added complexity of these regions. Performance was consistent across replicates and between RCH-ACV and SEM, indicating that the predictive rules learned by the model are robust across contexts.

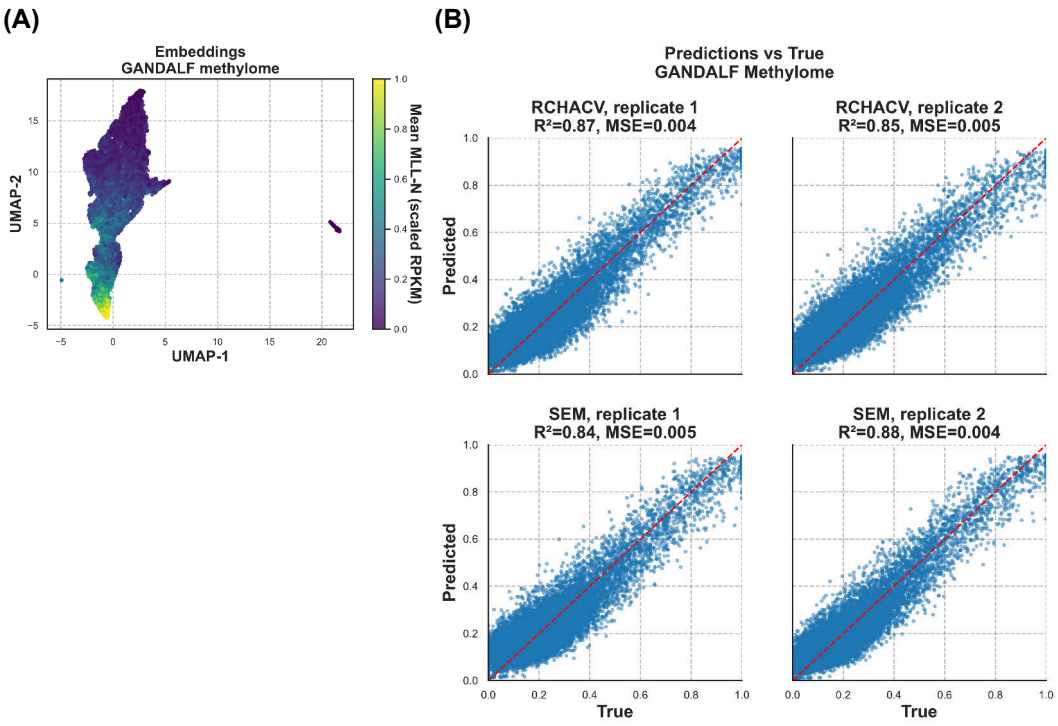


Figure 5.6: Continued next page.

Figure 5.6: (Previous page) Methylome GANDALF prediction performance. **(A)** UMAP of GANDALF methylome embeddings. Two-dimensional projection of the penultimate-layer embeddings for methylome-captured regions, coloured by mean scaled MLL-N RPKM. The model organises regions according to MLL-N intensity, though the structure is more heterogeneous than in promoters, reflecting diverse underlying chromatin states. **(B)** Predicted versus observed MLL-N signal for the methylome model. Scatter plots of predicted versus true MLL-N values for each replicate in RCH-ACV (top row) and SEM (bottom row). The red dashed line represents y = x. The model achieved R^2 = 0.84–0.88 and MSE = 0.004–0.005, indicating strong generalisation across diverse genomic regions.

The model also accurately captured cell-line-specific binding patterns at individual loci. For example, at the *GNAQ* locus (Figure 5.7), the model correctly predicted both shared and differential MLL-N occupancy between RCH-ACV and SEM, reflecting the influence of the MLL-AF4 fusion in SEM driving higher binding into the gene body. This locus exemplifies how the model integrates co-factor signals to predict nuanced binding patterns across MLL-AF4 and MLL contexts.

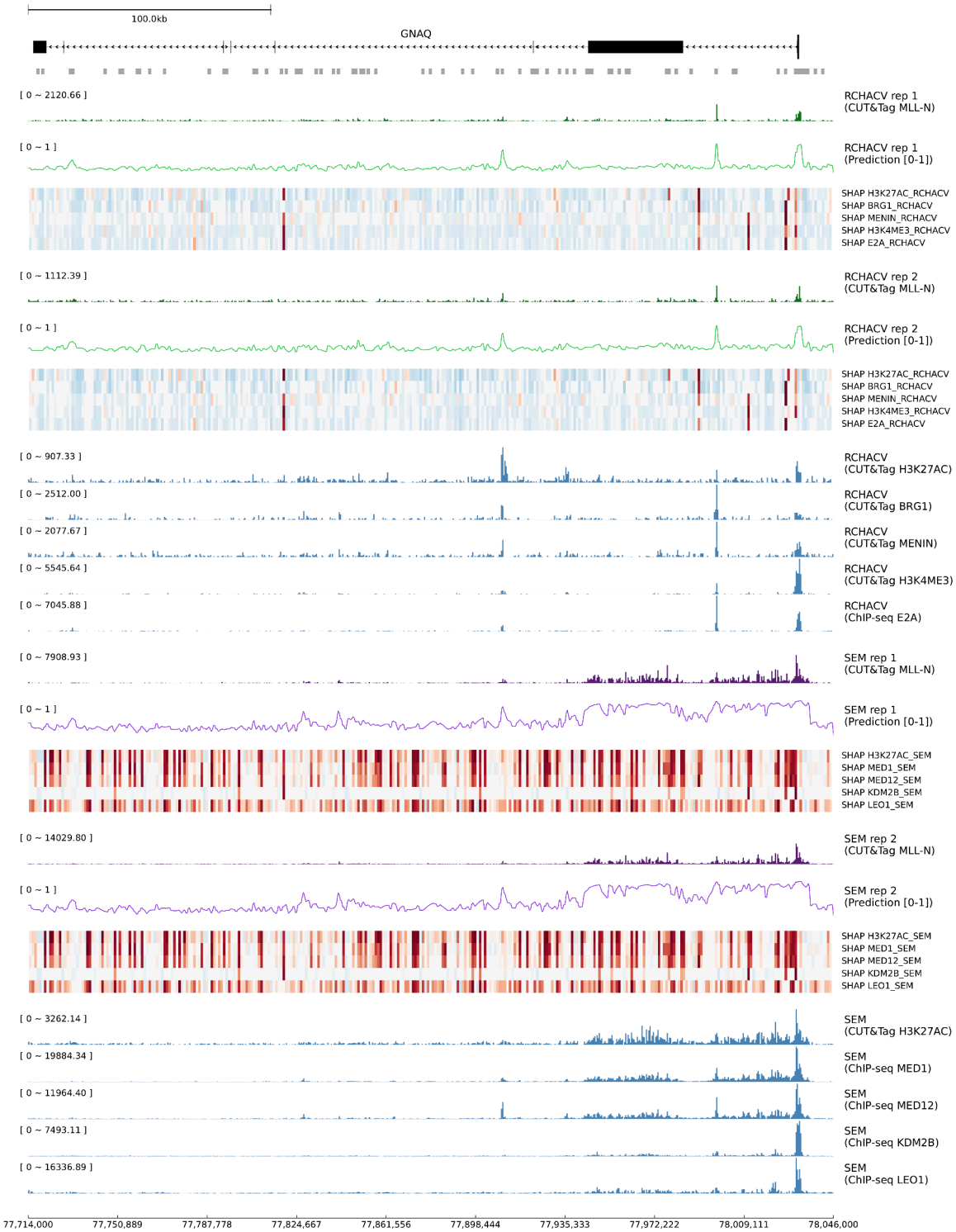


Figure 5.7: Continued next page.

Figure 5.7: (Previous page) Genome browser snapshot of the *GNAQ* locus (chr9:77,714,000–78,046,000). Per-replicate MLL-N CUT&Tag signal (RPKM) is shown alongside GANDALF predictions [0–1] from the methylome model used to predict tiled chr9 windows. For each replicate, SHAP contributions for selected features are displayed as heatmaps (red = positive contribution, blue = negative contribution, scaled per feature), followed by the corresponding ChIP-seq or CUT&Tag signal tracks for these features. RCH-ACV tracks are shown in green (MLL-N and predictions) and SEM tracks in purple. Training panel regions are shown in grey. The model captures both shared and cell-line-specific binding patterns, with predictions extending further into the gene body in the MLL-AF4 context (SEM).

Feature importance at methylome regions

To uncover the features driving MLL binding at these diverse regions, I applied Kernel SHAP to the final methylome model (Figure 5.8A). In RCH-ACV, the top predictors overlapped strongly with those identified at promoters. H3K27ac, BRG1, and Menin remained highly important, reflecting their central role in promoter tethering and active transcription. In addition, H3K4me3 was strongly predictive, consistent with it being deposited by MLL. Notably, E2A emerged as a top predictor in RCH-ACV, consistent with its fusion to PBX1 in this cell line and its known ability to recruit p300 and the Mediator complex to activate transcription (Pi, W.-C. et al. 2020). This shows that MLL recruitment in RCH-ACV is associated with actively transcribed genes.

In SEM, the pattern of predictors shifted markedly compared to promoters. The highest-ranked features included H3K27ac, Mediator subunits MED1 and MED12, LEO1 from the RNA polymerase II-associated PAF complex, and KDM2B. KDM2B, which binds unmethylated CpG islands and recruits variant PRC1 complexes (Farcas, A. M. et al. 2012; Blackledge, N. P. et al. 2014), was particularly enriched among SEM predictors. Its apparent importance likely reflects the shared occupancy of CpG-rich regions by both MLL-N and PRC-associated factors rather than a direct mechanistic link. Such co-localisation accords with a chromatin-sampling model, in which transient engagement of CpG islands by multiple factors produces time-averaged overlap without implying direct recruitment (Klose, R. J. et al. 2013). These features suggest that MLL binding in SEM is concentrated at CpG-rich regions and tightly coupled to enhancer activity and transcriptional elongation, consistent with the MLL-AF4 fusion driving activation at distal regulatory elements (Smith, A. et al. 2025a).

The top-ranked features for each cell line were strongly associated with measured MLL-N binding in the test set, with monotonic positive relationships observed in all

cases (Spearman's $\rho \geq 0.70$, $p < 10^{-200}$) (Figure 5.8B). This confirms that the features highlighted by SHAP reflect true biological drivers of MLL recruitment.

The methylome analysis demonstrated that GANDALF can accurately model MLL recruitment across a diverse set of genomic elements from co-factors, achieving high predictive accuracy while capturing meaningful biological patterns. In RCH-ACV, the model relied primarily on promoter-associated features, consistent with MLL binding at transcription start sites. In contrast, SEM predictions were driven by enhancer and elongation machinery, reflecting the fusion protein's role in hijacking distal regulatory elements. These results extend the promoter findings and highlight the context-dependent nature of MLL recruitment across the genome.

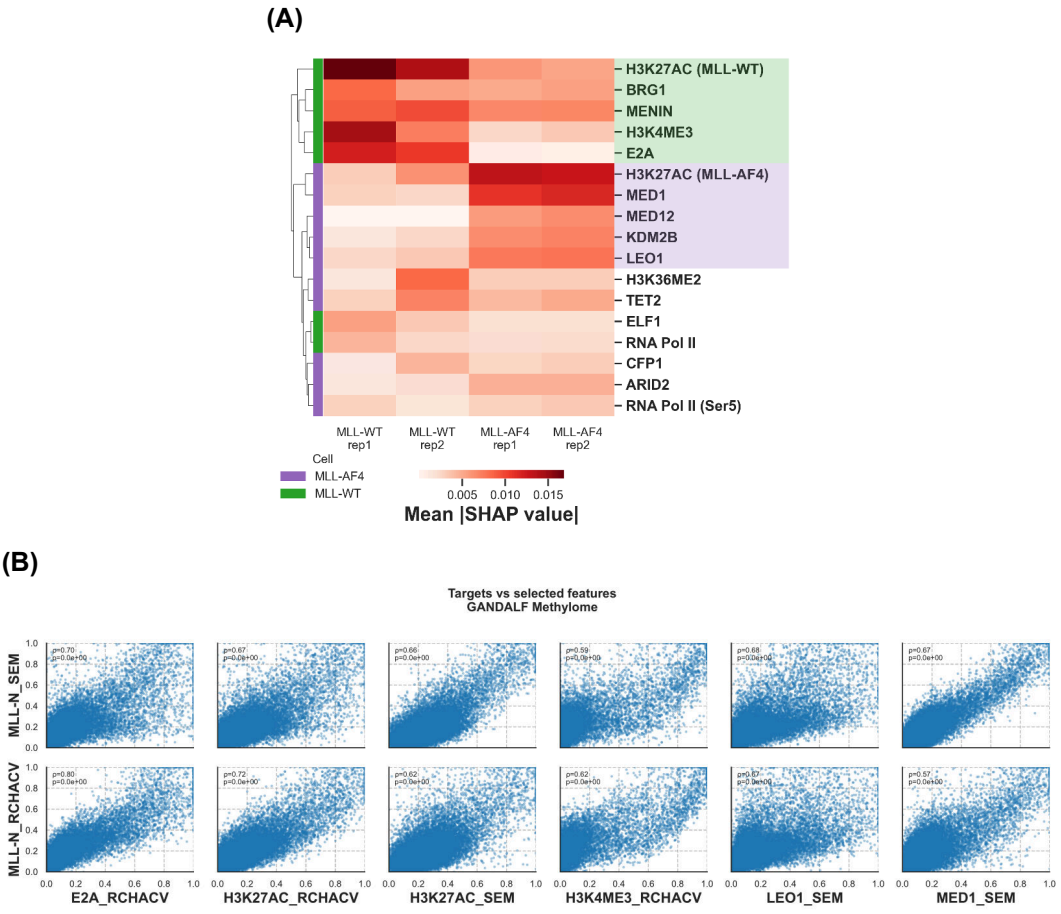


Figure 5.8: Continued next page.

Figure 5.8: (Previous page) SHAP feature importance and associations for the methylome GANDALF model. **(A)** Heatmap of mean absolute SHAP values across the four MLL-N outputs (MLL-WT and MLL-AF4 replicates). Features are colour-coded by source cell line (MLL-WT in green, MLL-AF4 in purple). Two distinct clusters emerge: a MLL-WT-specific group (green highlight) comprising H3K27ac, BRG1, Menin, H3K4me3, and E2A, reflecting promoter-linked recruitment; and a MLL-AF4-specific group (purple highlight) comprising H3K27ac, MED1, MED12, KDM2B, and LEO1, reflecting enhancer and elongation machinery. A third lower-importance group, including H3K36ME2, TET2, ELF1, RNA Pol II, CFP1, ARID2, and RNA Pol II (Ser5), shows lower importance across both contexts. **(B)** Scatter plots showing relationships between scaled RPKM values for the top three predictors per cell line and mean scaled MLL-N signal across the test set. All features show strong monotonic associations, confirming their direct link to MLL occupancy.

5.4 Discussion

I applied a ML framework to predict and interpret MLL binding across two genomic contexts: promoters and the broader Twist methylome regions. By integrating CUT&Tag and ChIP-seq datasets from both MLL (RCH-ACV) and MLL-AF4 (SEM) B-cell leukaemic cell lines, I trained the GANDALF neural network to model MLL recruitment as a function of co-binding transcription factors, histone modifications, and chromatin-associated proteins.

At promoters, the model achieved very high predictive performance (R^2 = 0.93-0.95), accurately capturing quantitative variation in MLL binding. Feature attribution using Kernel SHAP confirmed that the most important predictors were Menin and LEDGF, promoter-associated marks (H3K27ac, H3K4me3), RNA Polymerase II (Ser5-phosphorylated Pol II ChIP-seq from SEM, marking promoter-proximal initiation and pausing, was predictive for MLL binding in both cell lines; total Pol II ChIP-seq from RCH-ACV contributed less predictive signal), and SWI/SNF remodelling via BRG1 consistent with known mechanisms of MLL recruitment.

When applied to methylome regions, which includes CpG rich promoters as well as other distal CpG rich genomic regions, the model remained strong (R^2 = 0.84-0.88) but the feature landscape shifted markedly. In RCH-ACV, predictors were still promoter associated, whereas SEM predictions were dominated by enhancer and elongation associated features, including Mediator (MED1, MED12), the PAF complex component LEO1, BRD4, and KDM2B reflecting the MLL-AF4 fusion's role in hijacking transcriptional elongation machinery. These patterns were mirrored in correlation heatmaps and SHAP feature clusters, indicating distinct predictive feature landscapes in MLL occupancy in MLL compared with MLL-AF4 contexts.

Together, these results show that this approach not only predicts MLL binding with high accuracy but also provides mechanistic insight, linking co-binding patterns to MLL occupancy as a signature of active promoters. MLL was predicted by factors that are associated with active gene transcription and in a MLL-AF4 context these predictors shifted to elongation associated factors reflecting the contribution of the fusion partner protein to MLL binding site specificity.

5.4.1 Mechanistic interpretation of MLL recruitment

These analyses indicate to two interlinked modes of MLL recruitment that diverge between MLL and MLLfusion contexts. Factors associated with active gene pro-

moters were predictive of MLL binding such as H3K27ac and RNA polymerase but in the MLL-AF4 attributions the important features also included elongation-associated factors such as Mediator and LEO1 which is a component of the PAF1 complex.

At promoters, MLL binding tracked closely with Menin, consistent with a tethering mechanism that positions MLL at transcriptionally active, CpG-dense loci (Yokoyama, A. et al. 2005; Milne, T. A. et al. 2010). Prominent H3K4me3 and H3K27ac signals accord with MLL's canonical role in maintaining promoter competence through histone methylation (Ruthenburg, A. J. et al. 2011; Tie, F. et al. 2009). The importance of RNA polymerase II and BRG1 in both the correlation maps and the SHAP rankings supports a promoter-initiation module that couples chromatin remodelling, initiation and polymerase pausing (Clapier, C. R. and Cairns, B. R. 2009; Shilatifard, A. 2012). Although LEDGF is a recognised partner in this tethering complex, it did not rank among the top SHAP predictors here. This is a common behaviour when features are highly correlated, because kernel SHAP often assigns most importance to one member of a correlated set, especially when the data is not representative as in the case of LEDGF ChIP-seq due to data quality. In this case, importance was instead attributed to Menin and promoter-linked marks such as H3K4me3, H3K27ac and RNA polymerase II. Feature ablation, in which models are retrained on reduced feature subsets, could help disentangle whether LEDGF carries unique predictive information or is redundant with correlated features under these conditions by sequentially removing LEDGF or MENIN in the feature set and evaluating the impact of each on model performance and resulting SHAP scores for each. Functionally, LEDGF is dispensable for normal haematopoiesis but essential for MLLr leukaemogenesis, underscoring the centrality of the Menin-LEDGF-MLL axis (El Ashkar, S. et al. 2017). In line with this dependence on promoter-proximal transcriptional control, targeting chromatin readers that sustain high output, most notably BRD4, with BET inhibitors disrupts oncogenic transcription in MLLr leukaemia (Dawson, M. A. et al. 2011; Roe, J.-S. et al. 2015; Bhagwat, A. S. et al. 2016).

In methylome-captured regions that include enhancers and other distal elements, the recruitment landscape shifts. In RCH-ACV cells, MLL promoter-linked features such as Menin and BRG1 were predictive of MLL occupancy. In SEM cells, which harbour MLL-AF4, the strongest predictors include Mediator subunits MED1 and MED12, LEO1 from the PAF complex, BRD4, and KDM2B, factors associated with active enhancer architecture and the release of paused RNA polymerase II into productive elongation. Experimental work shows that MLL-AF4 cooperates directly

with PAF1 and the FACT complex to generate dense enhancer-promoter hubs that sustain oncogenic transcription, mirroring the elongation-associated signals captured by the SEM models (Crump, N. T. et al. 2023; Muntean, A. G. et al. 2010). MED1 and MED12 are two of the many subunits of the highly complex Mediator protein complex which is thought to bridge promoters and enhancers contributing to gene regulation, however the individual role of the subunits are not fully understood (Richter, W. F. et al. 2022). In haematopoietic contexts deletion of MED12 resulted in H3K27ac loss and enhancer inactivation through the loss of p300 (Aranda-Orgilles, B. et al. 2016). MED1 is a transcriptional co-activator and in haematopoiesis conditional knock-outs in mice showed that Med1 was essential for erythroid maturation through impaired activation of GATA1 dependent genes (Stumpf, M. et al. 2010). BRD4, a reader of enhancer acetylation, is likewise required for maintaining high-output transcription at fusion-driven loci, interacting with acetylated histones and mediator aligning with its prominence among SEM predictors (Kanno, T. et al. 2014; Bhagwat, A. S. et al. 2016).

Elongation-linked chromatin marks further connect MLL-AF4 to enhancer function. A subset of enhancers is dependent on DOT1L-mediated H3K79 methylation, and DOT1L inhibition collapses these elements and reduces expression of fusion targets, providing a mechanistic bridge between elongation marks and enhancer maintenance (Godfrey, L. et al. 2019). Together, these observations are consistent with MLL-AF4 operating within elongation-competent enhancer environments, where H3K79me contributes to maintaining enhancer activity and target-gene transcription.

Finally, enhancer control in MLLr acute lymphoblastic leukaemia is heterogeneous across patients. Single-cell and multi-omic analyses reveal variable enhancer usage at canonical oncogenes such as MEIS1 and RUNX family loci, which shapes transcriptional output between individuals (Smith, A. et al. 2025a). This clinical variability is consistent with the more heterogeneous latent space and slightly reduced predictive smoothness observed for methylome regions relative to promoters.

Taken together, the data support a model in which MLL remains largely promoter-focused through Menin-LEDGF tethering and promoter-associated histone marks, whereas MLL-AF4 extends recruitment into an enhancer circuit that engages Mediator, PAF, BRD4, DOT1L-linked elongation, and CpG-sensitive chromatin regulators. This shift rewires the chromatin landscape to favour distal enhancer activation and the formation of high-output transcriptional hubs that sustain leukaemic gene expression, in agreement with recent experimental studies and patient-level observations (Crump, N. T. et al. 2023; Godfrey, L. et al. 2019; Smith, A. et al. 2025a).

These mechanistic inferences have direct translational relevance. The promoter-tethering axis supports the rationale for Menin-targeted therapies in MLLr leukaemia, whereas the enhancer/elongation axis highlights vulnerability to perturbation of BRD4, Mediator and DOT1L, with preclinical data already demonstrating attenuation of fusion-driven transcription under BET or DOT1L inhibition (Dawson, M. A. et al. 2011; Pelish, H. E. et al. 2015; Godfrey, L. et al. 2019). In SEM, Mediator subunits and H3K27AC are among the strongest predictors of MLL-AF4 occupancy across both promoters and methylome-captured distal regions, a pattern absent in the MLL wild-type models where Menin and BRG1 dominate instead. This fusion-specific association with Mediator is independently supported by an unbiased model trained on the same cell line, in which MED1 signal was predictive of SEM-specific enhancer activity (Smith, A. et al. 2025a), and by functional evidence that MLL-AF4 knockdown reduces H3K27ac at SEM-specific enhancers in the same study. Together, these observations suggest a relationship between MLL-AF4 binding and Mediator-associated enhancer activity, and imply that patient-specific variation in MLL-AF4 binding would produce corresponding variation in enhancer-associated cofactor environments. Although causal directionality cannot be inferred from the cofactor model alone, this association is consistent with a role for MLL-AF4 in stabilising patient-specific enhancer activity, providing a mechanistic basis for the transcriptional heterogeneity observed between individuals in MLLr ALL. In this way, the model's ranked predictors provide testable hypotheses about which cofactors anchor MLL at promoters and which sustain fusion-dependent activity at distal elements, and they point to specific nodes for functional validation in future perturbation studies; however, as correlative models, they cannot establish causal directionality between cofactor co-occupancy and MLL recruitment.

5.4.2 Limitations

This study integrates co-binding and chromatin features to model MLL recruitment, but several constraints limit interpretation Firstly, because of data availability the feature sets differed between cell lines (RCH-ACV: 25 profiled factors; SEM: 62) Although source labels were retained and a uniform scaling pipeline applied, the unequal number of factors assayed for each cell line can affect model capacity and SHAP ranks, potentially accentuating SEM-specific signals.

Secondly, the analyses rely on CUT&Tag and ChIP-seq tracks generated with different antibodies and protocols; despite replicate concordance for MLL-N and chromosome-aware splits, residual batch effects and antibody biases may persist

and partially confound results.

Thirdly, track quality and background levels vary; features with elevated background (e.g., LEDGF, AF9, CDK6, SATB1) can be variance-compressed under zero-biased scaling, yielding weaker correlations and lower SHAP importance.

Fourthly, SHAP explains model behaviour, not causality. Under multicollinearity, kernel SHAP can concentrate attribution on one of several tightly correlated predictors (e.g., down-ranking LEDGF relative to Menin and promoter marks). The background distribution and value-function assumptions also shape SHAP magnitudes; importance should be read as predictive contribution under this model and feature set, not as proof of necessity.

Fifthly, ordinality remains unresolved: observed co-occurrence is compatible with scenarios in which MLL-AF4 is recruited to pre-existing, elongation-competent enhancers or, conversely, pre-assembled enhancer complexes attract MLL-AF4 these analyses are predictive, not causal.

Finally, generalisability was assessed in two B-ALL cell lines with distinct drivers (E2A-PBX1 in RCH-ACV; MLL-AF4 in SEM); while informative, this does not capture the spectrum of MLL fusions or patient variability, and cell-line idiosyncrasies may limit transfer to primary samples. The models also do not encode primary DNA sequence or 3D chromatin architecture explicitly, and the methylome regions were pre-selected, potentially biasing analyses towards CpG-rich elements and underrepresenting sequence-encoded recruitment logic and long-range constraints.

5.4.3 Conclusion

In this study I modelled MLL recruitment as a supervised prediction problem over co-binding proteins and chromatin features, training a single multi-target GANDALF model per genomic context and validating it across two B-ALL cell lines. The models generalised well to held-out data and their latent spaces ordered loci by MLL signal rather than by cell-line identity, indicating that they captured quantitative variation in occupancy rather than batch effects. Feature attributions closed the loop from prediction to mechanism. At promoters, Menin, RNA polymerase II, BRG1 and promoter marks such as H3K27ac and H3K4me3 dominated, consistent with a promoter-initiation module. Across methylome regions, especially in SEM with MLL-AF4, the signal shifted towards enhancer and elongation machinery, with Mediator, BRD4, PAF-associated factors, DOT1L-linked activity and CpG-sensing regulators such as KDM2B emerging as key determinants. Together these results support a

simple organising principle. MLL is largely promoter-focused, anchored at CpG-rich transcription start sites and coupled to transcription initiation. MLL-AF4 extends this recruitment network into distal elements, engaging co-activator and elongation modules. The agreement between unsupervised correlations, predictive performance and SHAP-based explanations argues that the models are recovering biologically meaningful rules rather than artefacts, while differences between cell lines highlight the context dependence of those rules. Beyond explaining existing data, the framework provides a practical way to generate ranked, testable hypotheses about co-factor dependencies at specific loci. By integrating ML with chromatin biology, this work offers a quantitative map of how co-factor constellations specify MLL occupancy across the genome and a foundation for targeted perturbation experiments that can distinguish necessary drivers from correlated bystanders.

6 The Role of DNA Methylation in MLL Recruitment

6.1 Introduction

6.1.1 Detecting DNA methylation

Hotchkiss (1948) first detected an extra pyrimidine in DNA hydrolysates, distinct from cytosine and the other bases. Wyatt (1951) identified this as 5-methylcytosine and quantified it in DNA, with broader cross-species measurements later achieved by mass spectrometry (Vanyushin, B. F. et al. 1970). Methylation-sensitive restriction enzymes then revealed that methylation is non-randomly distributed across the genome (Bird, A. P. and Southern, E. M. 1978). Through the 1980s, GC-MS/HPLC and restriction enzyme assays coupled to Southern blots or PCR (Singer-Sam, J. et al. 1990) improved sensitivity but remained limited to enzyme recognition motifs.

A major advance came with bisulphite sequencing (Frommer, M. et al. 1992), which converts unmethylated cytosines to Uracil (U) while leaving 5mC unchanged, enabling strand-specific, base-resolution maps. This chemistry underpinned locus-specific PCR assays, arrays, and ultimately WGBS. In *Arabidopsis*, Cokus et al. (2008) produced the first whole-genome methylome, showing context-specific methylation (CpG, CHG, CHH) and nucleosome-phased patterning. In humans, Lister et al. 2009 generated deeply covered methylomes (ES cells vs. fibroblasts), revealing widespread non-CpG methylation in pluripotent cells and large cell-state specific differences linked to regulatory programmes.

Despite its impact, bisulphite cannot distinguish 5mC from 5hmC, prompting oxidative bisulphite sequencing (oxBS), which first oxidises 5hmC to 5fC/5caC and then uses bisulphite to convert those to U (read as T in sequencing) while 5mC remains C, yielding a direct 5mC map. In TET-assisted bisulphite sequencing (TAB-seq), β-Glucosyltransferase (βGT) protects 5hmC and TET1 oxidises 5mC to 5caC, which

bisulphite converts to U (read as T), so sites that remain C correspond to 5hmC.

However, bisulphite treatment results in DNA fragmentation and an AT skew in the resulting data. Bisulphite free Tet-assisted Pyridine Borane Sequencing (TAPS) (Liu, Y. et al. 2019) enabled base-resolution detection of 5mC with markedly better DNA integrity, higher mapping efficiency, and fewer artefacts through TET oxidation of 5mC and 5hmC followed by pyridine-borane reduction of 5caC to dihydrouracil which is read as thymine. TAPSβ includes βGT protection of 5hmC prior to TET oxidation and the combination of both methods can resolve 5mC and 5hmC by comparing assays. Although this method does yield more coverage and higher mapping rates, multiple libraries are required in detecting both 5mC and 5hmC.

In this work, I performed simultaneous methylation sequencing (Füllgrabe, J. et al. 2023), a bisulphite-free approach that uses enzymatic conversion of unmodified cytosines coupled with strand copying and βGT protection and TET oxidisation to read A, C, G, T, 5mC, and 5hmC simultaneously (Figure 6.1). DNA is fragmented and hairpins are ligated to the double-stranded DNA and the strands are then separated. Each strand is then copied through synthesis using Klenow exo-polymerase and short sequencing adapters are ligated. 5hmC is protected from copying and deamination with βGT. 5mC is copied to the synthesised strand by DNA Methyltransferase 5 (DNMT5). TET2 oxidises the 5mC on both the original and copied strands preventing deamination. UvrD helicase opens the hairpin and unmodified cytosines to are converted to U with cytosine deaminase Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3A (APOBEC3A). Libraries are then PCR-amplified and paired-end sequenced. The bases are then resolved bioinformatically by comparison between the original strand (Read 1) and the copied strand (Read 2) as A (A/T), G (G/C), T (T/A), in addition to the unmodified C (T/T), 5mC (C/C), and 5hmC (C/T), allowing simultaneous detection of both 5mC and 5hmC without bisulphite-induced DNA damage or AT-skew.

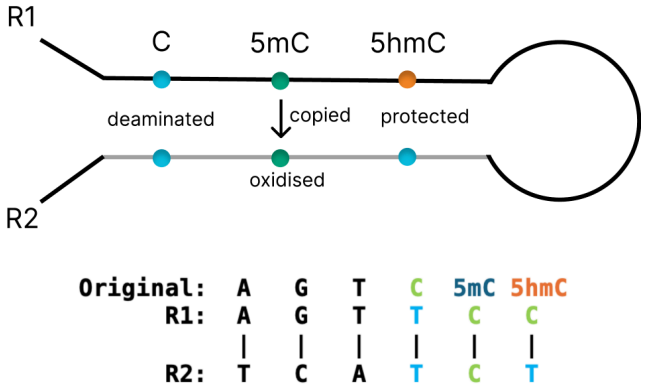


Figure 6.1: Simultaneous methylation sequencing assay overview. Hairpin adapters link the two strands of each fragmented DNA molecule; a copy strand (R2; grey) is synthesised from the original (R1). β-glucosyltransferase (βGT) protects 5hmC, then DNMT5 copies 5mC across 5mC onto R2. TET2 oxidises 5mC on both strands to 5caC, which is refractory to deamination. UvrD unwinds the hairpin and APOBEC3A deaminates only unprotected cytosines. Paired-end sequencing and bioinformatic comparison of R1/R2 resolve cytosine states at CpGs: T/T = unmodified C, C/C = 5mC, C/T = 5hmC. Adapted from (Füllgrabe, J. et al. 2023).

6.1.2 DNA methylation and DNA binding proteins

Cytosine methylation alters recognition motifs and local accessibility. Some proteins preferentially recognise methylated CpGs via a methyl-CpG binding domain, including members of the MBD family such as MeCP2 and MBD2 (Zhu, H. et al. 2016; Héberlé, É. and Bardet, A. F. 2019). By contrast, many sequence-specific transcription factors lose affinity when CpGs within their motifs are methylated; a classic example is c-Myc at the E-box (Prendergast, G. C. and Ziff, E. B. 1991; Yin, Y. et al. 2017). In some cases methylation can also create or shift specificity rather than acting purely as a blockade (Hu, S. et al. 2013).

5hmC further modulates binding. MeCP2 can bind both 5mC and 5hmC in neurons, with 5hmC enrichment at actively transcribed genes (Mellén, M. et al. 2012). UHRF2 also recognises 5hmC and was identified as a 5hmC reader; when co-expressed with TET1 in mESCs, 5hmC-rich regions undergo further oxidation within the TET pathway (Spruijt, C. G. et al. 2013). Interestingly, CEBPβ can bind to motifs containing 5mC, but its binding is selectively inhibited by 5hmC, whereas further oxidised bases (5fC/5caC) can restore or even enhance affinity (Khund Sayeed, S. et al. 2015).

Proteins with CXXC domains generally require unmethylated CpGs for binding. This class includes CFP1 (Shin Voo, K. et al. 2000), the CXXC -containing TET1 (Williams, K. et al. 2011), and the CXXC regulatory domain of DNMT1 (Pradhan,

M. et al. 2008). MLL is excluded from methylated CpG islands. Its CXXC domain specifically recognises uCpG sequences, and methylation disrupts this interaction and prevents promoter occupancy (Birke, M. 2002; Milne, T. A. et al. 2005a; Cierpicki, T. et al. 2010). Importantly, MLL occupancy also helps maintain these sites in a hypomethylated state. At the HOXA9 locus, for example, MLL binding has been proposed to protect CpG clusters from de novo methylation and sustains gene expression (Erfurth, F. E. et al. 2008).

This highlights a bidirectional coupling between methylation and factor binding: gains in 5mC reduce opportunities for MLL recruitment, while MLL binding itself helps preserve local hypomethylation. In this way, methylation patterns and MLL occupancy form a feedback loop that shapes the accessible subset of CpG islands in the genome (Zhu, H. et al. 2016; Héberlé, É. and Bardet, A. F. 2019).

6.1.3 Prediction of DNA binding proteins from methylation state

In recent years, there has been growing interest in using DNA methylation data to predict TF binding, complementing traditional motif or sequence based approaches. This stems from the well-established observation that methylation of cytosines within TF recognition motifs can directly inhibit binding for many factors, while unmethylated motifs are often required for occupancy (Yin, Y. et al. 2017). Because of this relationship, the local methylation landscape can provide informative context about potential TF accessibility and activity.

Several computational frameworks have been developed to integrate methylation profiles into TF binding prediction. For example, previous work intersecting predicted TF motifs with whole-genome bisulphite sequencing data significantly improves binding prediction across multiple cell types compared to motif data alone (Morgan, D. et al. 2024). Other tools, such as SEMplMe and MeDeMo, quantify how methylation at each position in a motif affects TF binding affinity, generating position-specific methylation sensitivity matrices (Nishizaki, S. S. and Boyle, A. P. 2022; Grau, J. et al. 2023). These studies reveal that the impact of methylation is highly TF-specific. While some TFs are strongly excluded by methylated cytosines, others tolerate or even prefer methylated sites, and in certain contexts, binding is unaffected. Methylation sensitivity can extend beyond the motif core to flanking regions, emphasizing the complexity of interpreting these signals (Luo, X. et al. 2021).

Most methylation aware models to date have focused exclusively on 5mC. However, several recent studies have begun explicitly incorporating 5hmC into predictive frameworks. For example, Deep5hmC constructs a deep learning model com-

bining DNA sequence and epigenetic context to predict genome-wide 5hmC enrichment across tissues, demonstrating that 5hmC patterns are sufficiently structured and reproducible to be accurately learned by ML models (Ma, X. et al. 2024). Similarly, 5hmC enrichment in both genic and nearby regulatory regions has been used to predict gene expression state across multiple cell types using neural networks (Gonzalez-Avalos, E. et al. 2024). These studies show that 5hmC is not merely a passive mark but can act as a predictive signal for regulatory activity, reinforcing its role as a functional component of the epigenome.

A few groups have extended motif models themselves to account for modified bases, including 5hmC, within TF binding sites. For example, expanding the standard DNA alphabet to model 5mC, 5hmC, and other oxidised bases, demonstrated that certain TFs exhibit differential affinity for motifs depending on which modified cytosine is present (Viner, C. et al. 2024). These approaches confirm that 5hmC can influence TF binding directly, though such models have typically been applied to gene expression or generic TFs rather than lineage-defining complexes such as MLL with the complexity of rearrangement in the case of MLL-AF4 causing aberrant DNA occupancy.

While these studies establish that 5hmC can be both predicted and predictive, few have systematically explored how 5mC and 5hmC together influence TF binding across the same genomic regions. This represents an important gap, as the interplay between these two modifications may reveal whether 5hmC acts independently of, or merely as a transition state within, the methylation cycle. Emerging technologies that simultaneously measure both modifications at base resolution (Füllgrabe, J. et al. 2023) now make such integrated modelling feasible, but computational frameworks that leverage these data are still in their infancy.

These questions are particularly relevant for MLLr leukaemia. MLL binds preferentially to unmethylated CpG-rich regions via its CXXC domain, and MLL fusion show altered binding patterns such as spreading into gene bodies (Kerry, J. et al. 2017). Because the recruitment of MLL complexes depends on the surrounding chromatin context, integrating both methylation and 5hmC data into predictive models offers a powerful way to determine whether these DNA modifications actively direct MLL targeting or simply reflect secondary chromatin changes.

In this chapter, I extend previous work by explicitly testing the predictive power of both 5mC and 5hmC for modelling MLL binding. By comparing MLL and MLL-AF4 fusion complexes, I aim to determine whether distinct methylation and 5hmC patterns underlie differential recruitment and to assess whether methylation data alone can accurately predict where MLL complexes bind.

6.1.4 Chapter Aims

The preceding sections outline how DNA methylation and 5hmC shape transcriptional regulation, with particular relevance to the MLL complex and its fusion derivatives. MLL preferentially occupies unmethylated CpG islands at promoters, whereas MLL fusion proteins are redirected to ectopic sites, often enhancers, through interactions with their fusion partners (Birke, M. 2002; Smith, A. et al. 2025a). These events occur within a chromatin landscape shaped by both 5mC and 5hmC, as well as by mutations in methylation regulators such as TET2 and DNMT3A.

A central unresolved question is whether DNA methylation actively impacts MLL binding, or whether it simply reflects a chromatin state established by other factors. Past work looking at non-methylated CpGs appears to be not predictive for MLL-AF4 binding (Kerry, J. et al. 2017), because many more CpG's are unmethylated than are bound by MLL-AF4. Here I revisit the relationship between both 5mC, and MLL binding including the MLL-AF4 fusion context, with the addition of 5hmC.

While global methylome changes in leukaemia have been well documented, there have been few direct comparisons of 5mC and 5hmC between MLL and fusion MLL contexts, and it remains unclear whether these modifications alone can predict MLL occupancy. To address this gap, here I ask three key questions:

1. What are the global and local patterns of 5mC and 5hmC in MLL (RCH-ACV) and MLL-AF4 (SEM) cells, particularly at CpG-rich promoters and regulatory regions targeted by the Twist Methylome capture panel?
2. How do methylation and 5hmC states differ between MLL and MLL-AF4 contexts, and how do these differences correlate with changes in MLL binding, especially at promoters and enhancers?
3. Can methylation and 5hmC data alone predict MLL recruitment, and do these marks play an instructive role in guiding MLL binding, or are they merely passive indicators of pre-existing chromatin states?

6.2 Methods

6.2.1 Samples and sequencing

Simultaneous methylation sequencing of 5mC and 5hmC was performed with targeted oligonucleotide hybridisation capture using the Twist Methylome panel (see Section 4.2.1 for panel composition; laboratory protocol in Chapter 2). Library preparation and sequencing followed the Biomodal six-letter sequencing workflow, and the resulting data were processed using the Duet pipeline (Füllgrabe, J. et al. 2023) (v1.4.1). This approach generates base-resolution measurements of cytosine, 5mC, and 5hmC across both DNA strands within targeted capture regions.

6.2.2 Differential binding and differential methylated regions

To determine differential binding in MLL-N between RCH-ACV and SEM cells, seqnado was used to process the CUT&Tag data as described in Chapter 3. This generated a merged consensus peaks set over both replicates for both cell lines which was used to count reads per peak using feature counts. DESeq2 (Love, M. I. et al. 2014) was then used on the counts table over 2 replicates to normalise the counts and apply Wald pairwise test to determine differentially bound peaks with adjusted p-value <0.05 and absolute log2 fold change >0. These differential peak regions were then extracted from the 5mC and 5hmC methylation calls with region mean methylation per CpG to be visualised as a heatmap using complex heatmap. Differential methylated regions for both 5mC and 5hmC were calculated using modality Differential Methylation Region (DMR) (Biomodal Ltd n.d.) (0.16.1) over promoter regions as well as all consensus MLL-N peaks of RCH-ACV and SEM cells.

6.2.3 Modelling MLL from methylation features

Dataset preparation

All analyses used the hg38 reference genome. Gene coordinates were derived from the GENCODE v44 basic annotation, as provided by the Modality software used for promoter tiling. A single representative TSS per gene was retained (canonical isoforms prioritised; ties resolved by transcript support level), and coordinates were oriented by transcript strand so that positions downstream of the TSS were positive. Promoter regions were tiled with fixed-span windows of 1,000 bp using a

predefined grid of TSS-relative centres. A dense core was tiled every 100 bp $\pm$1kb around the TSS, and symmetric flanks were sampled every 500 bp at $\pm$1,500, $\pm$2,000, $\pm$2,500, $\pm$3,000, $\pm$3,500 bp. The final ordered list of centres was the union of the dense and flank sets plus their sign-flipped counterparts. For each sliding window, per-CpG counts were strand-collapsed upstream and aggregated within the 1 kb span to produce window-level features. Methylation features were computed as coverage-weighted means of per-site fractions for 5mC and 5hmC a CpG-count covariate (number of CpGs in the window, from the reference) was added and treated as sample-invariant. Features were concatenated in genomic order so that column position encodes distance to the TSS. Windows without a defined aggregate value after filtering were left as NaN and then imputed as mean per feature on the full dataset using scikit-learn's SimpleImputer. For each promoter, MLL-N CUT&Tag signal was averaged over the same sliding-window span and log transformed then min-max scaled to [0,1] within replicate. The resulting design matrix contained 20,192 promoters and 372 predictors with four regression targets (two replicates per cell context)

Model specification and training

Gradient-boosted trees were fit using an XGBRegressor base estimator. Multi-target prediction was implemented via MultiOutputRegressor, yielding one booster per target trained on the same feature space. Hyper-parameters were selected by randomised search with grouped cross-validation. maximising uniform-average R^2 across outputs Cross-validation used 3 splits grouped by chromosome to prevent leakage across folds. To reduce computational load, 20 search iterations were conducted on a 10% random subset of the training promoters, preserving the chromosome groups within that subset. The best hyperparameters recovered by the search were then fixed for the final multi-output model, which was trained on the full training set (excluding the held-out test chromosome). Generalisation was assessed on a held-out chr9 test set, with all remaining promoters used for training/validation. Performance was reported per replicate using the (R^2), and MSE.

Model interpretation

Feature attributions were computed on the test set with SHAP TreeExplainer. For position-resolved summaries, SHAP values were converted to absolute values per promoter to quantify contribution magnitude, normalised so that promoter-wise contributions summed to 100%, and then averaged within fixed bins relative to the TSS

to obtain relative contribution profiles.

Software

All analyses were performed in Python and R. For specific software versions, see Appendix (Table B.6). Random seeds were set to 42.

6.3 Results

6.3.1 Simultaneous 5mC and 5hmC profiling

Cytosine modifications were profiled using Biomodal six-letter sequencing, which provides base-resolution discrimination of unmodified cytosine (C), 5-methylcytosine (5mC), and 5-hydroxymethylcytosine (5hmC) through a combination of selective oxidation, glucosylation, and deamination chemistries. As this is a relatively new technology, measurements were benchmarked against TAPS an established bisulphite-free method for methylation detection. Genome-wide sequencing of the six-letter methylation libraries was performed in single replicate for SEM and RCH-ACV cells, and matched samples were processed with TAPS were processed with SeqNado. Bigwig files analysed for both assays showed strong concordance in genome-wide methylation levels, and principal component analysis indicated that samples separated according to cell type rather than assay (Figure 6.2A, 6.2B). 5mC profiles at the FLT3-PAN3 locus (chr13:28,000,000-28,299,000) were also highly consistent between six-letter sequencing and TAPS (Figure 6.2C). These results confirm that six-letter sequencing yields methylation profiles comparable to TAPS, supporting its use in downstream analyses.

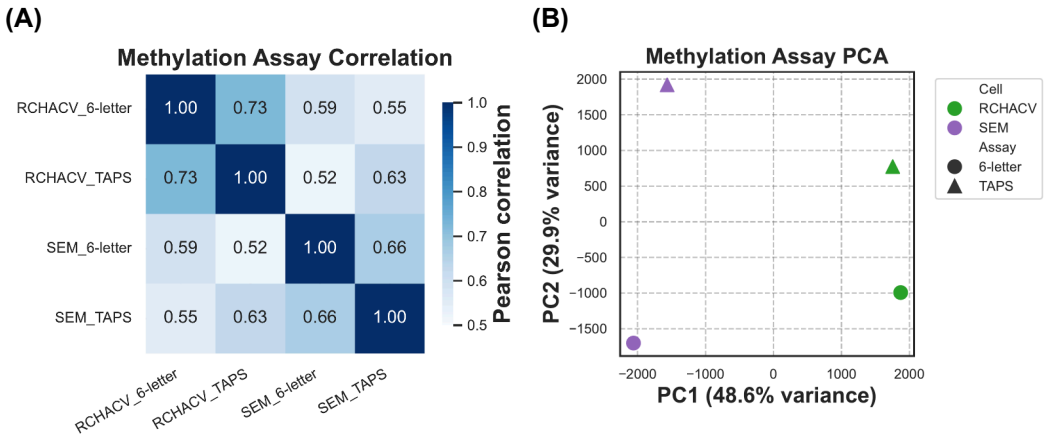


Figure 6.2: Continued next page.

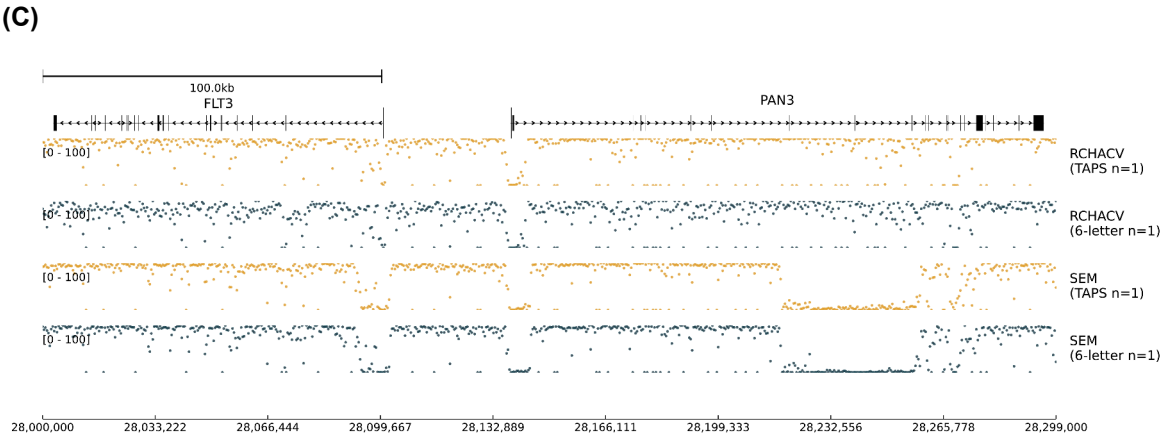


Figure 6.2: (Previous page) Comparison of six-letter sequencing and TAPS methylation profiles in matched samples. **(A)** Correlation of 5mC levels between six-letter sequencing and TAPS in matched samples. Pearson correlation coefficients (r) are indicated per sample. **(B)** Principal component analysis of six-letter sequencing and TAPS methylation profiles in matched samples. Samples cluster by cell type rather than assay, indicating concordant methylation patterns across methods. **(C)** Genome browser view of genome-wide, shallow 5mC profiles at the FLT3-PAN3 locus (chr13:28,000,000-28,299,000) in SEM and RCH-ACV cells, comparing six-letter sequencing (Green) and TAPS (Gold). Tracks show consistent methylation patterns across assays, with hypomethylated promoters and variable 5hmC enrichment.

Conversion accuracy was assessed using spike-in controls containing known modification states (Figure 6.3A). Sensitivity at synthetic oligos approached 95–100% for both 5mC and 5hmC, demonstrating excellent detection performance. Specificity was equally high: fully methylated λ phage DNA and the 5mC oligo generated $\sim$0% 5hmC calls, while unmethylated pUC19 and the 5hmC oligo yielded $\sim$0% 5mC calls. Cross-calling between 5mC and 5hmC was therefore negligible, with only minor variation observed between libraries.

Oligonucleotide hybridisation capture performance was robust across all libraries (Figure 6.3B). A high proportion of aligned reads mapped to the targeted Twist Methylome regions, with on-target rates consistently >80% across replicates, cell lines, and patient-derived samples. This indicates efficient and reproducible enrichment across experimental conditions.

Sequencing depth was well distributed across genomic features included in the panel (Figure 6.3C). Coverage was highest over CpG-dense regulatory elements, including CpG islands, promoters, and flanking regions, and extended to shores, shelves, open chromatin, transcription factor and CCCTC-binding factor (CTCF) binding sites, enhancer-like elements, and inter-CGI background regions. Most annotations achieved or exceeded a 30× coverage benchmark, ensuring reliable, site-

level quantification of both 5mC and 5hmC. Taken together, these results confirm that the six-letter sequencing assay generated high-quality, base-resolution methylation and 5hmC maps with excellent sensitivity, specificity, and uniform capture. These datasets provided a robust foundation for downstream comparative analyses of MLL versus MLL-AF4 contexts and for integration with MLL binding profiles described in later sections.

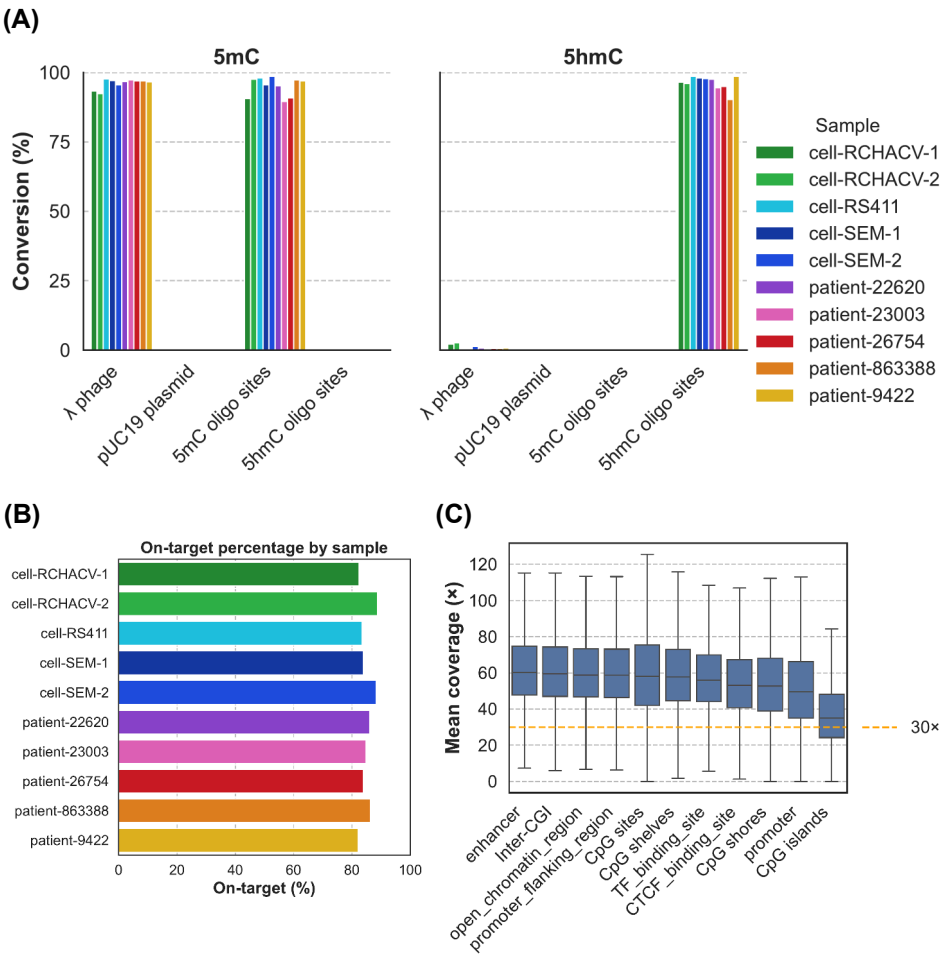


Figure 6.3: Overview of capture performance and methylation QC across samples. **(A)** Spike-in validation of 5mC conversion and 5hmC protection-oxidation (site-restricted).Bars report sensitivity on positives and specificity on negatives, with the oligo bars computed at CpGs of known state on the synthetic oligos ('5mC-oligo sites', '5hmC-oligo sites'). 5mC panel: λ phage and 5mC-oligo sites assess mC sensitivity; pUC19 and 5hmC-oligo sites assess mC specificity (hmC not miscalled as mC). 5hmC panel: 5hmC-oligo sites assess hmC sensitivity; λ phage, pUC19, and 5mC-oligo sites assess hmC specificity (no false-positive hmC). Sensitivities are $\sim$ 95–100% across samples with $\sim$ 0% on specificity controls, indicating accurate chemistry and calling. **(B)** On-target capture efficiency per sample, shown as the percentage of bases aligning to the targeted methylome panel. **(C)** Distribution of sequencing depth across genomic annotations represented in the Twist methylome panel (e.g., CpG islands/shores/shelves, promoters and flanking regions, open chromatin, TF/CTCF binding sites, enhancers, inter-CGI). A 30× benchmark is indicated in yellow.

6.3.2 Methylation Landscape and MLL Binding

Relationship between methylation and binding

To explore how DNA modifications relate to MLL occupancy, I first examined the average distribution of MLL-N signal and cytosine modifications around TSSs in RCH-ACV cells (MLL) than in SEM cells (MLL-AF4). A sharp, promoter-centred peak of MLL-N binding was observed precisely at the TSS (Figure 6.4A, upper panel). Both 5mC and 5hmC showed depletions relative to the flanking regions. In RCH-ACV cells, overall 5mC levels were lower than in SEM cells but displayed a similar trough centred on the TSS with elevated methylation in the surrounding regions (Figure 6.4A, middle panel).

Patterns of 5hmC differed between the cell lines. RCH-ACV showed higher overall 5hmC with a pronounced dip at the TSS, mirroring the 5mC pattern. In contrast, SEM cells exhibited substantially lower 5hmC levels overall, with only a shallow depletion at the TSS compared to the flanking regions (Figure 6.4A, lower panel).

To assess the global relationship between MLL-N and cytosine modifications, I computed Pearson correlations between MLL-N signal and both 5mC and 5hmC at promoters (±500 bp from the TSS) and enhancer-like regions (ATAC-defined peaks outside promoters) (Figures 6.4B, 6.4C). In both cell lines, MLL-N was negatively correlated with 5mC at promoters ($r \approx -0.58$ in SEM, -0.55 in RCH-ACV) and moderately so at enhancers ($r \approx -0.30$ in both), consistent with the known preference of MLL for unmethylated CpGs.

Correlations between 5hmC and MLL-N were weaker and more variable. At promoters, 5hmC correlated weakly negatively with MLL-N in RCH-ACV ($r \approx -0.26$) and SEM ($r \approx -0.16$), while enhancer regions showed almost no association ($r \approx -0.14$ and -0.07, respectively). Together, these results indicate that MLL-N binding is inversely associated with 5mC but shows no strong relationship with 5hmC, suggesting that hydroxymethylation occurs in nearby but not directly MLL-occupied regions.

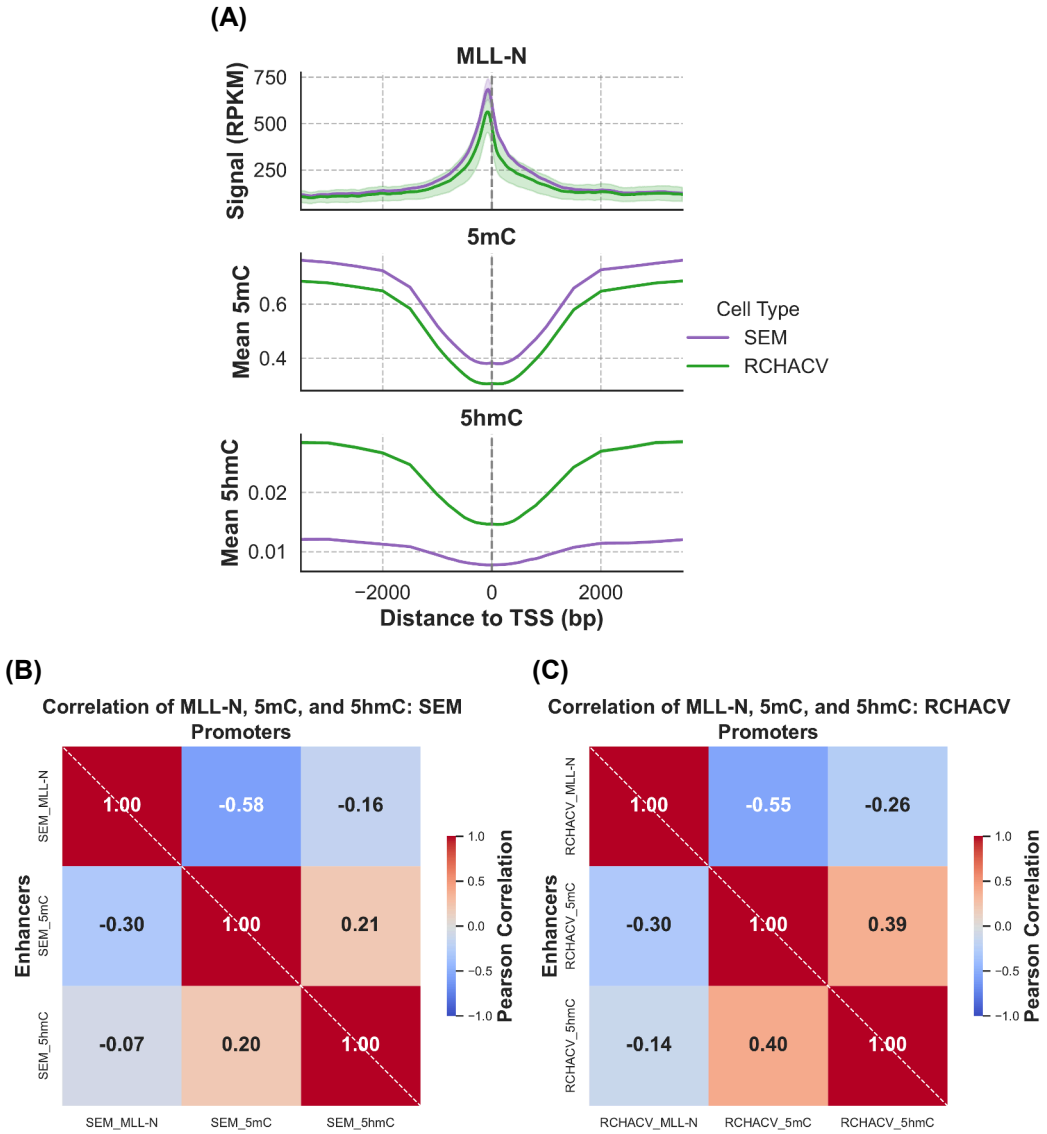


Figure 6.4: Relationship between DNA methylation and MLL-N binding. **(A)** Metagene profiles (±3 kb) around TSS showing MLL-N signal (top), mean 5mC (middle), and mean 5hmC (bottom) in SEM (purple) and RCH-ACV (green). **(B)** Correlation heatmap showing the relationship between mean 5mC, mean 5hmC, and MLL-N binding across promoters (Upper) and enhancer-like regions (Lower) defined by ATAC-seq and CUT&Tag in SEM cells. **(B)** Correlation heatmap showing the relationship between mean 5mC, mean 5hmC, and MLL-N binding across promoters (Upper) and enhancer-like regions (Lower) defined by ATAC-seq and CUT&Tag in RCH-ACV cells.

The relationship between MLL-N and 5mC were also evident at specific loci in patients with MLL-AF4 as well as in the RS4;11 cell line with MLL-AF4. However in some patients, such as 863388 and 9422, there was 5mC and still MLL-N binding, suggesting that this relationship is not absolute and other factors may also influence MLL-N recruitment. One possibility is sample heterogeneity: in bulk sequencing, apparent co-occurrence of 5mC and MLL-N binding may reflect averaging across

a mixed cell population, where methylation and occupancy arise from distinct cellular subpopulations rather than the same cell. Notably, this pattern was restricted to a subset of patients, which may reflect differences in the cellular composition of individual patient samples. For instance, at the *MEIS1* region, MLL-AF4 patient samples and cell lines showed prominent MLL-N peaks accompanied by local depressions in CpG methylation (Figure 6.5). Signal intensity varied across patients but was consistently higher than in RCH-ACV cells, which exhibited weaker MLL-N occupancy and relatively higher surrounding 5mC. These locus-level views reinforce the global analyses, illustrating how reduced methylation and loss of 5hmC accompany ectopic MLL-AF4 fusion binding.

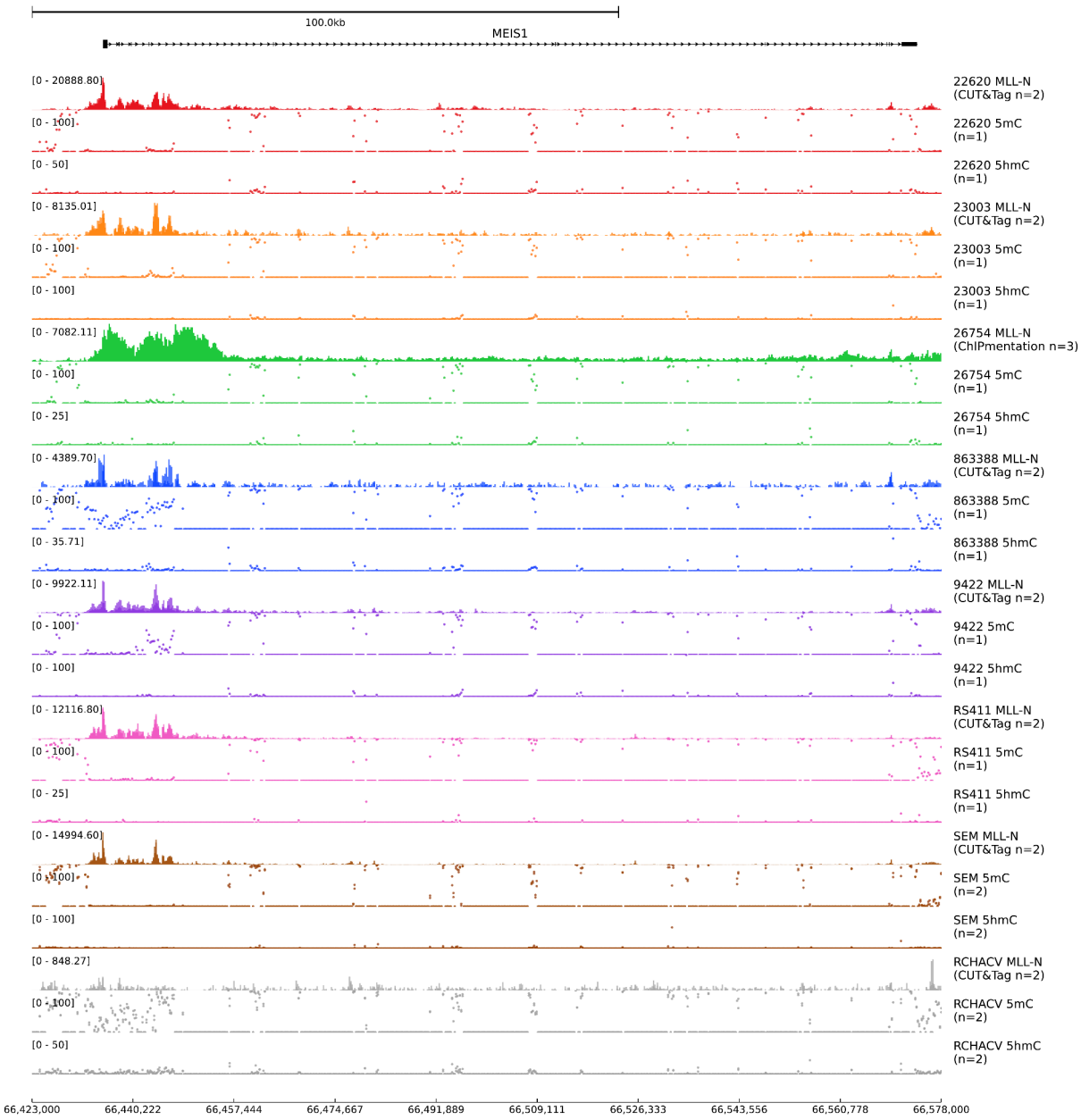


Figure 6.5: MLL-N binding and DNA methylation at the *MEIS1* locus. Genome browser snapshot (hg38) centred on *MEIS1* (chr2:66,423,000-66,578,000). For each sample, MLL-N CUT&Tag signal (RPKM), while 5mC and 5hmC are shown as methylation percentages (dots). Rows include five MLL-AF4 patient samples (22620, 23003, 26754, 863388, 9422), two MLL-AF4 cell lines (RS4;11, SEM), and the MLL cell line (RCH-ACV).

Differential methylation between MLL-AF4 and MLL contexts

In the metaplots for 5mC and 5hmC at promoters (Figure 6.4A), SEM cells showed overall higher 5mC and lower 5hmC compared with RCH-ACV. To determine whether these differences were statistically significant, I performed DMR analysis

for both 5mC and 5hmC at promoters comparing SEM and RCH-ACV. SEM cells had significantly higher 5mC at 10,151 promoters and significantly lower 5mC at only 2,189 promoters relative to RCH-ACV (Figures 6.6A, 6.6C). For 5hmC, the opposite trend was observed, with only 319 promoters showing higher 5hmC and 4,213 showing lower 5hmC in SEM compared with RCH-ACV (Figures 6.6B, 6.6C). The majority of promoters did not differ significantly in 5hmC levels. These results are consistent with the metaplots, which indicated higher 5mC and lower 5hmC at promoters in MLL-AF4 (SEM) relative to MLL (RCH-ACV).

To test whether this effect was promoter-specific or reflected a broader trend at MLL binding sites, I next performed DMR analysis over all consensus MLL-N peaks called in both SEM and RCH-ACV. Again, 5mC was elevated in SEM compared with RCH-ACV, and 5hmC was reduced. Among the 39,888 consensus peaks, 18,885 showed higher 5mC and 7,118 showed lower 5mC in SEM, while only 381 peaks had higher 5hmC and 3,456 had lower 5hmC in SEM compared with RCH-ACV (Figures 6.6D, 6.6E, 6.6F).

To examine how methylation related to MLL-N occupancy, I carried out differential binding analysis and visualised the significantly differentially bound peaks ($p < 0.05$) in a heatmap annotated with mean 5mC and 5hmC values (Figure 6.6G). Peaks with higher MLL-N in RCH-ACV (upper cluster) corresponded to lower 5mC and 5hmC in SEM, while peaks more strongly bound in SEM (lower cluster) had lower 5mC and 5hmC in RCH-ACV.

These global trends were also evident at specific loci. For example, at the *MEIS1* promoter, which is a target of MLL-AF4 binding, the promoter had significant differential 5mC and 5hmC between RCH-ACV and SEM that corresponded with significant differential MLL-N binding (Figure 6.6H).

Together, these results confirm that MLL preferentially binds hypomethylated regions, which also tend to have lower mean 5hmC. However, in the fusion context (MLL-AF4), there was overall higher 5mC and lower 5hmC compared with the non-fusion MLL context, suggesting reduced dependency on uCpGs, possibly through interactions with elongation-associated factors such as the SEC mediated by the fusion partner. Alternatively, some differences may reflect distinct cell-of-origin methylation patterns between these leukaemia cell lines.

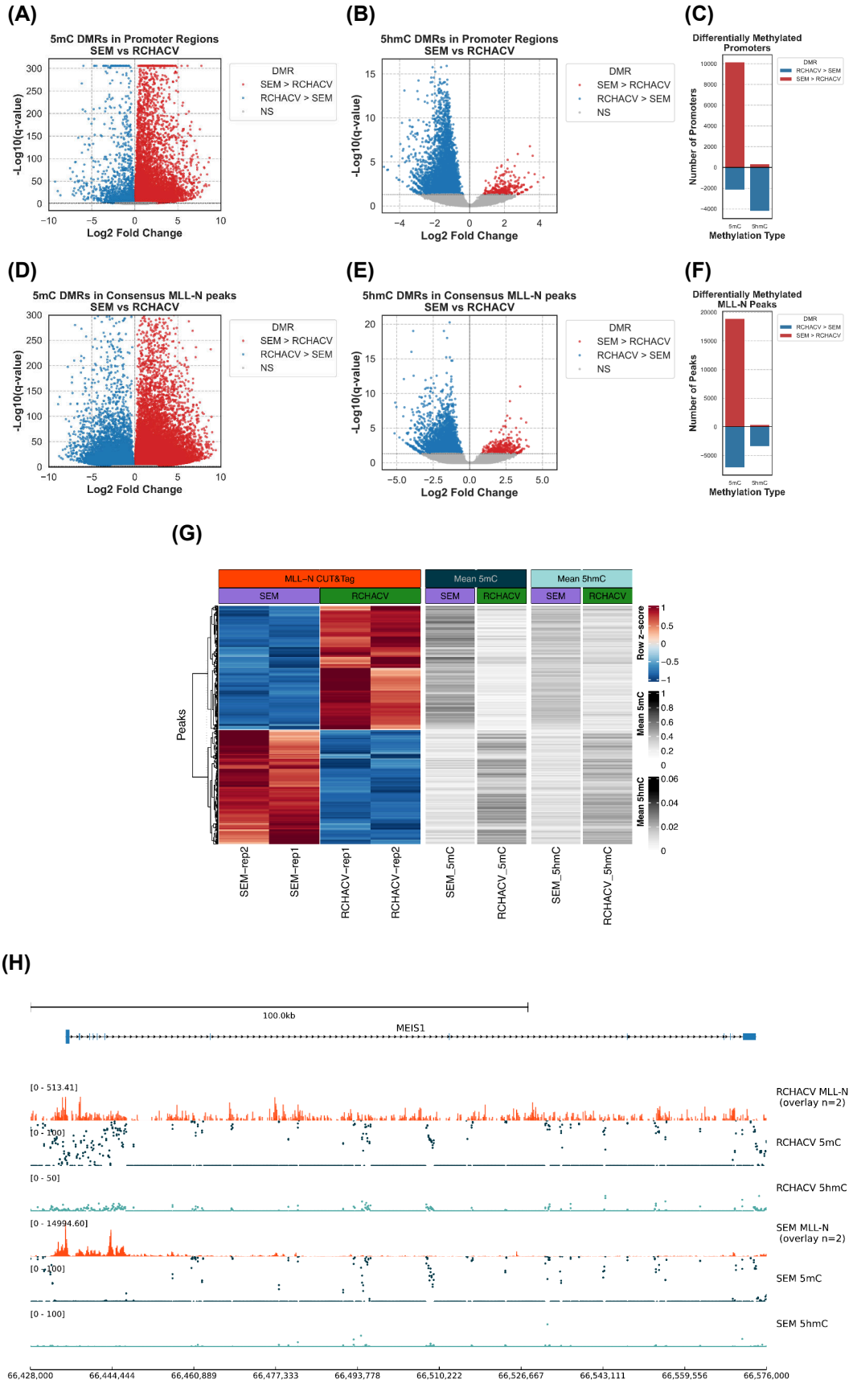


Figure 6.6: Continued next page.

Figure 6.6: Differential methylation between MLL-AF4 and MLL contexts. **(A-B)** Volcano plots of DMRs within promoters (±500 bp) for 5mC **(A)** and 5hmC **(B)**. **(C)** Barplot summarising the number of significant DMRs (adjusted p-value < 0.05) at promoters and for 5mC and 5hmC, showing a predominance of promoter hypermethylation (SEM > RCH-ACV) and 5hmC loss (RCH-ACV > SEM) in the MLL-AF4 context. **(D-E)** Volcano plots of DMRs within consensus MLL-N peaks for 5mC **(C)** and 5hmC **(D)**. **(F)** Barplot summarising the number of significant DMRs (adjusted p-value < 0.05) at consensus MLL-N peaks for 5mC and 5hmC, showing a predominance of hypermethylation (SEM > RCH-ACV) and 5hmC loss (RCH-ACV > SEM) in the MLL-AF4 context. **(G)** Heatmap of differential CUT&Tag MLL-N peaks (row z-scores) alongside mean 5mC and 5hmC per peak, for SEM and RCH-ACV. **(H)** MLL-N binding and DNA methylation at the *MEIS1* locus. Genome tracks (hg38) centred on *MEIS1* (chr2:66,423,000-66,578,000). For each (RCH-ACV; green) and MLL-AF4 (SEM; purple), CUT&Tag tracks display MLL-N RPKM signal, while mC and hmC tracks show CpG 5mC and 5hmC fractions, respectively. The inverse relationship between MLL-N binding and 5mC and 5hmC is evident at the promoter of *MEIS1* with lower MLL-N occupancy in RCH-ACV corresponding to higher 5mC and 5hmC, while in SEM, higher MLL-N binding corresponds to lower 5mC and 5hmC.

6.3.3 Predicting MLL Binding from Methylation State

In both SEM (MLL-AF4) and RCH-ACV (MLL), MLL occupancy anticorrelated with 5mC, whereas its association with 5hmC was weaker. Despite SEM's higher 5mC and lower 5hmC than RCH-ACV, the directionality between MLL and 5mC and 5hmC was conserved, prompting ML tests of how much these marks predict MLL binding across cell lines. To determine whether local DNA modification patterns alone are sufficient to predict MLL occupancy, I tiled genomic regions around annotated transcription start sites (±3 kb). Each promoter was divided into a dense central window flanked by progressively wider upstream and downstream bins to capture both the nucleosome-depleted region (NDR) and the first three nucleosomal positions. For each tile, I calculated features from the methylome capture data: strand-combined 5mC and 5hmC fractions, along with local CpG density. With this dataset, I trained a multi-target XGBoost multi-output regressor to predict MLL-N signal (scaled to 0-1) simultaneously for both cell types (RCH-ACV and SEM) and all biological replicates. This design allowed the model to share structure across outputs while preserving replicate-specific variation. Model performance was evaluated on a held-out chromosome (chr9) excluded from training and hyperparameter tuning. I quantified performance using R^2 and MSE, visualising the predicted versus observed MLL-N signal for each replicate.

To assess the predictive power of 5mC and 5hmC for MLL-N binding, I used this multi-output regression framework with three feature types: (1) CpG density, (2) 5mC, and (3) 5hmC. To determine the contribution of each feature type, I performed

ablation experiments in which models were trained with different subsets of features: (1) CpG density only. (2) 5mC only, (3) 5hmC only, (4) both 5mC and 5hmC, (5) all three features combined (Figure 6.7A). From this analysis, CpG density alone was the weakest predictor of MLL-N occupancy ($R^2 \approx 0.547$ in RCH-ACV and 0.495 in SEM). In contrast, 5mC alone was the strongest single feature ($R^2 \approx 0.715$ (RCH-ACV), 0.705 (SEM)), while 5hmC alone was moderate ($R^2 \approx 0.661$ (RCH-ACV), 0.647 (SEM)). Adding 5hmC to 5mC yielded a small but consistent gain ($\Delta R^2 \approx 0.002$ (RCH-ACV), 0.004 (SEM)). The best performance was obtained when CpG density was included alongside both 5mC and 5hmC ($R^2 = 0.737$ in RCH-ACV and 0.729 in SEM), and across all feature sets RCH-ACV was slightly higher than SEM. These results indicate that local methylation patterns at promoters, particularly 5mC, are highly informative for predicting MLL binding, with 5hmC providing additional complementary information.

Predictions closely tracked observed values in all outputs (Figure 6.7B). On the held-out chromosome, R^2 values ranged from 0.73 to 0.74, with MSE between 0.010 and 0.011, indicating that local methylation features captured the majority of between-region variance in MLL occupancy.

These results demonstrate that methylation and 5hmC patterns around promoters are highly predictive of MLL binding, even without including sequence or accessibility features such as ATAC-seq or H3K27ac. This establishes DNA modification state as a strong determinant of MLL recruitment.

Positional DNA methylation and 5hmC at TSSs

Promoter architecture was considered as a nucleosome-depleted region (NDR; -150 to +50 bp) flanked by the +1, +2 and +3 nucleosomes. To localise where the regression model relied on each feature group, SHAP values were computed on held-out data and summarised as mean absolute SHAP per position. For each promoter, SHAP values were absoluted and normalised to sum to 100%, then averaged in fixed bins relative to the TSS. Consequently, the resulting curves quantify contribution magnitude rather than effect direction (Figure 6.7C).

Using this magnitude metric, distinct spatial patterns were observed. For 5hmC, the maximal contribution in the MLL context (RCH-ACV) localised to the +2 nucleosome, whereas in the MLL-AF4 context (SEM) the principal 5hmC contribution shifted downstream towards +3. This distribution is compatible with a role for promoter-proximal 5hmC in supporting a more accessible chromatin environment immediately downstream of +1 in the MLL setting, with a downstream displacement

in the fusion setting suggestive of altered nucleosome phasing and/or redistribution of 5hmC activity. These mechanistic interpretations remain hypotheses pending direct experimental testing.

For 5mC, the dominant contribution was concentrated within the +1 nucleosome window. In complementary signed analyses, lower 5mC at +1 associated with higher predicted MLL-N signal, indicating that promoter-proximal methylation functions as a strong negative predictor of MLL occupancy. This observation accords with established coupling between nucleosome occupancy, DNMT activity and CpG methylation at promoter-proximal positions.

The CpG-count feature contributed minimally within the NDR and increased across the +2/+3 windows, indicating a sequence-encoded density effect that appears largely independent of cytosine modification state and is consistent with CpG-rich DNA supporting more stable downstream nucleosome positioning.

Taken together, these position-resolved reliance profiles support a working model in which firstly the +1-region 5mC provides the strongest repressive predictor with respect to MLL-N binding. Secondly, 5hmC contributes downstream in a context-dependent manner centred on +2 in MLL and shifted towards +3 in MLL-AF4 and thirdly, CpG density modulates the baseline chromatin organisation upon which methylation dynamics act. This model generates the hypothesis that perturbation of TET activity would reduce the downstream 5hmC contribution and sharpen +2/+3 nucleosome features, whereas inhibition of DNMT activity would diminish the +1-region 5mC contribution and increase predicted MLL-N occupancy at previously methylated promoters.

To determine the directionality of each feature's effect, I computed the Spearman correlation (ρ) between feature values and SHAP values across all positions (±3.5 kb from the TSS) (Figure 6.7D). Negative correlations indicate that higher feature values reduce the model's predicted MLL binding probability, while positive correlations indicate the opposite. In all replicates and both cell lines, 5mC (dark teal) showed a strong negative association centred on the TSS, indicating that higher 5mC strongly reduces predicted MLL-N binding, consistent with MLL's known preference for unmethylated CpG-rich promoters. 5hmC (light teal) showed a weaker but similar negative trend, suggesting that higher 5hmC also modestly reduces predicted MLL-N binding. In contrast, CpG density (orange) showed a positive correlation at the TSS, indicating that higher CpG density increases predicted MLL-N binding. These directional analyses confirm that 5mC is the dominant negative predictor of MLL-N binding, with 5hmC also contributing negatively but to a lesser extent, while CpG density positively influences MLL-N recruitment.

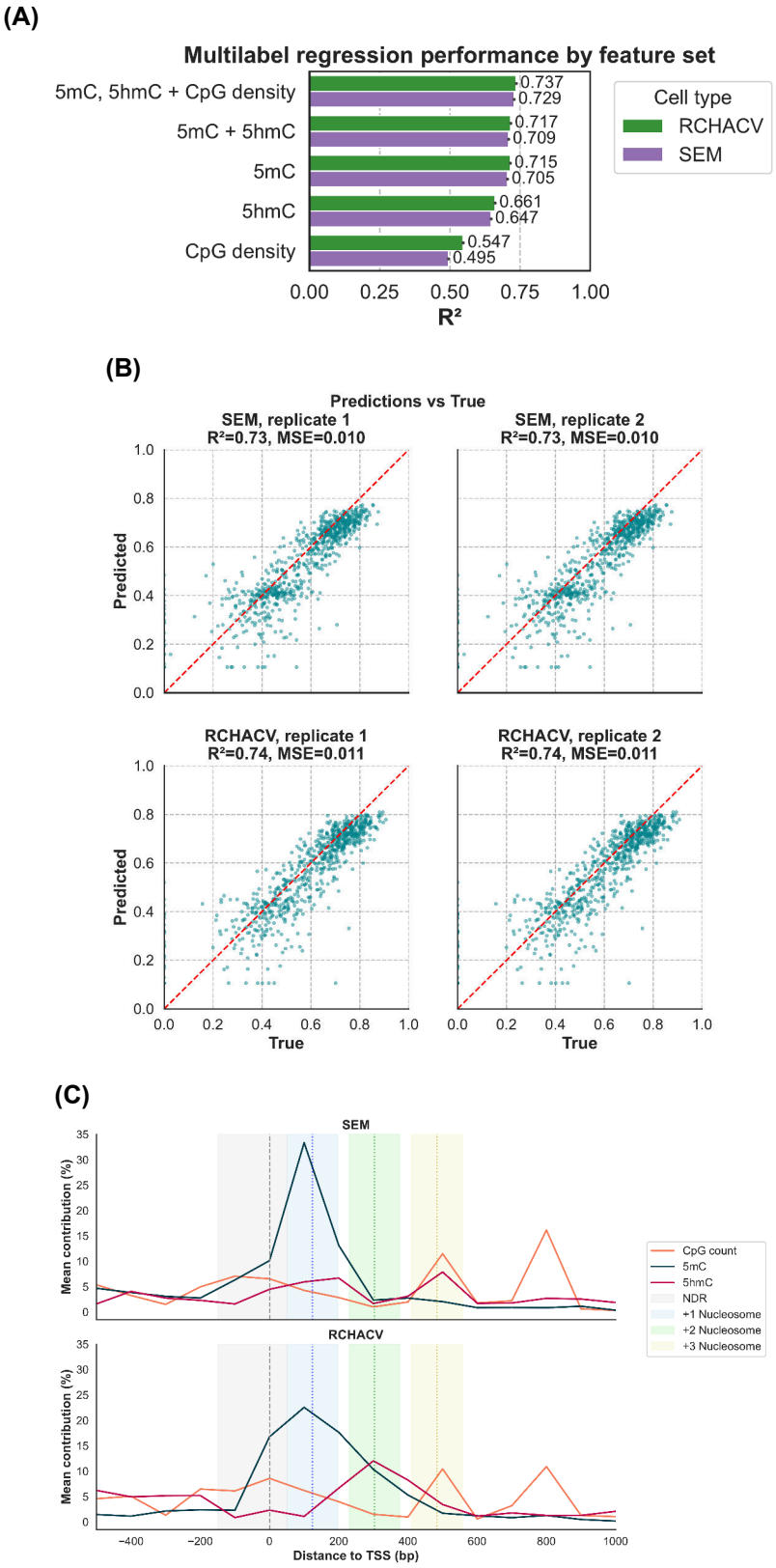


Figure 6.7: Continued next page.

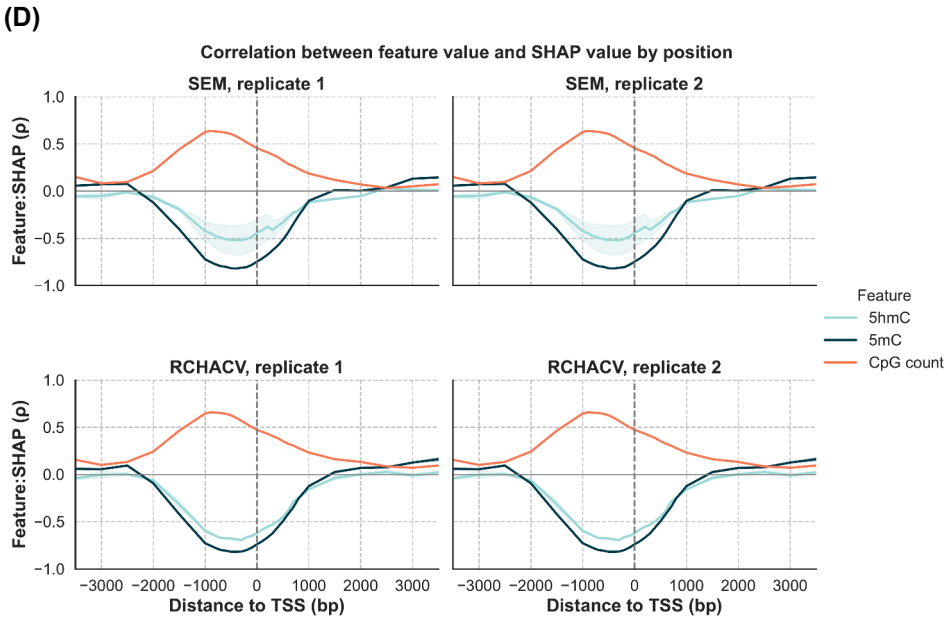


Figure 6.7: (Previous page) **(A)** Feature ablation results for predicting MLL-N binding from methylation features. Bar plot showing R^2 performance on held-out chr9 for models trained with different feature subsets: CpG density only, 5mC only, 5hmC only, both 5mC and 5hmC, and all three features combined. Results are shown separately for each replicate in RCH-ACV (Green) and SEM (Purple), Error bars indicate standard deviation across replicates. **(B)** XGBoost multi-target regression of MLL-N signal. Predicted versus observed scaled signal (0-1) on a held-out chr9 for each output (SEM and RCH-ACV, two replicates each). The multi-target model was trained jointly on all replicates and evaluated per replicate. Red dashed line, identity. Performance was consistent across outputs ($R^2 \approx$ 0.73-0.74; MSE $\approx$ 0.010-0.011), indicating robust generalisation from methylation features alone. **(C)** Position-resolved SHAP profiles of methylation features to MLL-N predictions around TSSs . Mean absolute SHAP contributions (expressed as % of total per promoter) plotted relative to the TSS for CpG count (orange), 5mC (teal), and 5hmC (magenta). SEM (top) and RCH-ACV (bottom) are shown separately. Shaded regions denote the NDR (-150 to +50 bp) and the +1, +2, and +3 nucleosomes, with dashed vertical lines indicating approximate nucleosome dyads. **(D)** Correlation between methylation feature values and SHAP values across promoter positions (±3.5 kb from the TSS). Each line shows the Spearman correlation (ρ) between a given feature's value and its SHAP contribution to predicted MLL-N binding. Negative correlations indicate that higher methylation reduces the model's predicted MLL binding probability. In all replicates and both cell lines, 5mC (dark teal) shows a strong negative association centred on the TSS, whereas 5hmC (light teal) shows a weaker but similar trend. CpG count (orange) shows a positive correlation with SHAP values, consistent with MLL's preference for unmethylated CpG-rich promoters.

6.3.4 TET2, MLL, and Promoter Methylation Dynamics

To investigate the relationship between MLL-N, TET2, and promoter methylation, I quantified MLL-N, TET2, 5mC, and 5hmC across 1 kb promoter windows in SEM cells. Promoter-averaged correlations revealed a very strong positive correlation between MLL-N and TET2 (r = 0.95). 5mC and 5hmC were similarly positively corre-

lated (r = 0.66). MLL-N was strongly negatively correlated with 5mC (r = –0.70) and moderately negatively correlated with 5hmC (r = –0.47). TET2 was also strongly negatively correlated with 5mC (r = –0.68) and negatively correlated with 5hmC (r = –0.45) (Figure 6.8A). These genome-wide summaries indicate that promoters co-occupied by MLL-N and TET2 are typically hypomethylated.

To resolve this relationship as a function of gene activity, I stratified promoters into expression quartiles (Q1-Q4) and plotted MLL-N against TET2, colouring points by methylation state (Figure 6.8B). Across all quartiles and most prominently in Q4 promoters with high TET2 co-displayed high MLL-N, while the corresponding colour scale showed low 5mC at these co-occupied sites. In contrast, low-occupancy promoters exhibited higher 5mC. When coloured by 5hmC, high-occupancy promoters tended to show lower 5hmC within the promoter windows, consistent with the global negative correlation and with depletion of 5hmC at the promoter core.

Positional analyses centred on TSSs further clarified these patterns (Figure 6.8C). In highly expressed genes, MLL-N and TET2 formed sharp, coincident peaks over the nucleosome-depleted region, whereas 5mC displayed a trough at the TSS that deepened with increasing expression. 5hmC was lowest at the TSS but increased across flanking nucleosomes, consistent with oxidation occurring around, rather than within, the NDR. Together, these profiles indicate that MLL-N/TET2 co-occupancy aligns with the locally hypomethylated promoter core of active genes. At *FLT3*, a highly expressed gene, MLL-N and TET2 show strong promoter occupancy coincident with depletion of 5mC and low 5hmC at the TSS (Figure 6.8D).

Overall, these results demonstrate that MLL-N and TET2 are tightly co-localised at promoters and that their co-occupancy is associated with promoter hypomethylation. The quartile-stratified scatter, TSS-centred metaplots, and representative loci together support a model in which a demethylated promoter environment consistent with TET2 activity is permissive for MLL-N binding.

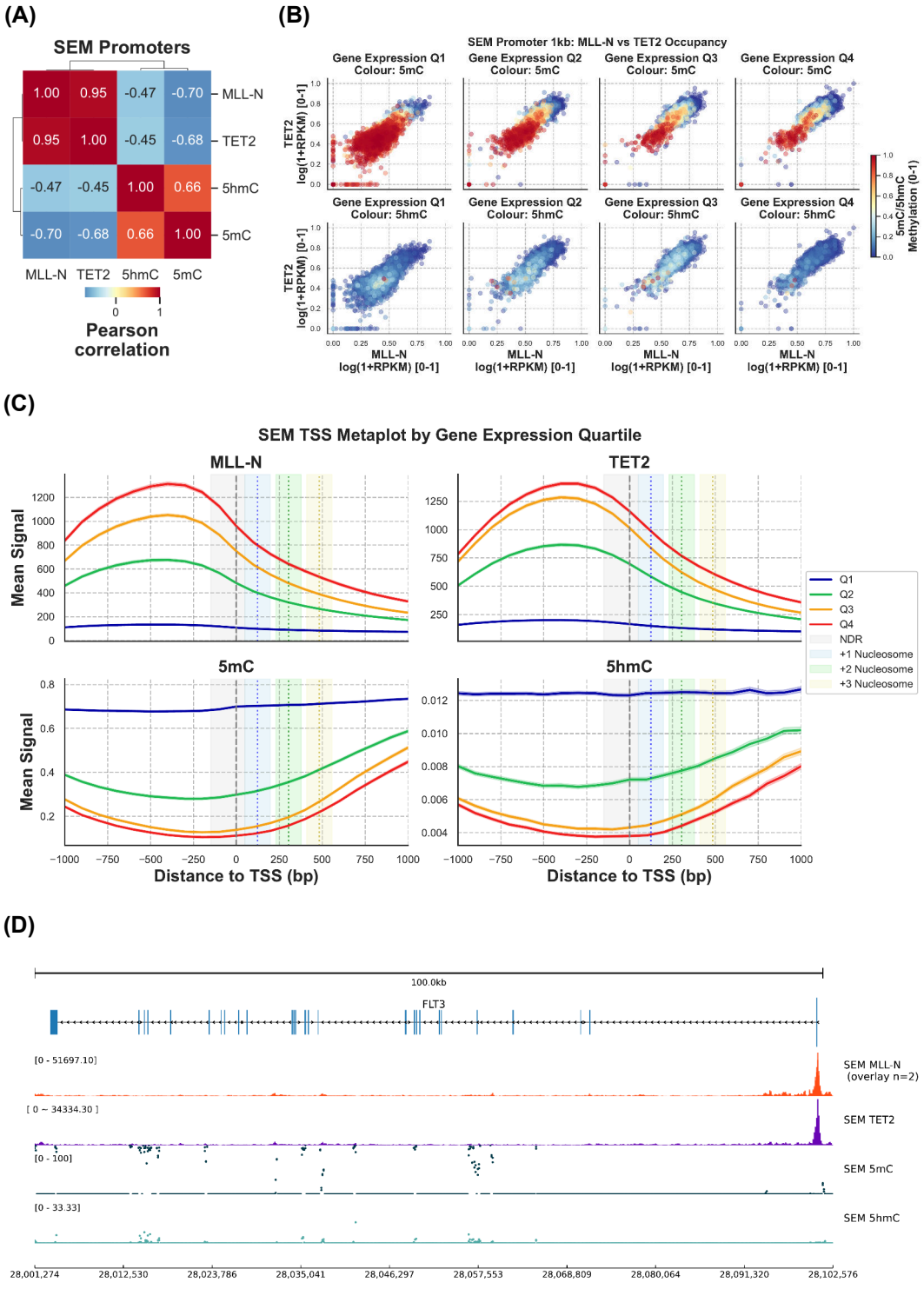


Figure 6.8: Continued next page.

Figure 6.8: (Previous page) Interplay between TET2, MLL-N, and promoter methylation in SEM cells. **(A)** Pearson correlation heatmap showing relationships between MLL-N, TET2, 5mC, and 5hmC across 1 kb promoter regions. MLL-N is positively correlated with TET2 and negatively correlated with 5mC, while TET2 is positively correlated with 5hmC and negatively correlated with 5mC, indicating coordinated occupancy with opposing relationships to methylation. **(B)** Scatter plots of MLL-N versus TET2 binding stratified by gene expression quartile (Q1 = lowest expression, Q4 = highest). Points are coloured by 5mC (top row) or 5hmC (bottom row). High TET2 and high MLL-N binding consistently co-occur with low 5mC and elevated 5hmC, particularly at highly expressed promoters. **(C)** TSS-centred metaplots of MLL-N, TET2, 5mC, and 5hmC signal by expression quartile. Highly expressed genes show sharp peaks of MLL-N and TET2 at the transcription start site, a nucleosome-depleted region (NDR), flanked by elevated 5hmC and depleted 5mC. **(D)** Genome browser snapshot of the *FLT3* locus (chr13:28,001,274-28,102,576) illustrating high MLL-N and TET2 co-occupancy at the promoter with corresponding low of 5mC and 5hmC.

6.4 Discussion

This study provides the first integrated view of 5mC and 5hmC landscapes in MLL and MLLr leukaemia, linking these epigenetic states to MLL complex binding patterns. By leveraging six-letter sequencing, I achieved base-resolution discrimination of 5mC and 5hmC across a targeted capture panel, enabling the exploration of methylation and MLL in both MLL and MLL-AF4 cell lines, as well as primary MLL-AF4 patient samples.

At promoters, MLL binding was correlated with hypomethylation, supporting the canonical model in which the MLL CXXC domain targets unmethylated CpG islands (Birke, M. 2002; Milne, T. A. et al. 2002). Low hydroxymethylation was also observed at MLL binding sites, however in SEM with MLL-AF4 fusion, 5hmC levels were reduced overall compared to RCH-ACV with MLL. However, in some patients, promoter 5mC was retained at MLL targets, suggesting that fusion MLL can bind methylated sites, potentially through altered recruitment mechanisms or co-factors.

I observed striking differences in both global and local DNA methylation profiles between MLL and fusion MLL contexts. SEM cells with MLL-AF4 displayed higher 5mC and reduced 5hmC compared with RCH-ACV with MLL, this was apparent in both promoter regions as well as all potential MLL binding sites across SEM and RCH-ACV cell lines. This difference may be due to differences in methylation stemming from the cell of origin. Although with the decreases 5hmC in SEM, this may also reflect reduced active demethylation possibly through reduced TET2 activity.

However, in SEM cells with MLL-AF4, I observed a high correlation between MLL-N and TET2 binding at promoters, with co-occupied sites showing low 5mC and 5hmC especially at the promoters of highly transcribed genes.

Predicting MLL binding from methylation state

Using ML, I demonstrated that MLL binding patterns at gene promoters can be predicted with high accuracy from methylation features alone. The XGBoost regression achieved R^2 values of $\approx$0.73-0.74 on held-out chromosomes, indicating that most of the variation in MLL occupancy is predictable from the methylation landscape. This provides strong evidence that DNA modification states are not merely correlative but contain sufficient information to infer binding directly.

Feature attribution analysis revealed a position-resolved model of methylation control at promoters. The strongest predictor of MLL binding was 5mC at the +1 nu-

cleosome, aligning with the established role of promoter-proximal methylation in blocking factor access. In contrast, 5hmC contributed downstream, centred on the +2 nucleosome in MLL cells but shifted to the +3 nucleosome in SEM. This spatial redistribution suggests altered nucleosome phasing or 5hmC targeting in the fusion context. These findings echo *in vitro* studies showing that 5hmC influences nucleosome conformation and stability (Mendonca, A. et al. 2014).

The successful prediction of MLL binding solely from methylation data parallels recent efforts to use 5hmC and 5mC to predict other regulatory outcomes, such as gene expression (Gonzalez-Avalos, E. et al. 2024). However, this work uniquely integrates both 5mC and 5hmC modifications at base resolution and applies them to a lineage-defining factor with known structural specificity for unmethylated CpG DNA. This represents a step toward models that can infer TF recruitment directly from the DNA modification landscape.

Mechanistic implications and future directions

These results suggest a model in which MLL recruitment is stabilised by promoter-proximal hypomethylation and supported by flanking 5hmC, whereas fusion proteins exploit alternative, less canonical sites as these features are lost or redistributed. This framework predicts that restoring TET activity or reducing DNMT-mediated methylation could shift the balance of MLL occupancy, potentially re-establishing normal promoter targeting.

Future work should test these predictions experimentally by manipulating TET proteins to ascertain if altering demethylation through TET activity impacts MLL binding patterns.

Additionally, applying this modelling framework to the patient samples and normal haematopoietic populations will clarify whether the patterns observed here generalise or are specific to these particular cell lines.

Deeper genome-wide sequencing of the methylation assay would allow for exploration of this approach in all genomic contexts, not just at promoters.

Conclusions

This chapter demonstrates that the methylation landscape alone provides a rich and predictive readout of MLL binding. The integration of 5mC and 5hmC highlights their complementary roles with 5mC as a restrictive barrier and 5hmC as a permissive signal. By linking these patterns to MLL versus fusion MLL recruitment, I provide

new insight into how epigenetic modifications influence transcriptional control in leukaemia and establish a foundation for future mechanistic studies.

7 Discussion

In MLLr leukaemias, fusion proteins bind CpG-rich DNA including at gene promoters, driving aberrant gene expression programmes through increased transcriptional activation. Although MLL binds unmethylated CpG-island promoters through its CXXC domain, only a subset of potential sites are occupied, and the rules governing this selectivity have remained unclear. This thesis set out to determine why MLL binds some CpG islands and not others, and how MLL-AF4 fusions alter this recruitment to engage distal enhancers. To address this, I focused on three fundamental regulatory layers; DNA sequence, Chromatin cofactors and context, and Cytosine methylation. By applying modular, interpretable ML models to each layer, I aimed to define their independent contributions before integrating them into a unified framework for MLL recruitment logic.

7.0.1 DNA Sequence as a Baseline Layer

Sequence models confirmed CpG density as a fundamental determinant of MLL binding, consistent with direct recognition of unmethylated CpG islands by the CXXC domain (Birke, M. 2002; Milne, T. A. et al. 2002). However, CpG content alone was insufficient: many CpG-rich promoters remained unbound, implying that additional sequence features refine this baseline.

Token-level attributions showed that in non-fusion contexts, CG-rich tokens were strongly positively associated with MLL binding. In fusion-positive lines this association was attenuated, and in MV4-11 it inverted, with a negative correlation between token-level CG percentage and MLL signal.

At the motif level, a baseline importance for CG-rich motifs was evident. Positively weighted motifs included VEZF1, KLFs, and MAZ, whereas MYC and KLF6 scored negatively consistent with competitive or restrictive effects. This pattern accords with motif grammar models from high-resolution sequence-to-profile frameworks (e.g., BPNet), where cooperative and competitive motif environments shape occu-

pancy (Avsec, Ž. et al. 2019).

In MLL-AF4 cells, the CpG baseline was diminished: binding was less tied to CpG-island context and more influenced by alternative motif environments. Negative attributions frequently mapped to lineage-restricted factors such as RUNX and SP family members, suggesting that fusion complexes exploit these contexts to engage distal regulatory elements. The shift from CpG-rich motifs in wild-type MLL to more heterogeneous sequence contexts in MLL-AF4 cells may reflect the documented fusion-context spreading phenotype. As MLL-AF4 complexes extend beyond CpG-rich promoters into gene bodies and distal enhancer-associated chromatin, the bound sequences become less CpG-dense, reducing the predictive weight of CG motifs (Kerry, J. et al. 2017; Smith, A. et al. 2025a).

Together, these results support a model in which MLL requires both CpG recognition and a favourable motif environment, whereas fusion proteins partially bypass CpG dependence and broaden the spectrum of potential binding sites potentially through interactions with components of the transcriptional machinery.

7.0.2 Chromatin Cofactors and Context

Cofactor modelling resolved two recruitment regimes. In the MLL setting, MLL associated most strongly with promoter-linked modules including Menin, LEDGF, and PAF1, together with H3K4me3 and H3K27ac. This is consistent with Menin acting as a scaffold and LEDGF stabilising promoter engagement of MLL (Yokoyama, A. and Cleary, M. L. 2008; Milne, T. A. et al. 2010).

By contrast, MLL-AF4 correlated with elongation machinery, including the SEC, BRD proteins, RNA polymerase II, and DOT1L with its characteristic H3K79 methylation signature aligning with models in which fusions hijack elongation to drive ectopic activation (Lin, C. et al. 2010; Deshpande, A. J. et al. 2013). The reduced dependency on Menin is consistent with recent findings showing that Menin stabilises MLL fusion protein binding and transcriptional activity only at a subset of target genes (Krivtsov, A. V. et al. 2019).

Together, these patterns indicate that while MLL relies on promoter-anchored cofactor networks to maintain stable occupancy, fusion complexes reprogramme cofactor dependencies toward elongation-linked modules, enabling retargeting to distal/ectopic sites and the formation of broad, transcriptionally active domains.

7.0.3 DNA Methylation as a Modulatory Layer

Analyses of cytosine methylation showed that MLL preferentially occupies unmethylated, CpG-rich promoters and is enriched at sites marked by 5-hydroxymethylcytosine (5hmC), consistent with TET-mediated demethylation stabilising promoter binding (Cimmino, L. et al. 2017; Füllgrabe, J. et al. 2023). In contrast, MLL-AF4 displayed reduced sensitivity to methylation state, aligning with its engagement of distal, transcriptionally active elements where CpG density and modification status are more variable.

When promoters were tiled and both 5mC and 5hmC were included as features, MLL binding could be predicted from cytosine methylation state. Attribution analyses highlighted 5hmC at the +2 nucleosome as a key positive signal for MLL, whereas this feature contributed little in the fusion context, indicating that 5hmC-linked regulation is retained for MLL but attenuated upon fusion. Notably, TET2 occupancy correlated strongly at promoters in fusion-positive cells, indicating frequent co-occupation; however, this did not translate into a comparable 5hmC importance for predicting fusion binding.

Together, these results suggest that MLL is constrained by the local 5mC/5hmC landscape via its CXXC -mediated preference for unmethylated CpGs, while MLL-AF4 partially escapes this regulation, diminishing the influence of active demethylation on recruitment.

7.0.4 Limitations and Future Directions

Despite the insights gained, several limitations should be acknowledged:

Sample scope and diversity. The analyses relied primarily on a small number of cell line samples, which may not fully represent the heterogeneity of MLLr leukaemias. Expansion to larger and more diverse cohorts will be essential to generalise these findings.

The Resolution of methylation data was limited. The Twist Methylome panel provided targeted coverage with a reasonable cost but limited genome-wide resolution, particularly at distal regulatory regions. Broader methylation with deep genome-wide datasets would capture rare or unanticipated patterns of 5hmC and 5mC more comprehensively.

Modular modelling approach. While separate models improved interpretability, they may have missed subtle interactions between layers. A fully integrated multimodal

model could capture cross-layer effects that remain unresolved.

Underrepresentation of active enhancers. While the methylome panel includes a substantial proportion of distal intergenic sequence (Figure 4.5), and CpG-rich enhancers are therefore likely represented, regions selected primarily for CpG island content may not systematically capture cell-line-specific active enhancers defined by accessibility rather than CpG density. TFs contributing to MLL recruitment may therefore act at distal enhancers in ways that are not fully resolved by the current panels.

Building on this work, several promising avenues emerge. Development of a unified multi-modal model that integrates sequence, cofactor, and methylation data could reveal higher-order interactions and improve insights into MLL recruitment logic.

Biological validation through experimental testing of predicted sequence motifs using base editing methods.

Follow-up work with TET protein perturbation studies to further understand the relationship between MLL binding and TET-mediated demethylation.

Extending cofactor modelling to accessibility-defined enhancer regions would test whether low TF importance in the current panels reflects panel composition or a genuine lack of discriminative signal. Cofactor attributions from the methylome model show lineage-restricted TFs including ELF1, RUNX family members, and PBX among detectable features but at relatively low importance, consistent with these factors being broadly bound at both MLL-occupied and unoccupied regions and therefore not highly discriminative for MLL recruitment. Practically, this requires separate per-cell-line models as enhancer landscapes differ substantially between cell lines (Smith, A. et al. 2025a).

7.0.5 Concluding Remarks

Although MLL binding to CpG islands via its CXXC domain is well established, the selective occupancy of only a subset of potential sites has remained enigmatic. This thesis shows that MLL binding specificity emerges from the interplay of sequence, cofactors, and methylation. MLL integrates CpG recognition, motif context, promoter-linked cofactors, and unmethylated or 5hmC-enriched landscapes to selectively bind developmental promoters.

My analyses suggest that selectivity reflects combinatorial sequence logic superimposed on CpG recognition. Sequence models (e.g., GROVER) point to CG-rich motif grammars that improve prediction of MLL occupancy, although the precise

motif interactions remain unresolved.

Complementary chromatin models emphasise promoter-proximal features associated with active transcription, with strong importance assigned to H3K27ac and related cofactor signals, whereas in the MLL-AF4 context the highest contributors shift towards elongation-linked factors, consistent with reduced reliance on CpG-centred rules and increased recruitment at distal enhancers.

Differential methylation analyses indicate higher promoter 5mC in SEM, potentially reflecting cell-of-origin differences, alongside globally reduced 5hmC despite correlated occupancy of MLL and TET2. Notably, the information from 5hmC at the +2 nucleosome emerges as a salient predictor, consistent with modulation of pause-release and early elongation rather than initiation alone.

Together, these results support a working model in which CpG recognition provides a baseline requirement, local motif grammar and promoter-linked cofactors refine targeting, and the 5mC/5hmC landscape finetunes elongation competence. MLL-AF4 perturbs this balance by hijacking elongation machinery, weakening CpG dependence, and broadening recruitment.

This framework does not fully explain why only a subset of CpG-rich promoters is bound, but it defines testable hypotheses regarding motif interactions, cofactor dependencies, and methylation modulation that can be explored in future work.

References

Alipanahi, B., A. Delong, M. T. Weirauch, and B. J. Frey (2015). “Predicting the Sequence Specificities of DNA- and RNA-binding Proteins by Deep Learning”. In: *Nature Biotechnology* 33.8, pp. 831–838. DOI: 10.1038/nbt.3300.

Allen, M. D., C. G. Grummitt, C. Hilcenko, S. Y. Min, L. M. Tonkin, C. M. Johnson, S. M. Freund, M. Bycroft, and A. J. Warren (2006). “Solution Structure of the Nonmethyl-CpG-binding CXXC Domain of the Leukaemia-Associated MLL Histone Methyltransferase”. In: *The EMBO Journal* 25.19, pp. 4503–4512. DOI: 10.1038/sj.emboj.7601340.

Amemiya, H. M., A. Kundaje, and A. P. Boyle (2019). “The ENCODE Blacklist: Identification of Problematic Regions of the Genome”. In: *Scientific Reports* 9.1, p. 9354. DOI: 10.1038/s41598-019-45839-z.

Andersson, A. K., J. Ma, J. Wang, X. Chen, A. L. Gedman, J. Dang, J. Nakitandwe, L. Holmfeldt, M. Parker, J. Easton, R. Huether, R. Kriwacki, M. Rusch, G. Wu, Y. Li, H. Mulder, S. Raimondi, S. Pounds, G. Kang, L. Shi, J. Becksfort, P. Gupta, D. Payne-Turner, B. Vadodaria, K. Boggs, D. Yergeau, J. Manne, G. Song, M. Edmonson, P. Nagahawatte, L. Wei, C. Cheng, D. Pei, R. Sutton, N. C. Venn, A. Chetcuti, A. Rush, D. Catchpoole, J. Heldrup, T. Fioretos, C. Lu, L. Ding, C.-H. Pui, S. Shurtleff, C. G. Mullighan, E. R. Mardis, R. K. Wilson, T. A. Gruber, J. Zhang, and J. R. Downing (2015). “The Landscape of Somatic Mutations in Infant MLL-rearranged Acute Lymphoblastic Leukemias”. In: *Nature Genetics* 47.4 (4), pp. 330–337. DOI: 10.1038/ng.3230.

Aranda-Orgilles, B., R. Saldaña-Meyer, E. Wang, E. Trompouki, A. Fassl, S. Lau, J. Mullenders, P. P. Rocha, R. Raviram, M. Guillamot, M. Sánchez-Díaz, K. Wang, C. Kayembe, N. Zhang, L. Amoasii, A. Choudhuri, J. A. Skok, M. Schober, D. Reinberg, P. Sicinski, H. Schrewe, A. Tsirigos, L. I. Zon, and I. Aifantis (2016). “MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis”. In: *Cell Stem Cell* 19.6, pp. 784–799. DOI: 10.1016/j.stem.2016.08.004.

Av Shrikumar, K. Tian, Annashcherbina, Žiga Avsec, Amr, C. McAnany, J. Kaczmarzyk, Pgreenside, Surag Nair, Mhfzsharmin, S. Holderbach, and R. Ma (2022). *Kundajelab/Tfmodisco: Bringing down Leiden Memory Use - Patch 1*. Version v0.5.16.2. Zenodo. DOI: 10.5281/ZENODO.5909083.

Avsec, Ž., V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley (2021). “Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions”. In: *Nature Methods* 18.10, pp. 1196–1203. DOI: 10.1038/s41592-021-01252-x.

Avsec, Ž., M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, and J. Zeitlinger (2019). *Base-Resolution Models of Transcription Factor Binding Reveal Soft Motif Syntax*. DOI: 10.1101/737981. URL: http://biorxiv.org/lookup/doi/10.1101/737981 (visited on 09/16/2025). Pre-published.

Ayton, P. M., E. H. Chen, and M. L. Cleary (2004). “Binding to Nonmethylated CpG DNA Is Essential for Target Recognition, Transactivation, and Myeloid Transformation by an MLL Oncoprotein”. In: *Molecular and Cellular Biology* 24.23, pp. 10470–10478. DOI: 10.1128/MCB.24.23.10470-10478.2004. PMID: 15542854.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble (2009). “MEME SUITE: Tools for Motif Discovery and Searching”. In: *Nucleic Acids Research* 37 (Web Server), W202–W208. DOI: 10.1093/nar/gkp335.

Bailey, T. L. and C. Elkan (1994). “Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers”. In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 2, pp. 28–36. PMID: 7584402.

Bannister, A. J. and T. Kouzarides (2011). “Regulation of Chromatin by Histone Modifications”. In: *Cell Research* 21.3, pp. 381–395. DOI: 10.1038/cr.2011.22.

Baubec, T., R. Ivánek, F. Lienert, and D. Schübeler (2013). “Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family”. In: *Cell* 153.2, pp. 480–492. DOI: 10.1016/j.cell.2013.03.011.

Bernstein, B. E., M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, T. R. Gingeras, S. L. Schreiber, and E. S. Lander (2005). “Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse”. In: *Cell* 120.2, pp. 169–181. DOI: 10.1016/j.cell.2005.01.001.

Bhagwat, A. S., J.-S. Roe, B. Y. Mok, A. F. Hohmann, J. Shi, and C. R. Vakoc (2016). “BET Bromodomain Inhibition Releases the Mediator Complex from Select Cis-Regulatory Elements”. In: *Cell Reports* 15.3, pp. 519–530. DOI: 10.1016/j.celrep.2016.03.054.

Bhardwaj, V., S. Heyne, K. Sikora, L. Rabbani, M. Rauer, F. Kilpert, A. S. Richter, D. P. Ryan, and T. Manke (2019). “snakePipes: Facilitating Flexible, Scalable and Integrative Epigenomic Analysis”. In: *Bioinformatics* 35.22. Ed. by B. Berger, pp. 4757–4759. DOI: 10.1093/bioinformatics/btz436.

Biomodal Ltd (n.d.). *Modality*. Version 0.16.1.

Bird, A. (2002). “DNA Methylation Patterns and Epigenetic Memory”. In: *Genes & Development* 16.1, pp. 6–21. DOI: 10.1101/gad.947102.

Bird, A. P. and E. M. Southern (1978). “Use of Restriction Enzymes to Study Eukaryotic DNA Methylation”. In: *Journal of Molecular Biology* 118.1, pp. 27–47. DOI: 10.1016/0022-2836(78)90242-5.

Birke, M. (2002). “The MT Domain of the Proto-Oncoprotein MLL Binds to CpG-containing DNA and Discriminates against Methylation”. In: *Nucleic Acids Research* 30.4, pp. 958–965. DOI: 10.1093/nar/30.4.958.

Blackledge, N. P., A. M. Farcas, T. Kondo, H. W. King, J. F. McGouran, L. L. Hanssen, S. Ito, S. Cooper, K. Kondo, Y. Koseki, T. Ishikura, H. K. Long, T. W. Sheahan, N. Brockdorff, B. M. Kessler, H. Koseki, and R. J. Klose (2014). “Variant PRC1 Complex-Dependent H2A Ubiquitylation Drives PRC2 Recruitment and Polycomb Domain Formation”. In: *Cell* 157.6, pp. 1445–1459. DOI: 10.1016/j.cell.2014.05.004.

Blayney, J. W., H. Francis, A. Rampasekova, B. Camellato, L. Mitchell, R. Stolper, L. Cornell, C. Babbs, J. D. Boeke, D. R. Higgs, and M. Kassouf (2023). “Super-Enhancers Include Classical Enhancers and Facilitators to Fully Activate Gene Expression”. In: *Cell* 186.26, 5826–5839.e18. DOI: 10.1016/j.cell.2023.11.030.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf (2013). “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling

of Open Chromatin, DNA-binding Proteins and Nucleosome Position". In: *Nature Methods* 10.12, pp. 1213–1218. DOI: 10.1038/nmeth.2688.

Cabal-Hierro, L., P. Van Galen, M. A. Prado, K. J. Higby, K. Togami, C. T. Mowery, J. A. Paulo, Y. Xie, P. Cejas, T. Furusawa, M. Bustin, H. W. Long, D. B. Sykes, S. P. Gygi, D. Finley, B. E. Bernstein, and A. A. Lane (2020). "Chromatin Accessibility Promotes Hematopoietic and Leukemia Stem Cell Activity". In: *Nature Communications* 11.1, p. 1406. DOI: 10.1038/s41467-020-15221-z.

Chahrour, M., S. Y. Jung, C. Shaw, X. Zhou, S. T. C. Wong, J. Qin, and H. Y. Zoghbi (2008). "MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription". In: *Science* 320.5880, pp. 1224–1229. DOI: 10.1126/science.1153252.

Cierpicki, T., L. E. Risner, J. Grembecka, S. M. Lukasik, R. Popovic, M. Omonkowska, D. D. Shultis, N. J. Zeleznik-Le, and J. H. Bushweller (2010). "Structure of the MLL CXXC Domain–DNA Complex and Its Functional Role in MLL-AF9 Leukemia". In: *Nature Structural & Molecular Biology* 17.1, pp. 62–68. DOI: 10.1038/nsmb.1714.

Cimmino, L., I. Dolgalev, Y. Wang, A. Yoshimi, G. H. Martin, J. Wang, V. Ng, B. Xia, M. T. Witkowski, M. Mitchell-Flack, I. Grillo, S. Bakogianni, D. Ndiaye-Lobry, M. T. Martín, M. Guillamot, R. S. Banh, M. Xu, M. E. Figueroa, R. A. Dickins, O. Abdel-Wahab, C. Y. Park, A. Tsirigos, B. G. Neel, and I. Aifantis (2017). "Restoration of TET2 Function Blocks Aberrant Self-Renewal and Leukemia Progression". In: *Cell* 170.6, 1079–1095.e20. DOI: 10.1016/j.cell.2017.07.032.

Clapier, C. R. and B. R. Cairns (2009). "The Biology of Chromatin Remodeling Complexes". In: *Annual Review of Biochemistry* 78.1, pp. 273–304. DOI: 10.1146/annurev.biochem.77.062706.153223.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen (2008). "Shotgun Bisulphite Sequencing of the Arabidopsis Genome Reveals DNA Methylation Patterning". In: *Nature* 452.7184, pp. 215–219. DOI: 10.1038/nature06745.

Consens, M. E., C. Dufault, M. Wainberg, D. Forster, M. Karimzadeh, H. Goodarzi, F. J. Theis, A. Moses, and B. Wang (2025). "Transformers and Genome Language Models". In: *Nature Machine Intelligence*. DOI: 10.1038/s42256-025-01007-9.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch (2010). "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State". In: *Proceedings of the National Academy of Sciences* 107.50, pp. 21931–21936. DOI: 10.1073/pnas.1016071107.

Crump, N. T., A. L. Smith, L. Godfrey, A. M. Dopico-Fernandez, N. Denny, J. R. Harman, J. C. Hamley, N. E. Jackson, C. Chahrour, S. Riva, S. Rice, J. Kim, V. Basrur, D. Fermin, K. Elenitoba-Johnson, R. G. Roeder, C. D. Allis, I. Roberts, A. Roy, H. Geng, J. O. J. Davies, and T. A. Milne (2023). "MLL-AF4 Cooperates with PAF1 and FACT to Drive High-Density Enhancer Interactions in Leukemia". In: *Nature Communications* 14.1, p. 5208. DOI: 10.1038/s41467-023-40981-9.

Daigle, S. R., E. J. Olhava, C. A. Therkelsen, C. R. Majer, C. J. Sneeringer, J. Song, L. D. Johnston, M. P. Scott, J. J. Smith, Y. Xiao, L. Jin, K. W. Kuntz, R. Chesworth, M. P. Moyer, K. M. Bernt, J.-C. Tseng, A. L. Kung, S. A. Armstrong, R. A. Copeland, V. M. Richon, and R. M. Pollock (2011). "Selective Killing of Mixed Lineage Leukemia Cells by a Potent Small-Molecule DOT1L Inhibitor". In: *Cancer Cell* 20.1, pp. 53–65. DOI: 10.1016/j.ccr.2011.06.009.

Dalla-Torre, H., L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. De Almeida, H. Sirelkhatim,

G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot (2024). “Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics”. In: *Nature Methods*. DOI: 10.1038/s41592-024-02523-z.

Dawson, M. A., R. K. Prinjha, A. Dittmann, G. Giotopoulos, M. Bantscheff, W.-I. Chan, S. C. Robson, C.-w. Chung, C. Hopf, M. M. Savitski, C. Huthmacher, E. Gudgin, D. Lugo, S. Beinke, T. D. Chapman, E. J. Roberts, P. E. Soden, K. R. Auger, O. Mirguet, K. Doehner, R. Delwel, A. K. Burnett, P. Jeffrey, G. Drewes, K. Lee, B. J. P. Huntly, and T. Kouzarides (2011). “Inhibition of BET Recruitment to Chromatin as an Effective Treatment for MLL-fusion Leukaemia”. In: *Nature* 478.7370, pp. 529–533. DOI: 10.1038/nature10509.

Deaton, A. M. and A. Bird (2011). “CpG Islands and the Regulation of Transcription”. In: *Genes & Development* 25.10, pp. 1010–1022. DOI: 10.1101/gad.2037511.

Deshpande, A. J., L. Chen, M. Fazio, A. U. Sinha, K. M. Bernt, D. Banka, S. Dias, J. Chang, E. J. Olhava, S. R. Daigle, V. M. Richon, R. M. Pollock, and S. A. Armstrong (2013). “Leukemic Transformation by the MLL-AF6 Fusion Oncogene Requires the H3K79 Methyltransferase Dot1l”. In: *Blood* 121.13, pp. 2533–2541. DOI: 10.1182/blood-2012-11-465120.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Version 2. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805 (visited on 09/28/2025). Pre-published.

Di Tommaso, P., M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame (2017). “Nextflow Enables Reproducible Computational Workflows”. In: *Nature Biotechnology* 35.4, pp. 316–319. DOI: 10.1038/nbt.3820.

Djabali, M., L. Selleri, P. Parry, M. Bower, B. D. Young, and G. A. Evans (1992). “A Trithorax–like Gene Is Interrupted by Chromosome 11q23 Translocations in Acute Leukaemias”. In: *Nature Genetics* 2.2, pp. 113–118. DOI: 10.1038/ng1092-113.

Dou, Y., T. A. Milne, A. J. Ruthenburg, S. Lee, J. W. Lee, G. L. Verdine, C. D. Allis, and R. G. Roeder (2006). “Regulation of MLL1 H3K4 Methyltransferase Activity by Its Core Components”. In: *Nature Structural & Molecular Biology* 13.8 (8), pp. 713–719. DOI: 10.1038/nsmb1128.

El Ashkar, S., J. Schwaller, T. Pieters, S. Goossens, J. Demeulemeester, F. Christ, S. Van Belle, S. Juge, N. Boeckx, A. Engelman, P. Van Vlierberghe, Z. Debyser, and J. De Rijck (2017). “LEDGF/P75 Is Dispensable for Hematopoiesis but Essential for MLL-rearranged Leukemogenesis”. In: *Blood*, blood-2017-05–786962. DOI: 10.1182/blood-2017-05-786962.

Erfurth, F. E., R. Popovic, J. Grembecka, T. Cierpicki, C. Theisler, Z.-B. Xia, T. Stuart, M. O. Diaz, J. H. Bushweller, and N. J. Zeleznik-Le (2008). “MLL Protects CpG Clusters from Methylation within the *Hoxa9* Gene, Maintaining Transcript Expression”. In: *Proceedings of the National Academy of Sciences* 105.21, pp. 7517–7522. DOI: 10.1073/pnas.0800090105.

Ewels, P. A., A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen (2020). “The Nf-Core Framework for Community-Curated Bioinformatics Pipelines”. In: *Nature Biotechnology* 38.3, pp. 276–278. DOI: 10.1038/s41587-020-0439-x.

Fang, C., S. Rao, J. D. Crispino, and P. Ntziachristos (2020). “Determinants and Role of Chromatin Organization in Acute Leukemia”. In: *Leukemia* 34.10, pp. 2561–2575. DOI: 10.1038/s41375-020-0981-z. PMID: 32690881.

Farcas, A. M., N. P. Blackledge, I. Sudbery, H. K. Long, J. F. McGouran, N. R. Rose, S. Lee, D. Sims, A. Cerase, T. W. Sheahan, H. Koseki, N. Brockdorff, C. P. Ponting, B. M. Kessler, and R. J. Klose (2012). “KDM2B Links the Poly-

comb Repressive Complex 1 (PRC1) to Recognition of CpG Islands". In: *eLife* 1, e00205. DOI: 10.7554/eLife.00205.

Felix, C. A., M. R. Hosler, N. J. Winick, M. Masterson, A. E. Wilson, and B. J. Lange (1995). "ALL-1 Gene Rearrangements in DNA Topoisomerase II Inhibitor-Related Leukemia in Children". In: *Blood* 85.11, pp. 3250–3256. PMID: 7756657.

Fong, C. Y., J. Morison, and M. A. Dawson (2014). "Epigenetics in the Hematologic Malignancies". In: *Haematologica* 99.12, pp. 1772–1783. DOI: 10.3324/haematol.2013.092007. PMID: 25472952.

Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul (1992). "A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands." In: *Proceedings of the National Academy of Sciences* 89.5, pp. 1827–1831. DOI: 10.1073/pnas.89.5.1827.

Füllgrabe, J., W. S. Gosal, P. Creed, S. Liu, C. K. Lumby, D. J. Morley, T. W. B. Ost, A. J. Vilella, S. Yu, H. Bignell, P. Burns, T. Charlesworth, B. Fu, H. Fordham, N. J. Harding, O. Gandelman, P. Golder, C. Hodson, M. Li, M. Lila, Y. Liu, J. Mason, J. Mellad, J. M. Monahan, O. Nentwich, A. Palmer, M. Steward, M. Taipale, A. Vandomme, R. S. San-Bento, A. Singhal, J. Vivian, N. Wójtowicz, N. Williams, N. J. Walker, N. C. H. Wong, G. N. Yalloway, J. D. Holbrook, and S. Balasubramanian (2023). "Simultaneous Sequencing of Genetic and Epigenetic Bases in DNA". In: *Nature Biotechnology* 41.10, pp. 1457–1464. DOI: 10.1038/s41587-022-01652-0.

Gale, K. B., A. M. Ford, R. Repp, A. Borkhardt, C. Keller, O. B. Eden, and M. F. Greaves (1997). "Backtracking Leukemia to Birth: Identification of Clonotypic Gene Fusion Sequences in Neonatal Blood Spots". In: *Proceedings of the National Academy of Sciences* 94.25, pp. 13950–13954. DOI: 10.1073/pnas.94.25.13950.

Godfrey, L., N. T. Crump, S. O'Byrne, I.-J. Lau, S. Rice, J. R. Harman, T. Jackson, N. Elliott, G. Buck, C. Connor, R. Thorne, D. J. H. F. Knapp, O. Heidenreich, P. Vyas, P. Menendez, S. Inglott, P. Ancliff, H. Geng, I. Roberts, A. Roy, and T. A. Milne (2021). "H3K79me2/3 Controls Enhancer-Promoter Interactions and Activation of the Pan-Cancer Stem Cell Marker PROM1/CD133 in MLL-AF4 Leukemia Cells." In: *Leukemia* 35.1, pp. 90–106. DOI: 10.1038/s41375-020-0808-y. PMID: 32242051.

Godfrey, L., N. T. Crump, R. Thorne, I.-J. Lau, E. Repapi, D. Dimou, A. L. Smith, J. R. Harman, J. M. Telenius, A. M. Oudelaar, D. J. Downes, P. Vyas, J. R. Hughes, and T. A. Milne (2019). "DOT1L Inhibition Reveals a Distinct Subset of Enhancers Dependent on H3K79 Methylation". In: *Nature Communications* 10.1, p. 2803. DOI: 10.1038/s41467-019-10844-3.

Gonzalez-Avalos, E., A. Onodera, D. Samaniego-Castruita, A. Rao, and F. Ay (2024). "Predicting Gene Expression State and Prioritizing Putative Enhancers Using 5hmC Signal". In: *Genome Biology* 25.1, p. 142. DOI: 10.1186/s13059-024-03273-z.

Grau, J., F. Schmidt, and M. H. Schulz (2023). "Widespread Effects of DNA Methylation and Intra-Motif Dependencies Revealed by Novel Transcription Factor Binding Models". In: *Nucleic Acids Research* 51.18, e95–e95. DOI: 10.1093/nar/gkad693.

Greenberg, M. V. C. and D. Bourc'his (2019). "The Diverse Roles of DNA Methylation in Mammalian Development and Disease". In: *Nature Reviews Molecular Cell Biology* 20.10, pp. 590–607. DOI: 10.1038/s41580-019-0159-6.

Greil, J., M. Gramatzki, R. Burger, R. Marschalek, M. Peltner, U. Trautmann, T. E. Hansen-Hagge, C. R. Bartram, G. H. Fey, K. Stehr, and J. Beck (1994). “The Acute Lymphoblastic Leukaemia Cell Line SEM with t(4;11) Chromosomal Rearrangement Is Biphenotypic and Responsive to Interleukin-7”. In: *British Journal of Haematology* 86.2, pp. 275–283. DOI: 10.1111/j.1365-2141.1994.tb04726.x.

Gupta, S., J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble (2007). “Quantifying Similarity between Motifs”. In: *Genome Biology* 8.2, R24. DOI: 10.1186/gb-2007-8-2-r24.

Hamley, J. C., H. Li, N. Denny, D. Downes, and J. O. J. Davies (2023). “Determining Chromatin Architecture with Micro Capture-C”. In: *Nature Protocols* 18.6, pp. 1687–1711. DOI: 10.1038/s41596-023-00817-8.

He, Y.-F., B.-Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, Y. Sun, X. Li, Q. Dai, C.-X. Song, K. Zhang, C. He, and G.-L. Xu (2011). “Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA”. In: *Science* 333.6047, pp. 1303–1307. DOI: 10.1126/science.1210944.

Héberlé, É. and A. F. Bardet (2019). “Sensitivity of Transcription Factors to DNA Methylation”. In: *Essays in Biochemistry* 63.6. Ed. by M. Blewitt, pp. 727–741. DOI: 10.1042/EBC20190033.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass (2010). “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities”. In: *Molecular Cell* 38.4, pp. 576–589. DOI: 10.1016/j.molcel.2010.05.004.

Heinz, S., C. E. Romanoski, C. Benner, and C. K. Glass (2015). “The Selection and Function of Cell Type-Specific Enhancers”. In: *Nature Reviews Molecular Cell Biology* 16.3, pp. 144–154. DOI: 10.1038/nrm3949.

Hentges, L. D., M. J. Sergeant, C. B. Cole, D. J. Downes, J. R. Hughes, and S. Taylor (2022). “LanceOtron: A Deep Learning Peak Caller for Genome Sequencing Experiments”. In: *Bioinformatics*, btac525. DOI: 10.1093/bioinformatics/btac525.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, A. A. Sigova, H. A. Hoke, and R. A. Young (2013). “Super-Enhancers in the Control of Cell Identity and Disease”. In: *Cell* 155.4, pp. 934–947. DOI: 10.1016/j.cell.2013.09.053.

Hotchkiss, R. D. (1948). “The Quantitative Separation of Purines, Pyrimidines, And Nucleosides by Paper Chromatography”. In: *Journal of Biological Chemistry* 175.1, pp. 315–332. DOI: 10.1016/S0021-9258(18)57261-6.

Houlsby, N., A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly (2019). *Parameter-Efficient Transfer Learning for NLP*. Version 2. DOI: 10.48550/ARXIV.1902.00751. URL: https://arxiv.org/abs/1902.00751 (visited on 09/28/2025). Pre-published.

Howard, J. and S. Ruder (2018). *Universal Language Model Fine-tuning for Text Classification*. Version 5. DOI: 10.48550/ARXIV.1801.06146. URL: https://arxiv.org/abs/1801.06146 (visited on 09/28/2025). Pre-published.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs]. URL: http://arxiv.org/abs/2106.09685 (visited on 08/08/2025). Pre-published.

Hu, S., J. Wan, Y. Su, Q. Song, Y. Zeng, H. N. Nguyen, J. Shin, E. Cox, H. S. Rho, C. Woodard, S. Xia, S. Liu, H. Lyu, G.-L. Ming, H. Wade, H. Song, J. Qian, and H. Zhu (2013). “DNA Methylation Presents Distinct Binding Sites for Human Transcription Factors”. In: *eLife* 2, e00726. DOI: 10.7554/eLife.00726.

Jack, I., R. Seshadri, M. Garson, P. Michael, D. Callen, H. Zola, and A. Morley (1986). “RCH-ACV: A Lymphoblastic Leukemia Cell Line with Chromosome Translocation 1;19 and Trisomy 8”. In: *Cancer Genetics and Cytogenetics* 19.3–4, pp. 261–269. DOI: 10.1016/0165-4608(86)90055-5.

Ji, Y., Z. Zhou, H. Liu, and R. V. Davuluri (2021). “DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-language in Genome”. In: *Bioinformatics* 37.15. Ed. by J. Kelso, pp. 2112–2120. DOI: 10.1093/bioinformatics/btab083.

Joseph, M. and H. Raj (2022). *GANDALF: Gated Adaptive Network for Deep Automated Learning of Features*. Version 6. DOI: 10.48550/ARXIV.2207.08548. URL: https://arxiv.org/abs/2207.08548 (visited on 09/03/2025). Pre-published.

Kanno, T., Y. Kanno, G. LeRoy, E. Campos, H.-W. Sun, S. R. Brooks, G. Vahedi, T. D. Heightman, B. A. Garcia, D. Reinberg, U. Siebenlist, J. J. O’Shea, and K. Ozato (2014). “BRD4 Assists Elongation of Both Coding and Enhancer RNAs by Interacting with Acetylated Histones”. In: *Nature Structural & Molecular Biology* 21.12, pp. 1047–1057. DOI: 10.1038/nsmb.2912.

Kaya-Okur, H. S., S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff (2019). “CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells”. In: *Nature Communications* 10.1, p. 1930. DOI: 10.1038/s41467-019-09982-5.

Kelley, D. R., J. Snoek, and J. L. Rinn (2016). “Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks”. In: *Genome Research* 26.7, pp. 990–999. DOI: 10.1101/gr.200535.115.

Kempynck, N., S. De Winter, C. H. Blaauw, V. Konstantakos, S. Dieltiens, E. C. Ekşi, V. Bercier, I. I. Taskiran, G. Hulselmans, K. Spanier, V. Christiaens, L. Van Den Bosch, L. Mahieu, and S. Aerts (2025). *CREsted: Modeling Genomic and Synthetic Cell Type-Specific Enhancers across Tissues and Species*. DOI: 10.1101/2025.04.02.646812. URL: http://biorxiv.org/lookup/doi/10.1101/2025.04.02.646812 (visited on 08/08/2025). Pre-published.

Kerry, J., L. Godfrey, E. Repapi, M. Tapia, N. P. Blackledge, H. Ma, E. Ballabio, S. O’Byrne, F. Ponthan, O. Heidenreich, A. Roy, I. Roberts, M. Konopleva, R. J. Klose, H. Geng, and T. A. Milne (2017). “MLL-AF4 Spreading Identifies Binding Sites That Are Distinct from Super-Enhancers and That Govern Sensitivity to DOT1L Inhibition in Leukemia.” In: *Cell reports* 18.2, pp. 482–495. DOI: 10.1016/j.celrep.2016.12.054. PMID: 28076791.

Khetan, S., B. S. Carroll, and M. L. Bulyk (2025). “Multiple Overlapping Binding Sites Determine Transcription Factor Occupancy”. In: *Nature*. DOI: 10.1038/s41586-025-09472-3.

Khund Sayeed, S., J. Zhao, B. K. Sathyanarayana, J. P. Golla, and C. Vinson (2015). “C/EBPβ (CEBPB) Protein Binding to the C/EBP|CRE DNA 8-Mer TTGC|GTCA Is Inhibited by 5hmC and Enhanced by 5mC, 5fC, and 5caC in the CG Dinucleotide”. In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1849.6, pp. 583–589. DOI: 10.1016/j.bbagrm.2015.03.002.

Klemm, S. L., Z. Shipony, and W. J. Greenleaf (2019). “Chromatin Accessibility and the Regulatory Epigenome”. In: *Nature Reviews Genetics* 20.4, pp. 207–220. DOI: 10.1038/s41576-018-0089-8.

Klose, R. J., S. Cooper, A. M. Farcas, N. P. Blackledge, and N. Brockdorff (2013). “Chromatin Sampling—An Emerging Perspective on Targeting Polycomb Repressor Proteins”. In: *PLoS Genetics* 9.8. Ed. by W. Reik, e1003717. DOI: 10.1371/journal.pgen.1003717.

Ko, M., H. S. Bandukwala, J. An, E. D. Lamperti, E. C. Thompson, R. Hastie, A. Tsangaratou, K. Rajewsky, S. B. Koralov, and A. Rao (2011). “Ten-Eleven-

Translocation 2 (TET2) Negatively Regulates Homeostasis and Differentiation of Hematopoietic Stem Cells in Mice". In: *Proceedings of the National Academy of Sciences* 108.35, pp. 14566–14571. DOI: 10.1073/pnas.1112317108.

Kouzarides, T. (2007). "Chromatin Modifications and Their Function". In: *Cell* 128.4, pp. 693–705. DOI: 10.1016/j.cell.2007.02.005. PMID: 17320507.

Krivtsov, A. V. and S. A. Armstrong (2007). "MLL Translocations, Histone Modifications and Leukaemia Stem-Cell Development". In: *Nature Reviews Cancer* 7.11 (11), pp. 823–833. DOI: 10.1038/nrc2253.

Krivtsov, A. V., K. Evans, J. Y. Gadrey, B. K. Eschle, C. Hatton, H. J. Uckelmann, K. N. Ross, F. Perner, S. N. Olsen, T. Pritchard, L. McDermott, C. D. Jones, D. Jing, A. Braytee, D. Chacon, E. Earley, B. M. McKeever, D. Claremon, A. J. Gifford, H. J. Lee, B. A. Teicher, J. E. Pimanda, D. Beck, J. A. Perry, M. A. Smith, G. M. McGeehan, R. B. Lock, and S. A. Armstrong (2019). "A Menin-MLL Inhibitor Induces Specific Chromatin Changes and Eradicates Disease in Models of MLL-Rearranged Leukemia". In: *Cancer Cell* 36.6, 660–673.e11. DOI: 10.1016/j.ccell.2019.11.001.

Lambert, S. A., A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch (2018). "The Human Transcription Factors". In: *Cell* 172.4, pp. 650–665. DOI: 10.1016/j.cell.2018.01.029.

Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder (2012). "ChIP-seq Guidelines and Practices of the ENCODE and modENCODE Consortia". In: *Genome Research* 22.9, pp. 1813–1831. DOI: 10.1101/gr.136184.111.

Lange, B., M. Valtieri, D. Santoli, D. Caracciolo, F. Mavilio, I. Gemperlein, C. Griffin, B. Emanuel, J. Finan, and P. Nowell (1987). "Growth Factor Requirements of Childhood Acute Leukemia: Establishment of GM-CSF-dependent Cell Lines". In: *Blood* 70.1, pp. 192–199. DOI: 10.1182/blood.V70.1.192.192.

Lara-Astiaso, D., A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, and I. Amit (2014). "Chromatin State Dynamics during Blood Formation". In: *Science* 345.6199, pp. 943–949. DOI: 10.1126/science.1256271.

Li, E. and Y. Zhang (2014). "DNA Methylation in Mammals". In: *Cold Spring Harbor Perspectives in Biology* 6.5, a019133–a019133. DOI: 10.1101/cshperspect.a019133.

Libbrecht, M. W. and W. S. Noble (2015). "Machine Learning Applications in Genetics and Genomics". In: *Nature Reviews Genetics* 16.6, pp. 321–332. DOI: 10.1038/nrg3920.

Lichtinger, M., R. Ingram, R. Hannah, D. Müller, D. Clarke, S. A. Assi, M. Lie-A-Ling, L. Noailles, M. S. Vijayabaskar, M. Wu, D. G. Tenen, D. R. Westhead, V. Kouskoff, G. Lacaud, B. Göttgens, and C. Bonifer (2012). "RUNX1 Reshapes the Epigenetic Landscape at the Onset of Haematopoiesis: RUNX1 Shifts Transcription Factor Binding Patterns". In: *The EMBO Journal* 31.22, pp. 4318–4333. DOI: 10.1038/emboj.2012.275.

Lin, C., E. R. Smith, H. Takahashi, K. C. Lai, S. Martin-Brown, L. Florens, M. P. Washburn, J. W. Conaway, R. C. Conaway, and A. Shilatifard (2010). "AFF4, a Component of the ELL/P-TEFb Elongation Complex and a Shared Subunit of

MLL Chimeras, Can Link Transcription Elongation to Leukemia". In: *Molecular Cell* 37.3, pp. 429–437. DOI: 10.1016/j.molcel.2010.01.026.

Lister, R., E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker (2013). "Global Epigenomic Reconfiguration During Mammalian Brain Development". In: *Science* 341.6146, p. 1237905. DOI: 10.1126/science.1237905.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker (2009). "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences". In: *Nature* 462.7271, pp. 315–322. DOI: 10.1038/nature08514.

Liu, Y., P. Siejka-Zielińska, G. Velikova, Y. Bi, F. Yuan, M. Tomkova, C. Bai, L. Chen, B. Schuster-Böckler, and C.-X. Song (2019). "Bisulfite-Free Direct Detection of 5-Methylcytosine and 5-Hydroxymethylcytosine at Base Resolution". In: *Nature Biotechnology* 37.4, pp. 424–429. DOI: 10.1038/s41587-019-0041-2.

Love, M. I., W. Huber, and S. Anders (2014). "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2". In: *Genome Biology* 15.12, p. 550. DOI: 10.1186/s13059-014-0550-8.

Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond (1997). "Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution". In: *Nature* 389.6648, pp. 251–260. DOI: 10.1038/38444.

Lundberg, S. and S.-I. Lee (2017). *A Unified Approach to Interpreting Model Predictions*. Version 2. DOI: 10.48550/ARXIV.1705.07874. URL: https://arxiv.org/abs/1705.07874 (visited on 09/28/2025). Pre-published.

Luo, X., T. Zhang, Y. Zhai, F. Wang, S. Zhang, and G. Wang (2021). "Effects of DNA Methylation on TFs in Human Embryonic Stem Cells". In: *Frontiers in Genetics* 12, p. 639461. DOI: 10.3389/fgene.2021.639461.

Ma, W., W. S. Noble, and T. L. Bailey (2014). "Motif-Based Analysis of Large Nucleotide Data Sets Using MEME-ChIP". In: *Nature Protocols* 9.6 (6), pp. 1428–1450. DOI: 10.1038/nprot.2014.083.

Ma, X., S. R. Thela, F. Zhao, B. Yao, Z. Wen, P. Jin, J. Zhao, and L. Chen (2024). "Deep5hmC: Predicting Genome-Wide 5-Hydroxymethylcytosine Landscape via a Multimodal Deep Learning Model". In: *Bioinformatics* 40.9. Ed. by M. Nikolski, btae528. DOI: 10.1093/bioinformatics/btae528.

Mellén, M., P. Ayata, S. Dewell, S. Kriaucionis, and N. Heintz (2012). "MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System". In: *Cell* 151.7, pp. 1417–1430. DOI: 10.1016/j.cell.2012.11.022.

Mendonca, A., E. H. Chang, W. Liu, and C. Yuan (2014). "Hydroxymethylation of DNA Influences Nucleosomal Conformation and Stability in Vitro". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1839.11, pp. 1323–1329. DOI: 10.1016/j.bbagrm.2014.09.014.

Meyer, C., P. Larghero, B. Almeida Lopes, T. Burmeister, D. Gröger, R. Sutton, N. C. Venn, G. Cazzaniga, L. Corral Abascal, G. Tsaur, L. Fechina, M. Emerenciano, M. S. Pombo-de-Oliveira, T. Lund-Aho, T. Lundán, M. Montonen, V. Juvonen, J. Zuna, J. Trka, P. Ballerini, H. Lapillonne, V. H. J. Van Der Velden, E. Sonneveld, E. Delabesse, R. R. C. De Matos, M. L. M. Silva, S. Bomken, K. Katsibardi, M. Keernik, N. Grardel, J. Mason, R. Price, J. Kim, C. Eckert, L. Lo Nigro, C. Bueno, P. Menendez, U. Zur Stadt, P. Gameiro, L. Sedék, T. Szczepański, A. Bidet, V. Marcu, K. Shichrur, S. Izraeli, H. O. Madsen, B. W. Schäfer, S. Kubetzko, R. Kim, E. Clappier, H. Trautmann, M. Brüggemann, P. Archer, J. Hancock, J.

Alten, A. Möricke, M. Stanulla, J. Lentes, A. K. Bergmann, S. Strehl, S. Köhrer, K. Nebral, M. N. Dworzak, O. A. Haas, C. Arfeuille, A. Caye-Eude, H. Cavé, and R. Marschalek (2023). “The KMT2A Recombinome of Acute Leukemias in 2023”. In: *Leukemia* 37.5, pp. 988–1005. DOI: 10.1038/s41375-023-01877-1.

Milne, T. A., S. D. Briggs, H. W. Brock, M. E. Martin, D. Gibbs, C. D. Allis, and J. L. Hess (2002). “MLL Targets SET Domain Methyltransferase Activity to Hox Gene Promoters.” In: *Molecular cell* 10.5, pp. 1107–1117. DOI: 10.1016/s1097-2765(02)00741-4. PMID: 12453418.

Milne, T. A., Y. Dou, M. E. Martin, H. W. Brock, R. G. Roeder, and J. L. Hess (2005a). “MLL Associates Specifically with a Subset of Transcriptionally Active Target Genes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.41, pp. 14765–14770. DOI: 10.1073/pnas.0503630102. PMID: 16199523.

Milne, T. A., J. Kim, G. G. Wang, S. C. Stadler, V. Basrur, S. J. Whitcomb, Z. Wang, A. J. Ruthenburg, K. S. J. Elenitoba-Johnson, R. G. Roeder, and C. D. Allis (2010). “Multiple Interactions Recruit MLL1 and MLL1 Fusion Proteins to the HOXA9 Locus in Leukemogenesis.” In: *Molecular cell* 38.6, pp. 853–863. DOI: 10.1016/j.molcel.2010.05.011. PMID: 20541448.

Milne, T. A., M. E. Martin, H. W. Brock, R. K. Slany, and J. L. Hess (2005b). “Leukemogenic MLL Fusion Proteins Bind across a Broad Region of the Hox A9 Locus, Promoting Transcription and Multiple Histone Modifications.” In: *Cancer research* 65.24, pp. 11367–11374. DOI: 10.1158/0008-5472.CAN-05-1041. PMID: 16357144.

Mölder, F., K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster (2021). “Sustainable Data Analysis with Snakemake”. In: *F1000Research* 10, p. 33. DOI: 10.12688/f1000research.29032.2.

Morgan, D., D. L. DeMeo, and K. Glass (2024). “Using Methylation Data to Improve Transcription Factor Binding Prediction”. In: *Epigenetics* 19.1, p. 2309826. DOI: 10.1080/15592294.2024.2309826.

Mosquera Orgueira, A., O. Krali, C. Pérez Míguez, A. Peleteiro Raíndo, J. Á. Díaz Arias, M. S. González Pérez, M. M. Pérez Encinas, M. Fernández Sanmartín, D. Sinnet, M. Heyman, G. Lönnerholm, U. Norén-Nyström, K. Schmiegelow, and J. Nordlund (2024). “Refining Risk Prediction in Pediatric Acute Lymphoblastic Leukemia through DNA Methylation Profiling”. In: *Clinical Epigenetics* 16.1, p. 49. DOI: 10.1186/s13148-024-01662-6.

Mueller, D., M.-P. García-Cuéllar, C. Bach, S. Buhl, E. Maethner, and R. K. Slany (2009). “Misguided Transcriptional Elongation Causes Mixed Lineage Leukemia”. In: *PLoS biology* 7.11, e1000249. DOI: 10.1371/journal.pbio.1000249. PMID: 19956800.

Muntean, A. G., J. Tan, K. Sitwala, Y. Huang, J. Bronstein, J. A. Connelly, V. Basrur, K. S. Elenitoba-Johnson, and J. L. Hess (2010). “The PAF Complex Synergizes with MLL Fusion Proteins at HOX Loci to Promote Leukemogenesis”. In: *Cancer Cell* 17.6, pp. 609–621. DOI: 10.1016/j.ccr.2010.04.012.

Nakamura, T., T. Mori, S. Tada, W. Krajewski, T. Rozovskaia, R. Wassell, G. Dubois, A. Mazo, C. M. Croce, and E. Canaani (2002). “ALL-1 Is a Histone Methyltransferase That Assembles a Supercomplex of Proteins Involved in Transcriptional Regulation”. In: *Molecular Cell* 10.5, pp. 1119–1128. DOI: 10.1016/S1097-2765(02)00740-2. PMID: 12453419.

Nan, X., H.-H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman, and A. Bird (1998). “Transcriptional Repression by the Methyl-CpG-binding Pro-

tein MeCP2 Involves a Histone Deacetylase Complex". In: *Nature* 393.6683, pp. 386–389. DOI: 10.1038/30764.

Nguyen, A. T. and Y. Zhang (2011). "The Diverse Functions of Dot1 and H3K79 Methylation". In: *Genes & Development* 25.13, pp. 1345–1358. DOI: 10.1101/gad.2057811.

Nishizaki, S. S. and A. P. Boyle (2022). "SEMplMe: A Tool for Integrating DNA Methylation Effects in Transcription Factor Binding Affinity Predictions". In: *BMC Bioinformatics* 23.1, p. 317. DOI: 10.1186/s12859-022-04865-x.

Okuda, H., M. Kawaguchi, A. Kanai, H. Matsui, T. Kawamura, T. Inaba, I. Kitabayashi, and A. Yokoyama (2014). "MLL Fusion Proteins Link Transcriptional Coactivators to Previously Active CpG-rich Promoters". In: *Nucleic Acids Research* 42.7, pp. 4241–4256. DOI: 10.1093/nar/gkt1394.

Pastor, W. A., U. J. Pape, Y. Huang, H. R. Henderson, R. Lister, M. Ko, E. M. McLoughlin, Y. Brudno, S. Mahapatra, P. Kapranov, M. Tahiliani, G. Q. Daley, X. S. Liu, J. R. Ecker, P. M. Milos, S. Agarwal, and A. Rao (2011). "Genome-Wide Mapping of 5-Hydroxymethylcytosine in Embryonic Stem Cells". In: *Nature* 473.7347, pp. 394–397. DOI: 10.1038/nature10102.

Pelish, H. E., B. B. Liau, I. I. Nitulescu, A. Tangpeerachaikul, Z. C. Poss, D. H. Da Silva, B. T. Caruso, A. Arefolov, O. Fadeyi, A. L. Christie, K. Du, D. Banka, E. V. Schneider, A. Jestel, G. Zou, C. Si, C. C. Ebmeier, R. T. Bronson, A. V. Krivtsov, A. G. Myers, N. E. Kohl, A. L. Kung, S. A. Armstrong, M. E. Lemieux, D. J. Taatjes, and M. D. Shair (2015). "Mediator Kinase Inhibition Further Activates Super-Enhancer-Associated Genes in AML". In: *Nature* 526.7572, pp. 273–276. DOI: 10.1038/nature14904.

Pi, W.-C., J. Wang, M. Shimada, J.-W. Lin, H. Geng, Y.-L. Lee, R. Lu, D. Li, G. G. Wang, R. G. Roeder, and W.-Y. Chen (2020). "E2A-PBX1 Functions as a Coactivator for RUNX1 in Acute Lymphoblastic Leukemia". In: *Blood* 136.1, pp. 11–23. DOI: 10.1182/blood.2019003312.

Pique-Regi, R., J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard (2011). "Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data". In: *Genome Research* 21.3, pp. 447–455. DOI: 10.1101/gr.112623.110.

Pomp, W., J. V. Meeussen, and T. L. Lenstra (2024). "Transcription Factor Exchange Enables Prolonged Transcriptional Bursts". In: *Molecular Cell* 84.6, 1036–1048.e9. DOI: 10.1016/j.molcel.2024.01.020.

Pradhan, M., P.-O. Estève, H. G. Chin, M. Samaranayke, G.-D. Kim, and S. Pradhan (2008). "CXXC Domain of Human DNMT1 Is Essential for Enzymatic Activity". In: *Biochemistry* 47.38, pp. 10000–10009. DOI: 10.1021/bi8011725.

Prendergast, G. C. and E. B. Ziff (1991). "Methylation-Sensitive Sequence-Specific DNA Binding by the c-Myc Basic Region". In: *Science* 251.4990, pp. 186–189. DOI: 10.1126/science.1987636.

Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O'Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy, and J. M. Ostell (2014). "RefSeq: An Update on Mammalian Reference Sequences". In: *Nucleic Acids Research* 42.D1, pp. D756–D763. DOI: 10.1093/nar/gkt1114.

Quang, D. and X. Xie (2016). "DanQ: A Hybrid Convolutional and Recurrent Deep Neural Network for Quantifying the Function of DNA Sequences". In: *Nucleic Acids Research* 44.11, e107–e107. DOI: 10.1093/nar/gkw226.

Quentmeier, H., M. P. Martelli, W. G. Dirks, N. Bolli, A. Liso, R. A. F. MacLeod, I. Nicoletti, R. Mannucci, A. Pucciarini, B. Bigerna, M. F. Martelli, C. Mecucci, H. G. Drexler, and B. Falini (2005). "Cell Line OCI/AML3 Bears Exon-12 NPM Gene Mutation-A and Cytoplasmic Expression of Nucleophosmin". In: *Leukemia* 19.10, pp. 1760–1767. DOI: 10.1038/sj.leu.2403899.

Rasmussen, K. D., G. Jia, J. V. Johansen, M. T. Pedersen, N. Rapin, F. O. Bagger, B. T. Porse, O. A. Bernard, J. Christensen, and K. Helin (2015). "Loss of *TET2* in Hematopoietic Cells Leads to DNA Hypermethylation of Active Enhancers and Induction of Leukemogenesis". In: *Genes & Development* 29.9, pp. 910–922. DOI: 10.1101/gad.260174.115.

Rauluseviciute, I., F. Drabløs, and M. B. Rye (2020). "DNA Hypermethylation Associated with Upregulated Gene Expression in Prostate Cancer Demonstrates the Diversity of Epigenetic Regulation". In: *BMC Medical Genomics* 13.1, p. 6. DOI: 10.1186/s12920-020-0657-6.

Richter, W. F., S. Nayak, J. Iwasa, and D. J. Taatjes (2022). "The Mediator Complex as a Master Regulator of Transcription by RNA Polymerase II". In: *Nature Reviews Molecular Cell Biology* 23.11, pp. 732–749. DOI: 10.1038/s41580-022-00498-3.

Roe, J.-S., F. Mercan, K. Rivera, D. J. Pappin, and C. R. Vakoc (2015). "BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia". In: *Molecular Cell* 58.6, pp. 1028–1039. DOI: 10.1016/j.molcel.2015.04.011.

Ruthenburg, A. J., H. Li, T. A. Milne, S. Dewell, R. K. McGinty, M. Yuen, B. Ueberheide, Y. Dou, T. W. Muir, D. J. Patel, and C. D. Allis (2011). "Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions." In: *Cell* 145.5, pp. 692–706. DOI: 10.1016/j.cell.2011.03.053. PMID: 21596426.

Sanabria, M., J. Hirsch, P. M. Joubert, and A. R. Poetsch (2024). "DNA Language Model GROVER Learns Sequence Context in the Human Genome". In: *Nature Machine Intelligence*. DOI: 10.1038/s42256-024-00872-0.

Schneider, T. D. and R. Stephens (1990). "Sequence Logos: A New Way to Display Consensus Sequences". In: *Nucleic Acids Research* 18.20, pp. 6097–6100. DOI: 10.1093/nar/18.20.6097.

Schneider, V. A., T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C.-S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, and D. M. Church (2017). "Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly". In: *Genome Research* 27.5, pp. 849–864. DOI: 10.1101/gr.213611.116.

Shilatifard, A. (2012). "The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis". In: *Annual Review of Biochemistry* 81.1, pp. 65–95. DOI: 10.1146/annurev-biochem-051710-134100.

Shin Voo, K., D. L. Carlone, B. M. Jacobsen, A. Flodin, and D. G. Skalnik (2000). "Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1". In: *Molecular and Cellular Biology* 20.6, pp. 2108–2121. DOI: 10.1128/MCB.20.6.2108-2121.2000.

Singer-Sam, J., M. Grant, J. M. Lebon, K. Okuyama, V. Chapman, M. Monk, and A. D. Riggs (1990). “Use of a *Hpa* II-Polymerase Chain Reaction Assay To Study DNA Methylation in the *Pgk-1* CpG Island of Mouse Embryos at the Time of X-Chromosome Inactivation”. In: *Molecular and Cellular Biology* 10.9, pp. 4987–4989. DOI: 10.1128/mcb.10.9.4987-4989.1990.

Smith, A., N. D. R. Denny, C. Chahrour, K. Sharp, M. Arachi, A. M. Dopico-Fernandez, N. Elliott, J. Harman, T. Jackson, H. Geng, O. Smith, J. Bond, I. Roberts, R. W. W. Stam, N. Crump, J. Davies, A. Roy, and T. A. Milne (2025a). “Enhancer Heterogeneity in Acute Lymphoblastic Leukemia Drives Differential Gene Expression in Patients”. In: *Blood Journal*, blood.2024028019. DOI: 10.1182/blood.2024028019.

Smith, T. and M. Waterman (1981). “Identification of Common Molecular Subsequences”. In: *Journal of Molecular Biology* 147.1, pp. 195–197. DOI: 10.1016/0022-2836(81)90087-5.

Smith, Z. D., S. Hetzel, and A. Meissner (2025b). “DNA Methylation in Mammalian Development and Disease”. In: *Nature Reviews Genetics* 26.1, pp. 7–30. DOI: 10.1038/s41576-024-00760-8.

Spitz, F. (2012). “Transcription Factors: From Enhancer Binding to Developmental Control”. In: *Nature Reviews Genetics*, p. 14.

Spruijt, C. G., F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. Jansen, C. Bauer, M. Münzel, M. Wagner, M. Müller, F. Khan, H. C. Eberl, A. Mensinga, A. B. Brinkman, K. Lephikov, U. Müller, J. Walter, R. Boelens, H. van Ingen, H. Leonhardt, T. Carell, and M. Vermeulen (2013). “Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives”. In: *Cell* 152.5, pp. 1146–1159. DOI: 10.1016/j.cell.2013.02.004.

Stong, R. C., S. J. Korsmeyer, J. L. Parkin, D. C. Arthur, and J. H. Kersey (1985). “Human Acute Leukemia Cell Line with the t(4;11) Chromosomal Rearrangement Exhibits B Lineage and Monocytic Characteristics”. In: *Blood* 65.1, pp. 21–31. PMID: 3917311.

Stormo, G. D., T. D. Schneider, L. Gold, and A. Ehrenfeucht (1982). “Use of the ’Perceptron’ Algorithm to Distinguish Translational Initiation Sites in E. Coli”. In: *Nucleic Acids Research* 10.9, pp. 2997–3011. DOI: 10.1093/nar/10.9.2997. PMID: 7048259.

Stormo, G. D. (2000). “DNA Binding Sites: Representation and Discovery”. In: *Bioinformatics* 16.1, pp. 16–23. DOI: 10.1093/bioinformatics/16.1.16.

Strahl, B. D. and C. D. Allis (2000). “The Language of Covalent Histone modiÆcations”. In.

Stumpf, M., X. Yue, S. Schmitz, H. Luche, J. K. Reddy, and T. Borggrefe (2010). “Specific Erythroid-Lineage Defect in Mice Conditionally Deficient for Mediator Subunit Med1”. In: *Proceedings of the National Academy of Sciences* 107.50, pp. 21541–21546. DOI: 10.1073/pnas.1005794107.

Sundararajan, M., A. Taly, and Q. Yan (2017). *Axiomatic Attribution for Deep Networks*. Version 2. DOI: 10.48550/ARXIV.1703.01365. URL: https://arxiv.org/abs/1703.01365 (visited on 09/28/2025). Pre-published.

Tahiliani, M., K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, and A. Rao (2009). “Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1”. In: *Science* 324.5929, pp. 930–935. DOI: 10.1126/science.1170116.

The Cancer Genome Atlas Research Network (2013). “Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia”. In: *New England Journal of Medicine* 368.22, pp. 2059–2074. DOI: 10.1056/NEJMoa1301689.

Tie, F., R. Banerjee, C. A. Stratton, J. Prasad-Sinha, V. Stepanik, A. Zlobin, M. O. Diaz, P. C. Scacheri, and P. J. Harte (2009). “CBP-mediated Acetylation of Histone H3 Lysine 27 Antagonizes *Drosophila* Polycomb Silencing”. In: *Development* 136.18, pp. 3131–3141. DOI: 10.1242/dev.037127.

Tsankov, A. M., H. Gu, V. Akopian, M. J. Ziller, J. Donaghey, I. Amit, A. Gnirke, and A. Meissner (2015). “Transcription Factor Binding Dynamics during Human ES Cell Differentiation”. In: *Nature* 518.7539, pp. 344–349. DOI: 10.1038/nature14233.

Tsuchiya, S., M. Yamabe, Y. Yamaguchi, Y. Kobayashi, T. Konno, and K. Tada (1980). “Establishment and Characterization of a Human Acute Monocytic Leukemia Cell Line (THP-1)”. In: *International Journal of Cancer* 26.2, pp. 171–176. DOI: 10.1002/ijc.2910260208.

Twist Bioscience (2022). *Twist Human Methylome Panel*. URL: https : / / www . twistbioscience.com/products/ngs/fixed-panels/human-methylome-panel.

Tyner, J. W., C. E. Tognon, D. Bottomly, B. Wilmot, S. E. Kurtz, S. L. Savage, N. Long, A. R. Schultz, E. Traer, M. Abel, A. Agarwal, A. Blucher, U. Borate, J. Bryant, R. Burke, A. Carlos, R. Carpenter, J. Carroll, B. H. Chang, C. Coblentz, A. d’Almeida, R. Cook, A. Danilov, K.-H. T. Dao, M. Degnin, D. Devine, J. Dibb, D. K. Edwards, C. A. Eide, I. English, J. Glover, R. Henson, H. Ho, A. Jemal, K. Johnson, R. Johnson, B. Junio, A. Kaempf, J. Leonard, C. Lin, S. Q. Liu, P. Lo, M. M. Loriaux, S. Luty, T. Macey, J. MacManiman, J. Martinez, M. Mori, D. Nelson, C. Nichols, J. Peters, J. Ramsdill, A. Rofelty, R. Schuff, R. Searles, E. Segerdell, R. L. Smith, S. E. Spurgeon, T. Sweeney, A. Thapa, C. Visser, J. Wagner, K. Watanabe-Smith, K. Werth, J. Wolf, L. White, A. Yates, H. Zhang, C. R. Cogle, R. H. Collins, D. C. Connolly, M. W. Deininger, L. Drusbosky, C. S. Hourigan, C. T. Jordan, P. Kropf, T. L. Lin, M. E. Martinez, B. C. Medeiros, R. R. Pallapati, D. A. Pollyea, R. T. Swords, J. M. Watts, S. J. Weir, D. L. Wiest, R. M. Winters, S. K. McWeeney, and B. J. Druker (2018). “Functional Genomic Landscape of Acute Myeloid Leukaemia”. In: *Nature* 562.7728, pp. 526–531. DOI: 10.1038/s41586-018-0623-z.

Uhlén, M., E. Björling, C. Agaton, C. A.-K. Szigyarto, B. Amini, E. Andersen, A.-C. Andersson, P. Angelidou, A. Asplund, C. Asplund, L. Berglund, K. Bergström, H. Brumer, D. Cerjan, M. Ekström, A. Elobeid, C. Eriksson, L. Fagerberg, R. Falk, J. Fall, M. Forsberg, M. G. Björklund, K. Gumbel, A. Halimi, I. Hallin, C. Hamsten, M. Hansson, M. Hedhammar, G. Hercules, C. Kampf, K. Larsson, M. Lindskog, W. Lodewyckx, J. Lund, J. Lundeberg, K. Magnusson, E. Malm, P. Nilsson, J. Ödling, P. Oksvold, I. Olsson, E. Öster, J. Ottosson, L. Paavilainen, A. Persson, R. Rimini, J. Rockberg, M. Runeson, Å. Sivertsson, A. Sköllermo, J. Steen, M. Stenvall, F. Sterky, S. Strömberg, M. Sundberg, H. Tegel, S. Tourle, E. Wahlund, A. Waldén, J. Wan, H. Wernérus, J. Westberg, K. Wester, U. Wrethagen, L. L. Xu, S. Hober, and F. Pontén (2005). “A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics”. In: *Molecular & Cellular Proteomics* 4.12, pp. 1920–1932. DOI: 10.1074/mcp.M500279-MCP200.

Vanyushin, B. F., S. G. Tkacheva, and A. N. Belozersky (1970). “Rare Bases in Animal DNA”. In: *Nature* 225.5236, pp. 948–949. DOI: 10.1038/225948a0.

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe (2009). “A Census of Human Transcription Factors: Function, Expression and Evolution”. In: *Nature Reviews Genetics* 10.4, pp. 252–263. DOI: 10.1038/nrg2538.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). *Attention Is All You Need*. Version 7. DOI: 10.48550/ARXIV . 1706 . 03762. URL: https : / / arxiv . org / abs / 1706 . 03762 (visited on 09/28/2025). Pre-published.

Viner, C., C. A. Ishak, J. Johnson, N. J. Walker, H. Shi, M. K. Sjöberg-Herrera, S. Y. Shen, S. M. Lardo, D. J. Adams, A. C. Ferguson-Smith, D. D. De Carvalho, S. J. Hainer, T. L. Bailey, and M. M. Hoffman (2024). “Modeling Methyl-Sensitive

Transcription Factor Motifs with an Expanded Epigenetic Alphabet". In: *Genome Biology* 25.1, p. 11. DOI: 10.1186/s13059-023-03070-0.

Vorontsov, I. E., I. A. Eliseeva, A. Zinkevich, M. Nikonov, S. Abramov, A. Boytsov, V. Kamenets, A. Kasianova, S. Kolmykov, I. S. Yevshin, A. Favorov, Y. A. Medvedeva, A. Jolma, F. Kolpakov, V. J. Makeev, and I. V. Kulakovskiy (2024). "HOCOMOCO in 2024: A Rebuild of the Curated Collection of Binding Models for Human and Mouse Transcription Factors". In: *Nucleic Acids Research* 52.D1, pp. D154–D163. DOI: 10.1093/nar/gkad1077.

Voss, K., G. V. D. Auwera, and J. Gentry (2017). *Full-Stack Genomics Pipelining with GATK4 + WDL + Cromwell*. DOI: 10.7490/F1000RESEARCH.1114634.1. URL: https://f1000research.com/slides/6-1381 (visited on 07/07/2025). Prepublished.

Wasserman, W. W. and A. Sandelin (2004). "Applied Bioinformatics for the Identification of Regulatory Elements". In: *Nature Reviews Genetics* 5.4, pp. 276–287. DOI: 10.1038/nrg1315.

Whalen, S., J. Schreiber, W. S. Noble, and K. S. Pollard (2022). "Navigating the Pitfalls of Applying Machine Learning in Genomics". In: *Nature Reviews Genetics* 23.3, pp. 169–181. DOI: 10.1038/s41576-021-00434-9.

Williams, K., J. Christensen, M. T. Pedersen, J. V. Johansen, P. A. C. Cloos, J. Rappsilber, and K. Helin (2011). "TET1 and Hydroxymethylcytosine in Transcription and DNA Methylation Fidelity". In: *Nature* 473.7347, pp. 343–348. DOI: 10.1038/nature10066.

Wilson, N. K., S. D. Foster, X. Wang, K. Knezevic, J. Schütte, P. Kaimakis, P. M. Chilarska, S. Kinston, W. H. Ouwehand, E. Dzierzak, J. E. Pimanda, M. F. De Bruijn, and B. Göttgens (2010). "Combinatorial Transcriptional Control In Blood Stem/Progenitor Cells: Genome-wide Analysis of Ten Major Transcriptional Regulators". In: *Cell Stem Cell* 7.4, pp. 532–544. DOI: 10.1016/j.stem.2010.07.016.

Wu, X. and Y. Zhang (2017). "TET-mediated Active DNA Demethylation: Mechanism, Function and Beyond". In: *Nature Reviews Genetics* 18.9, pp. 517–534. DOI: 10.1038/nrg.2017.33.

Wyatt, G. R. (1951). "Recognition and Estimation of 5-Methylcytosine in Nucleic Acids". In: *Biochemical Journal* 48.5, pp. 581–584. DOI: 10.1042/bj0480581.

Yin, Y., E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, K. R. Nitta, M. Taipale, A. Popov, P. A. Ginno, S. Domcke, J. Yan, D. Schübeler, C. Vinson, and J. Taipale (2017). "Impact of Cytosine Methylation on DNA Binding Specificities of Human Transcription Factors". In: *Science* 356.6337, eaaj2239. DOI: 10.1126/science.aaj2239.

Yokoyama, A. and M. L. Cleary (2008). "Menin Critically Links MLL Proteins with LEDGF on Cancer-Associated Target Genes". In: *Cancer Cell* 14.1, pp. 36–46. DOI: 10.1016/j.ccr.2008.05.003.

Yokoyama, A., I. Kitabayashi, P. M. Ayton, M. L. Cleary, and M. Ohki (2002). "Leukemia Proto-Oncoprotein MLL Is Proteolytically Processed into 2 Fragments with Opposite Transcriptional Properties". In: *Blood* 100.10, pp. 3710–3718. DOI: 10.1182/blood-2002-04-1015.

Yokoyama, A., M. Lin, A. Naresh, I. Kitabayashi, and M. L. Cleary (2010). "A Higher-Order Complex Containing AF4 and ENL Family Proteins with P-TEFb Facilitates Oncogenic and Physiologic MLL-Dependent Transcription". In: *Cancer Cell* 17.2, pp. 198–212. DOI: 10.1016/j.ccr.2009.12.040.

Yokoyama, A., T. C. Somervaille, K. S. Smith, O. Rozenblatt-Rosen, M. Meyerson, and M. L. Cleary (2005). "The Menin Tumor Suppressor Protein Is an Essen-

tial Oncogenic Cofactor for MLL-Associated Leukemogenesis". In: *Cell* 123.2, pp. 207–218. DOI: 10.1016/j.cell.2005.09.025.

Yokoyama, A., Z. Wang, J. Wysocka, M. Sanyal, D. J. Aufiero, I. Kitabayashi, W. Herr, and M. L. Cleary (2004). "Leukemia Proto-Oncoprotein MLL Forms a SET1-Like Histone Methyltransferase Complex with Menin To Regulate *Hox* Gene Expression". In: *Molecular and Cellular Biology* 24.13, pp. 5639–5649. DOI: 10.1128/MCB.24.13.5639-5649.2004.

Yu, B. D., J. L. Hess, S. E. Horning, G. A. J. Brown, and S. J. Korsmeyer (1995). "Altered Hox Expression and Segmental Identity in Mil-mutant Mice". In.

Yu, M., G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, and C. He (2012). "Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome". In: *Cell* 149.6, pp. 1368–1380. DOI: 10.1016/j.cell.2012.04.027.

Zaret, K. S. and S. E. Mango (2016). "Pioneer Transcription Factors, Chromatin Dynamics, and Cell Fate Control". In: *Current Opinion in Genetics & Development* 37, pp. 76–81. DOI: 10.1016/j.gde.2015.12.003.

Zaret, K. S. and J. S. Carroll (2011). "Pioneer Transcription Factors: Establishing Competence for Gene Expression". In: *Genes & Development* 25.21, pp. 2227–2241. DOI: 10.1101/gad.176826.111.

Zhao, S., C. D. Allis, and G. G. Wang (2021). "The Language of Chromatin Modification in Human Cancers". In: *Nature Reviews Cancer* 21.7 (7), pp. 413–430. DOI: 10.1038/s41568-021-00357-x.

Zhou, J. and O. G. Troyanskaya (2015). "Predicting Effects of Noncoding Variants with Deep Learning–Based Sequence Model". In: *Nature Methods* 12.10, pp. 931–934. DOI: 10.1038/nmeth.3547.

Zhou, Z., Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu (2023). *DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome*. Version 2. DOI: 10.48550/ARXIV.2306.15006. URL: https://arxiv.org/abs/2306.15006 (visited on 11/28/2024). Pre-published.

Zhu, H., G. Wang, and J. Qian (2016). "Transcription Factors as Readers and Effectors of DNA Methylation". In: *Nature Reviews Genetics* 17.9, pp. 551–565. DOI: 10.1038/nrg.2016.83.

Ziemin-van der Poel, S., N. R. McCabe, H. J. Gill, R. Espinosa, Y. Patel, A. Harden, P. Rubinelli, S. D. Smith, M. M. LeBeau, and J. D. Rowley (1991). "Identification of a Gene, MLL, That Spans the Breakpoint in 11q23 Translocations Associated with Human Leukemias." In: *Proceedings of the National Academy of Sciences* 88.23, pp. 10735–10739. DOI: 10.1073/pnas.88.23.10735.

Appendices

A Data preparation

A.1 Cell lines used

Cell-line	Disease	Genotype
RCH-ACV	B-ALL	E2A-PBX1
RS4;11	B-ALL	MLL-AF4
SEM	B-ALL	MLL-AF4
MV4-11	AML	MLL-AF4
THP-1	AML	MLL-AF9
OCI-AML3	AML	NPM1c+ and DNMT3A mutation

Table A.1: Cell lines and patient samples used

A.2 6-letter sequencing data

Cell-line/patient	Methylation mark	Prepared by
patient-22620	5mC & 5hmC	CC
patient-23003	5mC & 5hmC	CC
patient-26754	5mC & 5hmC	CC
patient-863388	5mC & 5hmC	CC
patient-9422	5mC & 5hmC	CC
RCH-ACV	5mC & 5hmC	CC
RS4;11	5mC & 5hmC	CC
SEM	5mC & 5hmC	CC

Table A.2: 6-letter sequencing data: Catherine Chahrour (CC)

A.3 CUT&Tag data

Cell-line/patient	Antibody	Prepared by
RCH-ACV	BRD4 Bethyl (A301-985A)	CC
RCH-ACV	BRG1 Bethyl (A300-813A)	CC
RCH-ACV	CTCF Cell Signaling (3418)	CC
RCH-ACV	ELF1 Abcam (ab64937)	CC
RCH-ACV	ENL Bethyl (A302-268A)	CC
RCH-ACV	ERG Cell Signaling (97249)	CC
RCH-ACV	H3K79ME3 Cell Signaling (9733)	CC
RCH-ACV	H3K4ME1 Diagenode (pAb-194-050)	CC
RCH-ACV	H3K4ME3 Diagenode (c15410003-A8034d)	CC
RCH-ACV	MED12 Bethyl (A300-774A)	CC
RCH-ACV	MENIN Bethyl (A300-105A)	CC
RCH-ACV	RNA pol2 Santa Cruz (sc899-k2712)	CC
RCH-ACV	PU1 Cell Signaling (2258S)	CC
RCH-ACV	RUNX1 Active Motif (39000)	CC
RCH-ACV	RUNX2 Cell Signaling (8486)	CC
RCH-ACV	TET2 Bethyl (A304-247A)	CC
RCH-ACV	H3K27AC Diagenode (C15410196)	AS
RCH-ACV	MLL-N Bethyl (A300-086A)	AS
MV4-11	MLL-N Bethyl (A300-086A)	AS
OCI-AML3	MLL-N Bethyl (A300-086A)	RM
RS4;11	MLL-N Bethyl (A300-086A)	AS
SEM	H3K27AC Diagenode (C15410196)	AS
SEM	MLL-N Bethyl (A300-086A)	AS
THP-1	MLL-N Bethyl (A300-086A)	AS
patient-22620	RUNX1 Active Motif (39000)	AS
patient-22620	MLL-N Bethyl (A300-086A)	AS
patient-23003	MLL-N Bethyl (A300-086A)	AS
patient-863388	MLL-N Bethyl (A300-086A)	AS
patient-9422	MLL-N Bethyl (A300-086A)	AS

Table A.3: CUT&Tag data: Catherine Chahrour (CC), Alastair Smith (AS), Rebecca Maynard (RM)

A.4 ChIPmentation data

Cell-line/patient	Antibody	Prepared by
patient-26754	MLL-N Bethyl (A300-086A)	AS

Table A.4: ChIPmentation data: Alastair Smith (AS)

A.5 ChIP-seq data

Cell-line/patient	Antibody	Prepared by
SEM	AF9	Milne Archive
SEM	ARID2	Milne Archive
SEM	BAF180	Milne Archive
SEM	BAZ2B	Milne Archive
SEM	BCL6	Milne Archive
SEM	BRD2	Milne Archive
SEM	BRD3	Milne Archive
SEM	BRD4	Milne Archive
SEM	BRG1	Milne Archive
SEM	CDK6	Milne Archive
SEM	CFP1	Milne Archive
SEM	CTCF	Milne Archive
SEM	CYCLINT1	Milne Archive
SEM	DOT1L	Milne Archive
SEM	E2A	Milne Archive
SEM	ELF1	Milne Archive
SEM	ELL2	Milne Archive
SEM	ENL	Milne Archive
SEM	ERG	Milne Archive
SEM	EZH2	Milne Archive
SEM	FLI1	Milne Archive
SEM	GCN5	Milne Archive
SEM	H3K36ME2	Milne Archive
SEM	H3K36ME3	Milne Archive
SEM	H3K4ME1	Milne Archive
SEM	H3K4ME3	Milne Archive
SEM	H3K79ME2	Milne Archive
SEM	H3K79ME3	Milne Archive
SEM	H3K9ME3	Milne Archive
SEM	JMJD1C	Milne Archive
SEM	JMJD2A	Milne Archive
SEM	JMJD6	Milne Archive
SEM	KDM2B	Milne Archive
SEM	KSRP	Milne Archive
SEM	LEDGF	Milne Archive
SEM	LEO1	Milne Archive
SEM	MAZ	Milne Archive
SEM	MED1	Milne Archive
SEM	MED12	Milne Archive
SEM	MED26	Milne Archive
SEM	MENIN	Milne Archive
SEM	NSD3	Milne Archive
SEM	OGT	Milne Archive
SEM	P300	Milne Archive
SEM	PAF1	Milne Archive
SEM	PHF6	Milne Archive
SEM	POL2SER5	Milne Archive
SEM	PU1	Milne Archive
SEM	RBBP5	Milne Archive

Continued on next page

Cell-line/patient	Antibody	Prepared by
SEM	RING1B	Milne Archive
SEM	RUNX1	Milne Archive
SEM	RUNX2	Milne Archive
SEM	RUNX3	Milne Archive
SEM	SATB1	Milne Archive
SEM	SMYD2	Milne Archive
SEM	SPT16	Milne Archive
SEM	SSRP1	Milne Archive
SEM	SUZ12	Milne Archive
SEM	TET2	Milne Archive
SEM	UTX	Milne Archive
RCH-ACV	BAZ2B	Milne Archive
RCH-ACV	BCL6	Milne Archive
RCH-ACV	DOT1L	Milne Archive
RCH-ACV	E2A	Milne Archive
RCH-ACV	EZH2	Milne Archive
RCH-ACV	P300	Milne Archive
RCH-ACV	PAF1	Milne Archive
RCH-ACV	PBX	Milne Archive

Table A.4: ChIP-seq data

B Software Environments

B.1 Notebook and Jupyter Packages

Package	Version
ipykernel	6.30
ipywidgets	8.1.7
jupyter	1.1.1
jupyterlab	4.4
notebook	7.4

Table B.1: Jupyter and notebook packages used in all environments

B.2 Environment: Crested

Channels: nvidia, bioconda, conda-forge

Package	Version
anndata	0.12.1
crested	1.5.0
cuda-version	12.9
keras	3.10.0
modisco-lite	2.4.0
pybigwig	0.3.22
pyranges	0.1.4
pysam	0.23.3
tensorflow	2.18.0
umap-learn	0.5.9.post2

Table B.2: Conda environment for Crested

B.3 Environment: Transformer Datasets

Channels: bioconda, conda-forge

Package	Version
anndata	0.12.1
crested	1.5.0
datasets	4.0.0
modisco-lite	2.4.0
pybigwig	0.3.22
pyranges	0.1.4
torch	2.7.1
transformers	4.54.1

Table B.3: Environment for transformer datasets

B.4 Environment: Transformer Training

Channels: bioconda, conda-forge, pytorch

Package	Version
accelerate	0.20.3
anndata	0.12.1
captum	0.8.0
cuda-version	11.7
cudatoolkit	11.7.0
datasets	2.19.2
evaluate	0.4.0
modisco-lite	2.4.0
peft	0.3.0
pybigwig	0.3.22
pyranges	0.1.4
pytorch	1.13.1
safetensors	0.6.2
tokenizers	0.13.3
transformers	4.29.2
umap-learn	0.5.9.post2

Table B.4: Environment for transformer training

B.5 Environment: PyTorch Tabular

Channels: conda-forge

Package	Version
anndata	0.12.2
captum	0.7.0
crested	1.5.0
cuda-version	12.9
keras	3.11.3
modisco-lite	2.4.0
pyranges	0.1.4
pysam	0.23.3
pytorch-lightning	2.4.0
pytorch-tabnet	4.1.0
pytorch-tabular	1.1.1
shap	0.48.0
torch	<2.6
umap-learn	0.5.9.post2

Table B.5: Environment for PyTorch Tabular

B.6 Environment: Methylation Model

Channels: bioconda, conda-forge

Package	Version
anndata	0.11.4
keras	3.11.3
modality	0.16.1
py-xgboost	3.0.5
pyranges	0.1.4
shap	0.48.0
xgboost	3.0.5

Table B.6: Environment for methylation model

B.7 Environment: XGBoost

Channels: conda-forge

Package	Version
anndata	0.11.4
pyranges	0.1.4
shap	0.48.0
torch	2.7.1
xgboost	3.0.2

Table B.7: Environment for XGBoost

C Supplementary Tables

C.1 GEO metadata for CUT&Tag samples

library name	title	organism	cell line	cell type	ChIP antibody	molecule	single or paired-end	instrument model	description	processed data file	processed data file	raw file	raw file
CAT-RCHACV_BRD4	CAT-RCHACV_BRD4	Homo sapiens	–	–	BRD4	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_BRD4_lanceotron.bed	CAT-RCHACV_BRD4_deeptools.bigWig	CAT-RCHACV_BRD4_1.fastq.gz	CAT-RCHACV_BRD4_2.fastq.gz
CAT-RCHACV_BRG1	CAT-RCHACV_BRG1	Homo sapiens	–	–	BRG1	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_BRG1_lanceotron.bed	CAT-RCHACV_BRG1_deeptools.bigWig	CAT-RCHACV_BRG1_1.fastq.gz	CAT-RCHACV_BRG1_2.fastq.gz
CAT-RCHACV_CTCF	CAT-RCHACV_CTCF	Homo sapiens	–	–	CTCF	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_CTCF_lanceotron.bed	CAT-RCHACV_CTCF_deeptools.bigWig	CAT-RCHACV_CTCF_1.fastq.gz	CAT-RCHACV_CTCF_2.fastq.gz
CAT-RCHACV_ELF1	CAT-RCHACV_ELF1	Homo sapiens	–	–	ELF1	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_ELF1_lanceotron.bed	CAT-RCHACV_ELF1_deeptools.bigWig	CAT-RCHACV_ELF1_1.fastq.gz	CAT-RCHACV_ELF1_2.fastq.gz
CAT-RCHACV_ENL	CAT-RCHACV_ENL	Homo sapiens	–	–	ENL	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_ENL_lanceotron.bed	CAT-RCHACV_ENL_deeptools.bigWig	CAT-RCHACV_ENL_1.fastq.gz	CAT-RCHACV_ENL_2.fastq.gz
CAT-RCHACV_ERG	CAT-RCHACV_ERG	Homo sapiens	–	–	ERG	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_ERG_lanceotron.bed	CAT-RCHACV_ERG_deeptools.bigWig	CAT-RCHACV_ERG_1.fastq.gz	CAT-RCHACV_ERG_2.fastq.gz
CAT-RCHACV_H3K27ME3	CAT-RCHACV_H3K27ME3	Homo sapiens	–	–	H3K27ME3	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_H3K27ME3_lanceotron.bed	CAT-RCHACV_H3K27ME3_deeptools.bigWig	CAT-RCHACV_H3K27ME3_1.fastq.gz	CAT-RCHACV_H3K27ME3_2.fastq.gz
CAT-RCHACV_H3K4ME1	CAT-RCHACV_H3K4ME1	Homo sapiens	–	–	H3K4ME1	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_H3K4ME1_lanceotron.bed	CAT-RCHACV_H3K4ME1_deeptools.bigWig	CAT-RCHACV_H3K4ME1_1.fastq.gz	CAT-RCHACV_H3K4ME1_2.fastq.gz
CAT-RCHACV_H3K4ME3	CAT-RCHACV_H3K4ME3	Homo sapiens	–	–	H3K4ME3	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_H3K4ME3_lanceotron.bed	CAT-RCHACV_H3K4ME3_deeptools.bigWig	CAT-RCHACV_H3K4ME3_1.fastq.gz	CAT-RCHACV_H3K4ME3_2.fastq.gz
CAT-RCHACV_MAZ	CAT-RCHACV_MAZ	Homo sapiens	–	–	MAZ	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_MAZ_lanceotron.bed	CAT-RCHACV_MAZ_deeptools.bigWig	CAT-RCHACV_MAZ_1.fastq.gz	CAT-RCHACV_MAZ_2.fastq.gz
CAT-RCHACV_MED1	CAT-RCHACV_MED1	Homo sapiens	–	–	MED1	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_MED1_lanceotron.bed	CAT-RCHACV_MED1_deeptools.bigWig	CAT-RCHACV_MED1_1.fastq.gz	CAT-RCHACV_MED1_2.fastq.gz
CAT-RCHACV_MED12	CAT-RCHACV_MED12	Homo sapiens	–	–	MED12	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_MED12_lanceotron.bed	CAT-RCHACV_MED12_deeptools.bigWig	CAT-RCHACV_MED12_1.fastq.gz	CAT-RCHACV_MED12_2.fastq.gz
CAT-RCHACV_MENIN	CAT-RCHACV_MENIN	Homo sapiens	–	–	MENIN	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_MENIN_lanceotron.bed	CAT-RCHACV_MENIN_deeptools.bigWig	CAT-RCHACV_MENIN_1.fastq.gz	CAT-RCHACV_MENIN_2.fastq.gz
CAT-RCHACV_POL2	CAT-RCHACV_POL2	Homo sapiens	–	–	POL2	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_POL2_lanceotron.bed	CAT-RCHACV_POL2_deeptools.bigWig	CAT-RCHACV_POL2_1.fastq.gz	CAT-RCHACV_POL2_2.fastq.gz
CAT-RCHACV_PU1	CAT-RCHACV_PU1	Homo sapiens	–	–	PU1	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_PU1_lanceotron.bed	CAT-RCHACV_PU1_deeptools.bigWig	CAT-RCHACV_PU1_1.fastq.gz	CAT-RCHACV_PU1_2.fastq.gz
CAT-RCHACV_RUNX1	CAT-RCHACV_RUNX1	Homo sapiens	–	–	RUNX1	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_RUNX1_lanceotron.bed	CAT-RCHACV_RUNX1_deeptools.bigWig	CAT-RCHACV_RUNX1_1.fastq.gz	CAT-RCHACV_RUNX1_2.fastq.gz
CAT-RCHACV_RUNX2	CAT-RCHACV_RUNX2	Homo sapiens	–	–	RUNX2	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_RUNX2_lanceotron.bed	CAT-RCHACV_RUNX2_deeptools.bigWig	CAT-RCHACV_RUNX2_1.fastq.gz	CAT-RCHACV_RUNX2_2.fastq.gz
CAT-RCHACV_TET2	CAT-RCHACV_TET2	Homo sapiens	–	–	TET2	genomic DNA	paired-end	Illumina NovaSeq X	–	CAT-RCHACV_TET2_lanceotron.bed	CAT-RCHACV_TET2_deeptools.bigWig	CAT-RCHACV_TET2_1.fastq.gz	CAT-RCHACV_TET2_2.fastq.gz

Table C.1: GEO metadata produced using SeqNado