
UNIVERSITY OF ICELAND

**Ph.D. Thesis**

**in Biochemistry**

# The Structural Dynamics of Sox2

## Deciphering the Role of Intrinsically Disordered Regions in Pioneer Transcription Factors

**Sveinn Bjarnason**

May 2024

**FACULTY OF LIFE AND ENVIRONMENTAL SCIENCES**

# The Structural Dynamics of Sox2

Sveinn Bjarnason

Dissertation submitted in partial fulfillment of a
*Philosophiae Doctor* degree in biochemistry

Advisor
Pétur Orri Heiðarsson

Ph.D. Committee
Magnús Már Kristjánsson
Erna Magnúsdóttir

Opponents
Sarah Shammas
Sebastian Deindl

Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, May 2024

The Structural Dynamics of Sox2
Deciphering the Role of Intrinsically Disordered Regions in Transcription Factors
Dissertation submitted in partial fulfillment of a *Ph.D.* degree in biochemistry



Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Sturlugata 7
102, Reykjavik
Iceland

Telephone: 525 4000

Bibliographic information:
Sveinn Bjarnason, 2024, *The Structural Dynamics of Sox2*, PhD dissertation, Faculty of Life and Environmental Sciences, University of Iceland, 162 pp.

Author ORCID: 0000-0001-8088-6590
Printing: Háskólaprent ehf.
ISBN: 978-9935-9769-4-9

# Abstract

This thesis explores the role of intrinsically disordered proteins and regions in the eukaryotic nucleus, focusing particularly on the pioneer transcription factor Sox2. Sox2 is well known for its crucial role in maintaining stem cell pluripotency and initiating gene expression changes critical for development and cellular reprogramming. Through a series of studies, we examined the dynamic interplay between Sox2, DNA, and nucleosomes. Employing advanced methodologies such as single-molecule Förster Resonance Energy Transfer and Nuclear Magnetic Resonance spectroscopy, integrated with molecular simulations, we delved into the complex changes in IDR ensembles and activation domain accessibility induced by DNA and nucleosome binding. Our findings reveal that the highly disordered C-terminal IDR of Sox2, has its conformational dynamics guided by charged intramolecular interactions and Sox2's interactions with DNA, causes substantial rearrangements in the IDR without affecting DNA binding affinity. Furthermore, we identified a specific site within Sox2's IDR that interacts directly with core histones, reduced rapid intrachain dynamics, and possibly influences nucleosome conformation. Using both Widom-601 and native-like nucleosome sequences, we demonstrate Sox2's ability to remodel nucleosomes and shed light on Sox2's capacity to displace histone H1. By unravelling the ensembles of Sox2 interacting with chromatin, this thesis contributes to the foundational knowledge that could lead to the development of targeted strategies to improve Sox2's reprogramming abilities.

# Útdráttur

Þessi ritgerð skoðar hlutverk ómótaðra próteina og svæða í kjarna heilkjörnunga, með áherslu á frumkvöðla umritunarþáttinn Sox2. Sox2 er þekkt fyrir hlutverk sitt í viðhaldi á fjölhæfni stofnfrumna og fyrir að hefja breytingar á tjáningu gena sem eru nauðsynlegar fyrir þróun og endurforritun frumna. Við skoðuðum hið dýnamíska samspil á milli Sox2, DNA og litnisagna. Með notkun háþróaðra aðferða eins og staksameinda FRET og kjarnsegulómum, ásamt tölvu hermunum, köfuðum við ofan í þær flóknu breytingar sem verða á aðgengileika ómótaðra virkjunarsvæða af völdum bindingar við DNA og litnisagnir. Niðurstöður okkar sýna að hreyfanleiki C-enda IDR Sox2 er stýrt af samskiptum innansameinda hleðslna og samskipti Sox2 við DNA veldur verulegri breytingu á mengi bygginga próteinsins án þess að hafa áhrif á DNA bindigetu. Auk þess greindum við stutta amínósýruröð innan ómótaðs svæðis Sox2 sem víxlverkar við histón próteinin í litnisögnum, dregur úr innankeðju hreyfanleika, og hefur mögulega áhrif á byggingu litnisagna. Með því að nota bæði Widom-601 og náttúrulegar litnisagna raðir, sýnum við hvernig Sox2 endurmótar litnisagnir og vörpum ljósi á eiginleika Sox2 til að fjarlægja histón H1. Með því að afhjúpa byggingarlegan breytileika Sox2 í samskiptum við litni, leggur þessi ritgerð grunn að þekkingu sem gæti leitt til þróunar á hnitmiðuðum aðferðum til að bæta endurforritunarhæfni Sox2.

*To my wife Sucharita,*

*thanks for keeping all my pieces together*

# Table of Contents

**Chapter 3**
**DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2**



**Chapter 4**
**Pioneer factor Sox2 remodels nucleosomes and displaces histone H1**

# List of Figures

# List of Tables

# List of Publications

Part of thesis:

**Sveinn Bjarnason**, Sarah F. Ruidiaz, Jordan McIvor, Davide Mercadante, Pétur O. Heidarsson, Protein intrinsic disorder on a dynamic nucleosomal landscape, Progress in molecular biology and translational science. **183**, 295-354 (2021)

**Sveinn Bjarnason**, Jordan McIvor, Andreas Prestel, Kinga S. Demény, Jakob T. Bullerjahn, Birthe B. Kragelund, Davide Mercadante, Pétur O. Heidarsson. DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2, Nature Communications. **15**, 1445 (2024)

**Sveinn Bjarnason**, Jens Nicolai Vilstrup Decker, Eliška Koutná, Kinga S. Demény, Vaclav Veverka, Pétur O. Heidarsson, Tentative title: Pioneer factor Sox2 remodels nucleosomes and displaces histone H1, Manuscript in preparation.

Not part of thesis:

Dhawal Choudhary, Laura Mediani, Mario J. Avellaneda, **Sveinn Bjarnason**, Simon Alberti, Edgar E. Boczek, Pétur O. Heidarsson, Alessandro Mossa, Serena Carra, Sander J. Tans, Ciro Cecconi. Human Small Heat Shock Protein B8 Inhibits Protein Aggregation without Affecting the Native Folding Process, Journal of the American Chemical Society. **145**, 15188-15196 (2023)

Daniele Montepietra, **Sveinn Bjarnason**, Kristinn Ragnar Óskarsson, Pétur O. Heidarsson, Ciro Cecconi, Giorgia Brancolini, 2024. Tentative title: Unveiling disrupted salt bridge networks in human HSPB8 mutant through molecular simulations and smFRET, Manuscript in preparation.

# Acknowledgements

I would like to extend my deepest gratitude to all those who have supported me throughout the journey of completing this thesis. First and foremost, I am incredibly thankful to my supervisor, Pétur Orri Heiðarsson. Without his grand ideas and his ability to follow through with them, none of this would have been possible. You introduced me to a whole new world of biochemistry and inspired me to do my best through it. Thank you!

I am also grateful to my doctoral committee members, Magnús Már Kristjánsson and Erna Magnúsdóttir, for showing such an interest in my work, for their many excellent questions and, for always being in my corner.

I would like to thank all of my colleagues in the Heidarsson group. Without you this journey would have been very dull. Our many evenings at the student cellar after a long day in the lab were a real balm for the soul. Special shoutout to Kiddi for our many “Sode Time” breaks, during which many of my best ideas came to me, both for the lab and for life in general. Last, but certainly not least, special thanks to my colleague and friend Kinga S. Demény for her invaluable assistance with everything nucleosome related.

I must also express my gratitude to my collaborators, their expertise, insight, and support have been invaluable to the success of this project. To each of you, thank you for being an integral part of this journey!

I must acknowledge the financial support I received from The Icelandic Research Fund, Doctoral scholarship fund of The University of Iceland, European Research Council and ML4NGP (COST).

Finally, I would like to thank my wife and best friend for her consistently wavering support, you really kept me on my toes! I have to thank my parents, Bjarni and Ragnheiður, and my siblings, Halli and Marín, for all their love and care. Then, I must thank my friends for all of their support through the years.

# 1 Introduction

## 1.1 The Emerging Significance of Intrinsically Disordered Proteins in the Human Proteome

Throughout most of the history of biochemistry, molecular and structural biology, and protein physics the prevailing belief has been that a proteins function arises from its stable three dimensional structure.[1] An understandable view, considering the challenges that have been overcome by analysing protein structures, both in fundamental research and pharmaceutical development. The Protein Data Bank[2] serves as a testament to this success with over 66,063 human protein structures to date. Recently, interest has been growing on proteins and domains that lack a stable three dimensional structure under physiological conditions[3]. These proteins range from fully intrinsically disordered proteins[4,5] (IDPs) to partly structured proteins that contain functional intrinsically disordered regions[6-8] (IDRs) (**Figure 1.1**). IDRs, and by extension IDPs, are characterized by an amino acid sequence that prevents a stable three dimensional structure.[9]

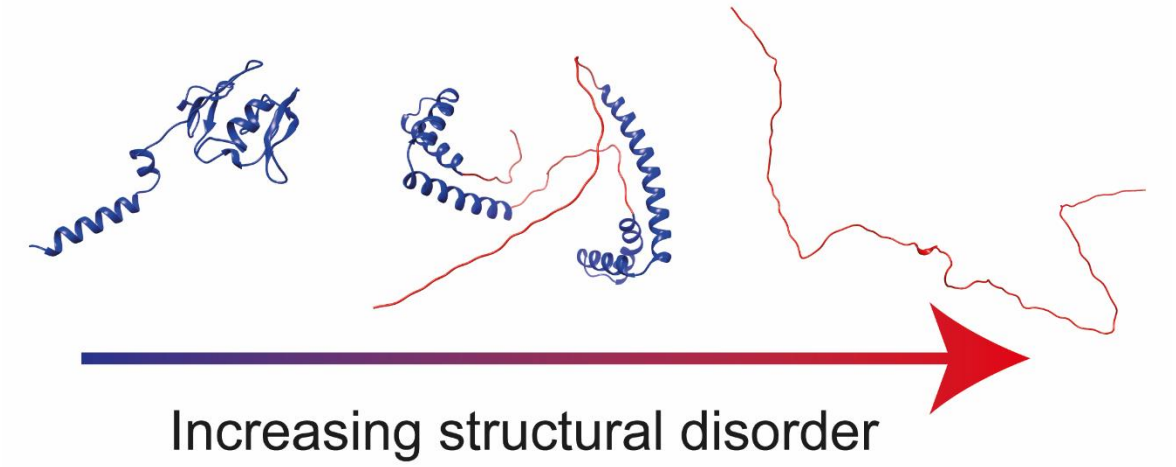



***Figure 1.1. Spectrum of protein structural disorder.*** *Displayed here are representative structures predicted by AlphaFold[10], showcasing varying degrees of protein organization. On the left is EPPIN, a protein which is highly structured, in the center is HMGB2, which features significant IDRs, and on the right is HMGA2, a prototypical IDP, representing the extreme end of structural disorder.*

Compared to folded proteins, IDRs have fewer hydrophobic residues and are enriched in low-complexity sequence repeats, charged and hydrophilic residues.[11] Despite the inability to form a stable fold, IDRs are essential for cellular function, with around 70% of proteins

in the human proteome predicted to contain one or more IDRs of 30 residues or more.[12] They are found in all cellular compartments performing roles in cell signalling, genome maintenance, immune surveillance, transcription and much more.[13-16] Rather than a stable three dimensional structure, IDRs can be better described as a collection of structurally distinct conformers that can rapidly interconvert[17], or simply put, by an ensemble of structures. The characteristics of such ensembles are quantifiable parameters that describe three dimensional features derived from the ensemble. These include parameters such as radius of gyration ($R_g$), inter residue distances and secondary structure propensity. But what purpose do IDRs serve and why has there been an evolutionary selection for these regions? IDRs have been demonstrated to serve various functions, but a common feature shared by many IDRs is their ability to enable multivalent molecular interactions that are challenging to achieve with folded domains.[12] In other words, IDPs are interaction masters, being able to capitalize on their structural disorder to form a variety of macromolecular interactions. These can be in the form of a fuzzy complex where disorder is preserved[18], folding upon binding where binding induces structural formation[19] and multivalent interactions that drive phase separation[20] to count a few. The interactions of IDPs need to be finely tuned as their dysregulation/dysfunction has been associated with several diseases[21,22]. IDPs have thus also created a need to adjust the process of drug discovery since they render the conventional structure-based drug design unfeasible.[23,24] For further reading on IDPs, their properties and functions, there exist several excellent reviews.[9,12,16]

## 1.2 Disorder is Highly Enriched in the Nucleus

The cell nucleus, a key organelle within the eukaryotic cell, serves as the epicentre of genetic information and regulatory control.[25] Within its double membraned boundary[26], it orchestrates the complex systems of gene expression, DNA replication, and chromosomal maintenance that underpins the functionality of life.[27,28] IDPs are highly enriched in the nucleus, with more than 90% of the proteins being predicted to contain significant amount of disorder.[29] This substantial enrichment of IDPs within the nucleus is clearly not a coincidence, but a reflection of the unique functions needed in the nucleus. The nucleus, being the hub of genetic regulation, necessitates a high degree of flexibility and adaptability, characteristics that IDPs embody.[30] Unlike their structured counterparts, the rapidly interconverting structures of IDPs allows them to engage in a diverse array of interactions with DNA, RNA, and other nuclear proteins.[31] This adaptability allows for a dynamic regulation of gene expression, where a rapid response to cellular signals is critical. IDPs can engage in multiple interactions simultaneously, often with high specificity but at low affinity, allowing for transient interactions to drive processes such as transcriptional regulation and chromatin remodeling.[32] Having established the role of IDPs in the nucleus, we can turn our sights to the cornerstone of chromatin organization: the nucleosome.

In the architecture of chromatin, nucleosomes emerge as the fundamental units, orchestrating the compact organization of DNA.[33] Each nucleosome is made up of a segment of DNA wound around a core of histone proteins, which are not just structural entities but also dynamic participants in gene regulation.[34] Central to this dynamic interplay are the IDRs found in the core histones (H3, H4, H2A, H2B), which play an important role in the nucleosome function (**Figure 2.1**).[35]

Histones, particularly through their IDRs, exhibit a remarkable capacity for post translational modifications and interactions with other nuclear factors.[36,37] These modifications and interactions are crucial for regulating DNA accessibility and, consequently, gene expression. The disordered regions in histones, characterized by their flexible and adaptable structure, allow for a diverse array of functional conformations.[38] This flexibility is key to understanding how chromatin structure modulates and responds to the cellular environment. A prime example of this functional versatility is histone H1.0, often referred to as the linker histone.[39] H1.0 differs from core histones in that it binds to the DNA between nucleosomes, influencing higher order chromatin structure. The presence of the disordered regions in H1.0 facilitates its ability to engage in various interactions[40], thereby impacting chromatin compaction and gene expression.

## 1.3 Transcription Factors Orchestrate Gene Regulation

The process of translating DNA into mRNA that is further translated into protein, is controlled by a vast network of around 1600 TFs in humans. Similar to histones, TFs consist of both structured DNA binding domains (DBD) and long IDRs, which confer upon them a flexibility essential for their regulatory roles.[13,41] To accurately regulate gene expression, TFs must bind at specific genomic locations. The ability to recognize specific DNA sequences was thought to arise solely from their DBDs[42], but recently there has been evidence that IDRs in TFs also influence DNA binding specificity.[43] However, despite the abundance of these specific DNA sequences in the genome, most remain unbound and unregulated by transcription factors.[44] This highlights an important balance in genomic regulation, the mere presence of these sequences throughout the genome does not guarantee binding. The selectivity of TFs is influenced by several factors such as chromatin context, the presence of other regulatory proteins and post translational histone modifications.[45] The histone modifications occur primarily on the histone tails and include methylation, acetylation, phosphorylation and ubiquitination, where each carries distinct regulatory signals.[46] In their quest to access specific DNA sequences TFs must therefore navigate a complex and dynamic chromatin landscape. The key to successful navigation lies in the ability of TFs to recognize not just the DNA sequence but also the chromatin context in which it resides. Some TFs are equipped with domains that can identify and bind to specific histone modifications[47], effectively "reading" the chromatin state. Others may recruit chromatin remodeling complexes[48], which can actively reposition, evict or restructure nucleosomes, altering the chromatin configuration to either expose or occlude DNA binding sites.[49]

Pioneer transcription factors (pTFs) differentiate themselves by their remarkable ability to access and bind DNA within tightly packed chromatin[50], a trait not commonly shared by all TFs. Unlike typical TFs that bind to DNA sequences made accessible through prior chromatin remodeling, pioneer factors have the unique capability to engage with condensed chromatin regions (**Figure 2.6A)**.[51] They possess the ability to bind directly to DNA sequences occluded by nucleosomes[52], initiating the first steps of chromatin opening.[53] This property allows pTFs to prepare chromatin for subsequent binding by other factors and activate otherwise inaccessible genes.[54] The mechanism by which most pTFs achieve this feat involves a combination of their DBDs and interactions with chromatin modifiers, which

often occur through the IDRs of the pTFs.[43] The action of pTFs are crucial in cellular processes such as development, differentiation and response to environmental changes, where precise and timely gene regulation is critical.[55,56]

The pTFs that perhaps best exemplify their role in cell differentiation are the so-called Yamanaka factors. Comprising Sox2, Oct4, c-Myc and Klf4, these pTFs have become renowned for their pivotal role in inducing pluripotency, a process that has revolutionized our understanding of cell differentiation and development.[57] The groundbreaking discovery by Shinya Yamanaka and his team revealed that these four TFs, when introduced into somatic cells, could reprogram them to become pluripotent stem cells. This process, known as cellular reprogramming, underscores the influence TFs exert on cellular fate and identity. Among the Yamanaka factors, Sox2 and Oct4 are particularly notable for their roles in maintaining stem cell pluripotency and regulating developmental genes.[58] These pTFs demonstrate an extraordinary ability to remodel the chromatin landscape of a cell. By binding to specific DNA sequences within chromatin, even those occluded by nucleosome rich regions, they can initiate a cascade of changes that erase somatic cell memory and establish a pluripotent state.[50]

## 1.4 Sox2: The Molecular Architect of Cellular Identity

Among transcription factors critical to cell fate and identity, Sox2 has emerged as a key player. A member of the Sox (sex determining region Y (SRY)-related high mobility group (HMG)-box) family of transcription factors, Sox2 shares a highly conserved DBD (80 residues) with 19 other family members.[59] The 20 Sox proteins can be further categorized into eight subgroups based on sequence identity and function, with Sox2 sharing the SOXB1 subgroup with Sox1 and Sox3.[60] In addition to their shared DBD, members of the SOXB1 family also share general composition features in their IDRs, such as the number and distribution of charges. This sequence similarity among SOXB1 members is associated with closely related biological activities, suggesting a degree of functional redundancy[61]. Like other TFs, members of the SOXB1 family are predicted to be mostly disordered.[62]

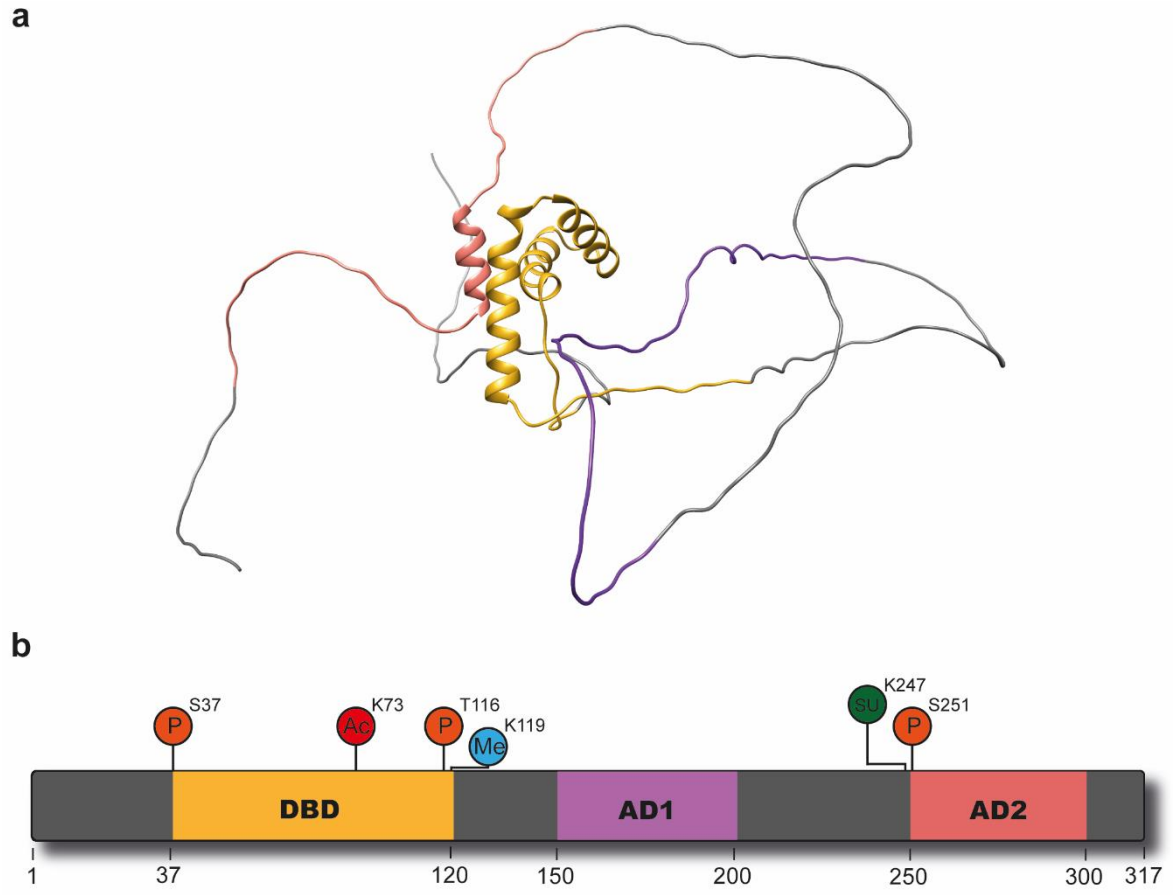



***Figure 1.2. Predicted structure of Sox2 by AlphaFold[10]. a)*** *The depicted model illustrates regions of very high model confidence in yellow (pLDDT > 90), signifying a robust prediction of Sox2's structure, while regions of lower confidence are highlighted in gray (pLDDT < 90), indicating areas where the structural conformation is less certain. Within these lower confidence regions, the activation domains AD1 and AD2 are colored purple and salmon, respectively. This emphasizes their significance despite the uncertainty in structural prediction.* ***b)*** *Model of Sox2 detailing some of the experimentally confirmed post-translational modifications found in Sox2[63,64], with P in orange corresponding to phosphorylation, Ac in red to acetylation, Me in blue to methylation and SU in green to sumoylation.*

In embryonic development, Sox2 is expressed as early as the 2-cell stage of murine embryos and is crucial for the formation of the pluripotent inner cell mass during early embryonic development.[65] Deletion of Sox2 leads to failure in embryoblast formation and early embryonic lethality.[66] As development progresses, Sox2 becomes restricted to specific cell populations, playing a vital role in determining cell fate, especially in the neural lineage.[67] Given the impact that Sox2 can have on cell identity, it should come as no surprise that Sox2 has its expression levels precisely regulated by a network of transcriptional[68], and post-translational systems.[63] Sox2 is subjected to various post-translational modifications, such as phosphorylation, SUMOylation, methylation, acetylation, ubiquitylation and more, which affect its activity, localization and stability.[69-71] Notably, a majority of these modifications are located on Sox2 IDRs, suggesting a potential for these regions to facilitate rapid responses to cellular signals and environmental changes.

Despite the presence of numerous regulatory systems, Sox2 can still become dysregulated, leading to serious consequences. Its overexpression has been correlated with poor survival rates in several types of cancers[72-74], underscoring its role as an oncogene. By driving tumor growth, enhancing metastatic potential, and contributing to chemotherapy resistance, Sox2 plays a critical part in cancer progression. However, the role of Sox2 in cancer prognosis of several cancers is not straightforward but Sox2 appears to exert a suppressive effect on tumorigenesis.[75] This dual role of Sox2 in cancer has sparked interest in targeting it for anticancer therapies.[76] Efforts to modulate Sox2 activity are underway, with several therapeutic strategies, ranging from small molecule inhibitors to T-cell immunotherapy.[77] There is however still no reliable anticancer therapy that targets Sox2. This is not unexpected as TFs have often been described as "undruggable".[78] This stems from their general lack of a well-defined pocket or surface that are amenable to small molecule binding, which are essential for traditional drug design. Drug design is further complicated by TF function involving protein-protein interactions through their IDRs[79], interactions that have proven difficult to disrupt, in part because most fuzzy interactions remain poorly characterized.

While the research of targeting Sox2 in cancer therapy continues, despite the many obstacles, the role of Sox2 in cell reprogramming and regenerative medicine has opened a new and promising frontier. As a key factor in the induction of pluripotency, Sox2 has been pivotal in the groundbreaking development of induced pluripotent stem cells[57] (iPSCs). This process, though revolutionary, is marked by a notable inefficiency.[80] This inefficiency is potentially influenced by several factors, including the variability in Sox2 expression levels, the interplay of Sox2 with other cellular components and TFs. Nevertheless, the process of reverting differentiated cells back to a pluripotent state has not only revolutionized our understanding of cell lineage and differentiation but also paved the way for innovative approaches in regenerative medicine. In tissue regeneration and wound healing, Sox2's role in cellular plasticity and differentiation is crucial.[81] Its involvement in promoting the regeneration of various systems, particularly those with inherently limited natural regenerative abilities, such as heart tissue[82], spinal cords[83] and auditory systems[84], highlights its potential as a therapeutic target. This could be achieved by manipulating the expression or function of Sox2.

Despite the array of potential applications of Sox2, ranging from regenerative medicine and production of iPSCs to cancer therapeutics, progress is hindered by a common bottleneck. The lack of a comprehensive structural description of Sox2. Apart from its structured DBD, there is a significant gap in our understanding of its structure, given that roughly 80% of Sox2 residues are predicted to be disordered. This absence of detailed structural data of Sox2, impedes biophysical studies and by extension targeted therapeutic interventions. A well-defined structural ensemble would enable researchers to elucidate the conformational dynamics of Sox2, shedding light on how it interacts with various molecular partners. This knowledge is crucial for designing molecules/mutations that can specifically modulate Sox2's function, whether for increasing iPSCs production efficiency or inhibiting its oncogenic potential.

## 1.5 Probing the Dynamic World of IDPs

Investigating the structural dynamics of IDPs, such as Sox2, requires a multifaceted approach using several techniques. Techniques like Nuclear Magnetic Resonance[85] (NMR)

spectroscopy, Small-Angle X-ray Scattering[86] (SAXS), Circular Dichroism[87] (CD) spectroscopy, coarse-grained simulations[88] and single-molecule Förster Resonance Energy Transfer[89] (smFRET), all have their place in the study of IDPs. NMR offers unrivalled insight into protein structure at the atomic level, providing information on the local environment and dynamics of individual amino acids.[90,91] SAXS, on the other hand, excels in providing low-resolution structural information, making it particularly useful when studying large and flexible proteins. It offers an overall shape and size of proteins, complementing the high resolution of smFRET and NMR[92]. CD spectroscopy contributes to our understanding of secondary structures and their conformational shifts. Coarse-grained simulations contribute to this array of techniques by offering a way to model systems over extended timescales. These simulations simplify the protein structure to focus on larger scale motions and interactions, providing a view of protein behaviour that can be difficult to capture with more detailed models.[93] Among these, smFRET stands out as one of the most powerful methods to study structural disorder.[94] smFRET has proved to be a powerful tool in the recreation of structural ensembles, offering the potential to unravel the conformational dynamics of IDPs at the single-molecule level.[95,96]

At its core, smFRET is based on the non-radiative transfer of energy between two fluorescent molecules, a donor and an acceptor, which occurs when they are in close proximity, typically within a range of 1-10 nanometres.[97] The fundamental principle of smFRET is captured in the Förster equation, which describes the efficiency of energy transfer between the donor and acceptor molecules. This efficiency is highly dependent on the distance between these molecules, making FRET an extremely sensitive method for measuring distances at the nanoscale. The Förster equation is given as:

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6}$$

where $E$ is the FRET efficiency, $r$ is the distance between the donor and acceptor and $R_0$ is the Förster radius, the distance at which the energy transfer efficiency is 50%. The Förster radius is specific to the pair of fluorescent molecules used and depends on factors like the overlap of the donor emission and acceptor absorption spectra, the quantum yield of the donor and the relative orientation of the donor and acceptor dipoles.[98]

In smFRET, these principles are applied at the level of individual molecules, allowing researchers to observe and measure dynamic processes and conformational changes in real time. Unlike ensemble FRET, which averages signals over a large population of molecules, smFRET provides detailed information about heterogeneity and transient states in biomolecules. By attaching donor and acceptor fluorophores to specific sites within a molecule, smFRET can reveal information about structural changes, folding dynamics and interactions with other molecules. The change in FRET efficiency as the molecule undergoes conformational changes provides insights into the distance and spatial relationships between different parts of the molecule, or between molecules.[99] Experimentally, FRET efficiency is determined by counting photons emitted and detected in the donor and acceptor detector channels.

$$E = \frac{n_A}{(n_A + n_D)}$$

Where $n_A$ and $n_D$ are the number of acceptor and donor photons, respectively. This quantification allows for the calculation of the energy transfer efficiency ($E$) and from it the distances and conformational states of biomolecules can be inferred with high precision.

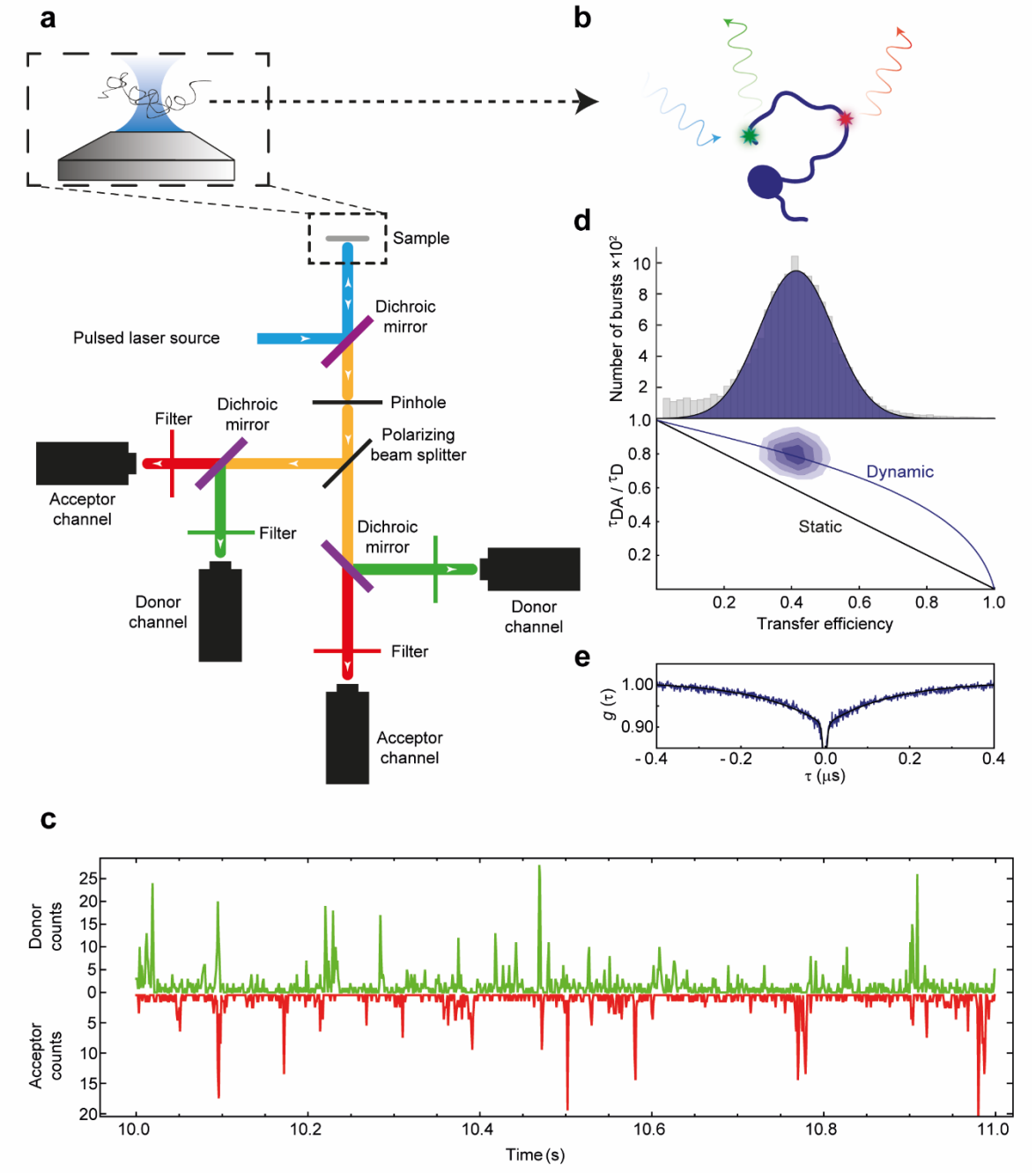



***Figure 1.3. Overview of instrumentation and data analysis for smFRET. a)*** *Schematic of a four-channel confocal single-molecule fluorescence microscope, designed to segregate and tally individual photon events by both polarization and wavelength. This setup enables the recording of arrival times for each photon.* ***b)*** *Depiction of a protein in solution, labelled with smFRET compatible fluorescent labels.* ***c)*** *A snapshot of a photon trajectory, aggregated into bins, capturing the fluorescence bursts from single protein molecules as they move through the confocal volume, each burst represents a single molecular encounter.* ***d)*** *A composite data representation consisting of a transfer efficiency histogram alongside*

*a two-dimensional histogram correlating the relative donor fluorescence lifetime, $\tau_{DA}$ /$\tau_D$, with transfer efficiency.* ***e)*** *Nanosecond fluorescence correlation spectroscopy (nsFCS) reporting on the nanosecond reconfiguration dynamics of the labelled protein. Figure adapted from "Single-molecule spectroscopy of protein folding dynamics—expanding scope and timescales", 2013, Current Opinion in Structural Biology, 23 (1), 36-47.*[100]

The primary advantages of single-molecule experiments include their ability to measure specific long range distances within and between molecules, differentiate between static and dynamic variations and study subpopulations simultaneously at equlibrium.[101] Another, often overlooked, advantage is the small sample volume and concentration used in smFRET measurements, generally in the range of tens to hundreds of pM in only a few µL of sample. This offers benefits beyond merely reducing the amount of protein needed, it also enables the study of proteins that are typically insoluble at concentrations required for other methods.[102]

Together, these methods create a comprehensive toolkit for probing structures and dynamics. smFRET's ability to measure distances and conformational changes at the single-molecule level is enhanced by NMR's atomic scale resolution and the macroscopic perspective offered by coarse-grained simulations. This integrative approach enriches our understanding of dynamic structural ensembles, with each technique compensating for the limitations of the others.

## 1.6 Objectives

The primary aim of this thesis is to recreate the structural ensembles of the pTF Sox2, in its free form, DNA bound and finally nucleosome bound. Specifically, it seeks to investigate the dynamic interactions between Sox2, DNA and nucleosomes, to understand how these interactions contribute to the regulation of gene expression. Through a combination of experimental approaches and computational modelling, this research aims to provide insight into the conformational dynamics of Sox2, particularly focusing on its IDRs and their significance in mediating chromatin remodeling. This study strives to uncover the molecular underpinnings of Sox2's pioneering activity on chromatin, with the goal of enhancing Sox2's reprogramming ability and contributing to the development of targeted cancer therapies.

The first objective of this thesis is to conduct a thorough review of IDPs within the nucleus, focusing on their roles, mechanisms, and implications for cellular function (**Chapter 2**). This objective aims to establish a foundation for understanding the complex roles of IDPs on the nucleosomal landscape, setting the stage for detailed investigations into specific IDPs like Sox2 in subsequent objectives. Through this review, we seek to highlight the importance of integrated approaches in the study of IDPs.

The next objective of this thesis is to dissect and compare the structural ensembles of Sox2 in both its unbound state and when bound to DNA (**Chapter 3**). We will employ advanced biophysical techniques, such as smFRET, NMR spectroscopy and computational modelling to recreate the structural ensembles, shedding light on the dynamic interplay between ordered and disordered domains of Sox2. By achieving this objective, we aim to provide a detailed molecular level insight into the structural dynamics of Sox2, providing a platform for mapping the effects of mutations and binding partners.

The final objective of this thesis is to recreate Sox2's structural ensembles when bound to nucleosomes, probing for interactions with the core histones (**Chapter 4**). Furthermore, we will assess Sox2's ability to remodel chromatin by displacement of histone H1, elucidating Sox2's role in transcriptional activation. By fulfilling this objective, the research will advance our understanding of the complex role Sox2 plays in chromatin dynamics and gene expression regulation.

# 2 Chapter 2

## Protein intrinsic disorder on a dynamic nucleosomal landscape

Sveinn Bjarnason, Sarah F. Ruidiaz, Jordan McIvor, Davide Mercandante, Pétur O. Heidarsson

# About this chapter

This chapter was published as a review in Volume 183 of "Progress in Molecular Biology and Translational Science"[103]. In it we highlight the latest developments in the study of IDPs focusing on nuclear proteins, outlining the challenges these proteins present for study, and explore various methods to investigate them.

# Author contribution

S.B. wrote the subchapter "Chromatin remodelers and histone-modifying enzymes" and made substantial, contributions to sections of other subchapters. Additionally, S.B. either produced or assisted in generation of figures and their captions. S.F.R., J.M., D.M. and P.O.H. wrote the rest of the manuscript and produced the remaining figures. All authors revised the text and contributed to error correction.

# Keywords

Chromatin; intrinsically disordered proteins; histones; chromatin remodelers; transcription; integrative experimental and simulation approaches nucleosome.

## 2.1 Abstract

The complex nucleoprotein landscape of the eukaryotic cell nucleus is rich in dynamic proteins that lack a stable three-dimensional structure. Many of these intrinsically disordered proteins operate directly on the first fundamental level of genome compaction: the nucleosome. Here we give an overview of how disordered interactions with and within nucleosomes shape the dynamics, architecture, and epigenetic regulation of the genetic material, controlling cellular transcription patterns. We highlight experimental and

computational challenges in the study of protein disorder and illustrate how integrative approaches are increasingly unveiling the fine details of nuclear interaction networks. We finally dissect sequence properties encoded in disordered regions and assess common features of disordered nucleosome-binding proteins. As drivers of many critical biological processes, disordered proteins are integral to a comprehensive molecular view of the dynamic nuclear milieu.

## 2.2 Introduction

As organisms become increasingly complex, so too must they evolve a more sophisticated molecular alphabet. The recent discovery of proteins that can adopt multiple structural states is one way of addressing this complexity and it has dramatically changed our view of the protein structure-function paradigm.[104,105] Intrinsically disordered proteins (IDPs) either do not contain any well-defined secondary structure element or have long unstructured regions (IDRs), and they fluctuate between a multitude of isoenergetic structural states. These proteins, which comprise an estimated third of the human proteome[106], are particularly prominent in the nucleus where as much as 70% have been shown or predicted to be IDPs.[29] The cell nucleus, which encompasses the genetic material, is a complex and moldable nucleoprotein landscape, shaped by frequent epigenetic changes that regulate the pattern of gene expression, and ultimately the organismal phenotype. It is thus unsurprising that a multivalent and dynamic nuclear proteome is needed to steer such a diverse environment. The conformational plasticity mediated by intrinsic disorder has been suggested to provide additional levels of functionality to complex cellular regulatory mechanisms.

In this chapter, we highlight protein disorder in the nucleus and emphasize the interplay between IDPs and the nucleosomal landscape that leads to a functional output. We first define the general components of the nuclear environment, before discussing the challenges and recent advances in understanding structural disorder within the context of transcription. We then compile and dissect a subset of important molecular systems in the nucleus that involve disordered interactions, including the effects of chemical modifications, and overview the resulting biological consequences. We exclusively review the interactions of structural disorder within nucleosomes and chromatin, but for reviews on IDP interactions with nucleic acids, we refer to excellent work on those topics.[16,107,108] Deciphering the complexity of molecular disorder in the chromatin landscape is an exceedingly challenging task. Yet, recent work has begun to map the functions of many constituent proteins of the nucleus by using innovative biophysical strategies, moving us ever closer to a comprehensive molecular view of the cell nucleus.

## 2.3 Protein intrinsic disorder on a nucleosomal landscape

### 2.3.1 Components of the nuclear environment

The importance of IDPs and IDRs in cell biology is now well established, and their prevalence in signaling and regulatory pathways has been clearly demonstrated.[16] It is their unique conformational properties that make them ideally suited for their roles. High

structural heterogeneity, a consequence of their low complexity and biased amino-acid sequences[109], imparts IDPs with multivalency in many cases, allowing them to interact with more than one biomolecular partner.[110] Even though the presence of disordered proteins in the nucleus has been recognized for decades, it is only relatively recently that their functions have surfaced. IDPs, which have sometimes been called constituents of the dark proteome[111], are now increasingly being illuminated as key players in the nucleus of eukaryotes.

To appreciate the many roles played by IDPs and IDRs in the nucleus, we first need to clearly define the nuclear architecture that they operate within (**Figure 2.1**). The genetic material for a typical human cell is composed of ~4.6 million basepairs of DNA, which contain the instructions for generating the cell´s proteome. The DNA is substantially compacted to fit this enormous amount into the relatively tiny nucleus, and at all stages of DNA compaction we encounter dynamic protein disorder in one form or another. The first level of compaction is to wrap the DNA around an octamer of the core histones (H2A, H2B, H3, and H4) containing two copies of each, forming the nucleosome.[112] A chromatosome is then constructed by binding of linker histone H1 (H1), which attaches to the dyad of a nucleosome (**Figure 2.1**).[113] Both the core and linker histones contain a large amount of disorder regulating nucleosome structure and dynamics and ultimately impacting global

chromatin structure.[114] There are many histone variants, some cell- or tissue-specific, that can be

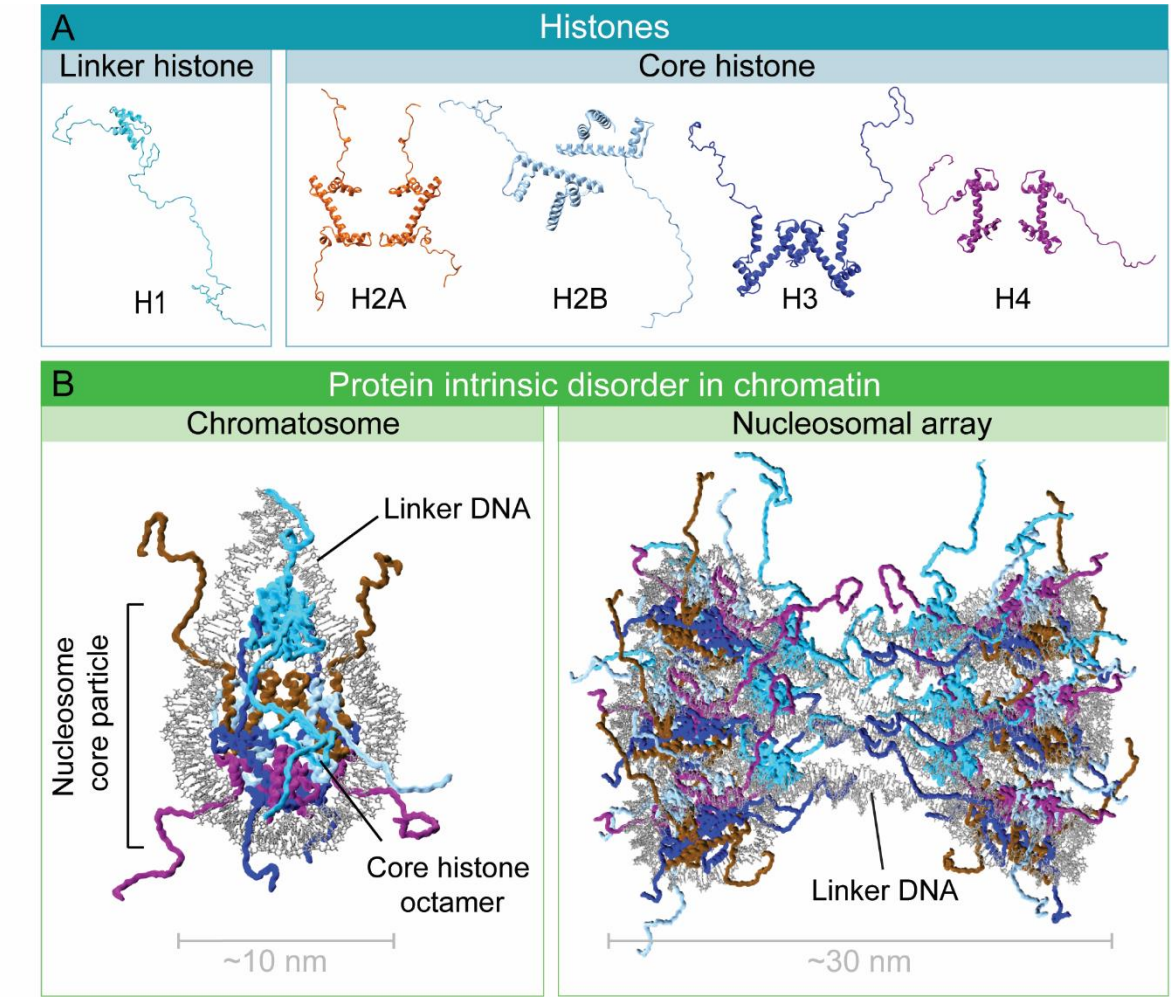



**Figure 2.1.** ***Protein disorder on a nucleosomal landscape.*** *Intrinsic disorder is a large component of the nucleosomal landscape, contributing to chromatin architecture, dynamics and overall function. A-B) The nucleosome core particle (NCP) is composed of an octamer of core histones (H2A, H2B, H3, and H4), around which ~147 bp of DNA (grey) is wound in a left-handed super-helical manner.[115] Within the NCP, H2A (orange), H2B (light blue), H3 (dark blue) and H4 (magenta) homodimerize via interactions in the structured domains, while the intrinsically disordered N- and C-terminal regions extend into the local nucleosomal space. Linker histone H1 (cyan) binds on or close to the nucleosomal dyad, forming the chromatosome (B) and uses its long disordered and highly basic C-terminal domain to drive conformational changes in linker DNA, impacting the overall structure of poly-nucleosomal arrays and ultimately chromatin fibers.*

dynamically exchanged to impart nucleosomes with distinct structural properties.[116] H1 rapidly exchanges between nucleosomes on the second to minute timescale *in vivo*[117], using largely its positively charged and disordered C-terminal tail to drive orientational changes in linker DNA connecting adjacent nucleosomes.[118] Local interactions between nucleosomes, involving the disordered histone regions, and binding of various regulatory

proteins modulate nucleosomal structure and dynamics (recently reviewed in[119]), and subsequently chromatin condensation into higher-order structures. Protein disorder thus plays an integral role in the formation, regulation, recognition, and modification of genome architecture.

### 2.3.2 Post-translational modifications fine-tune disordered interactions

To add yet another layer of complexity, most IDPs are chemically and reversibly modified after translation from the ribosome[120]. Histones and their variants have multiple post-translational modification (PTM) sites, mostly in their IDRs, where the pattern and number of modifications can fine-tune their interactions with nucleosomes and other biomolecules.[121] In general, PTMs render the proteome far more vast than the genome, with hundreds of thousands or even up to a million chemically distinct proteins at any given time in the cell.[122] Chemical modifications can change stability, concentration, localization, conformations, and interaction patterns of proteins, providing an important form of regulation and signaling. The most common modifications include (but are not limited by) covalent yet reversible chemical additions such as phosphorylations, acetylations, methylations, hydroxylations, and amidations, as well as attachments of sugar moieties or entire proteins involving sumoylation or ubiquitinylation.[120] In addition to protein modifications, DNA can be modified, most commonly involving cytosine methylation, and when located in CpG islands on promoters, this covalent modification is normally associated with gene repression.[123] Together, these modifications form an almost unfathomably complex and constantly evolving molecular surroundings that dictate the state of a cell.

Protein PTM sites are frequently located in IDRs, partly due to their accessibility to modifying enzymes such as kinases, acetylases, and methylases.[122] PTMs can induce or relieve secondary structure propensity or have a global effect on the structural ensemble sampled by the disordered region, potentially shifting the ensemble to a certain functional state, resembling conformational selection. They can also affect disorder-to-order transitions, which are a common interaction-mode for IDPs[124], or affect the degree of disorder in fuzzy[125] or fully[40] disordered complexes. PTMs that affect charges will influence intrachain electrostatic interactions, which have an important role in determining the compactness of a disordered region.[126] In general, PTMs modulate the structural and dynamical properties of IDPs, fine-tuning their functional repertoire. We now explore the arsenal of experimental and computational approaches that can and have been used to engage with IDPs, ranging from simple gel-based binding experiments to sophisticated atomistic models.

### 2.3.3 Challenges in studying disordered protein interactions with nucleosomes

Quantitative measurements of structurally heterogeneous polypeptides binding to the dynamic nucleoprotein landscape is a daunting task. Nonetheless, technological advances that enable access to various levels of molecular detail are continuously emerging.[127] As an initial characterization of protein-DNA interactions, classical binding experiments have often involved using an electrophoretic mobility shift assay (EMSA). EMSA is a simple and rapid way to monitor the binding of proteins (structured or disordered) to DNA by observing the changed migration pattern of DNA as a result of protein binding.[128] The EMSA can

provide information on binding affinity and specificity but may underestimate these parameters as during the electrophoresis the system is out of equilibrium. In addition, the

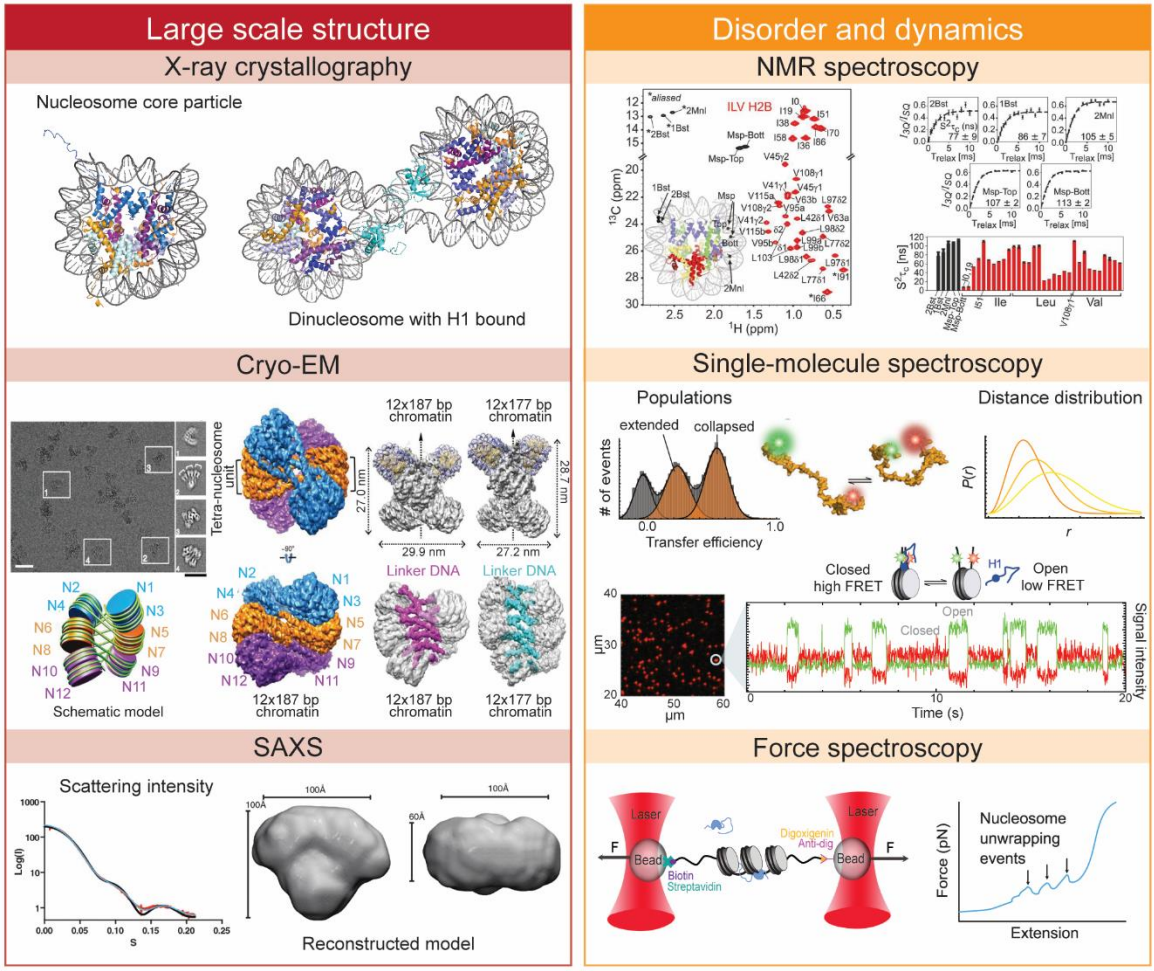



**Figure 2.2.** ***Methods to study chromatin and intrinsically disordered proteins.*** *Atomic-resolution structures can be determined from X-ray crystallography and Cryo-EM while SAXS gives lower resolution information on the overall dimensions of molecules. NMR spectroscopy yields both atomic resolution three-dimensional models of biomolecules and their inter- and intramolecular dynamics, as well as providing residue-specific information on protein-protein or protein-DNA interactions. Single-molecule techniques can be used to study heterogeneous conformational ensembles, at equilibrium, intra- and intermolecular distance distributions, and reaction kinetics. The X-ray crystal structures show the nucleosome core particle*[129] *and a dinucleosome*[130] *with bound H1 (PDB codes 1AOI and 6LAB). Cryo-EM, SAXS, and NMR data shown is reproduced from*[131-133] *with permission.*

EMSA does not give direct information on actual binding sites, i.e., it does not detect the exact base pair sequence which is recognized. Exact sequence with base-pair resolution can be determined using footprinting assays (e.g., hydroxyl radical footprinting[134]) or nuclease digestion (e.g., Micrococcal nuclease or MNase). Isothermal titration calorimetry (ITC) and more recently microscale thermophoresis, enable quantitative determination of protein-DNA binding affinity and specificity.[135] Chromatin immunoprecipitation, which relies on

chemical crosslinking of the target protein to DNA, combined with sequencing (ChIP-Seq[136]) is a powerful method to find protein binding sites *in vivo*. Similarly, ATAC-seq[137] (Assay for Transposase-Accessible Chromatin using sequencing) reveals genome-wide chromatin accessibility as a consequence of chromatin remodeling or other processes. In this elegant method a transposase is used to incorporate next-generation sequencing adapters into chromatin, which after sequencing provides a map of genome-wide chromatin accessibility. To understand local contributions from the polypeptide sequence, the beforementioned approaches can be combined with genetic and biochemical modifications of target proteins, such as introducing domain deletions/additions, charge reversal, domain swapping or local mutations. Still, without a view into microscopic molecular-level details, the underlying physical principles of protein function can be challenging to deconvolute. Cryogenic electron microscopy (cryo-EM) and X-ray crystallography enable determining atomic-resolution three-dimensional structures of macromolecules.[138] X-ray crystallography determines structures from diffraction patterns and it is the most widely used technique in structural biology.[139] Modern cryo-EM is rapidly catching up through recent advances in deep-frozen sample preparations, direct electron detection cameras and sophisticated image analysis[140], which take advantage of graphics processing units (GPU) acceleration. Recent studies using these methods have supplied us with an impressive view of large molecular assemblies, such as a translating ribosome[141] and entire chromatin fibers (Fig 2).[130,131,142] However, biomolecular processes involving extensive disordered interactions lie outside the scope of current structural biology efforts and thus require different approaches to understand their molecular underpinnings.

### 2.3.4 Integrative modelling of disordered protein interactions

Modern research on structurally heterogeneous systems such as IDPs often combines multiple techniques to decipher their underlying physical mechanisms (**Figure 2.2,2.3**). To study dynamic and disordered systems, techniques that can resolve conformational subpopulations in bulk have proven particularly useful. Nuclear magnetic resonance (NMR) spectroscopy can be used to obtain three-dimensional structural models of well-folded proteins, and it has also been extensively used to study protein dynamics and disorder[40,143,144], even in live cells.[145] After assignments of chemical shifts, protein NMR gives residue-specific information on structure, stability, binding sites, and dynamics on a wide timescale.[146] Despite still being limited to relatively small to medium-sized systems for structure determination, NMR spectroscopy has revealed dynamical movements of the disordered core histones and their interactions, even within entire nucleosomes (Fig 2).[133,147,148] Small-angle X-ray scattering (SAXS), the solution-state counterpart to X-ray crystallography, gives information on the shapes of molecules, including IDPs and their dynamic populations, often aided by computer simulations.[149,150] Especially relevant to DNA binding proteins, fluorescence recovery after photobleaching (FRAP) probes the mobility of fluorescently labeled proteins, inside the cell nucleus, and has been used to study the dynamic exchange of histone H1 between nucleosomes.[151] These and other methods have over the years been extraordinarily influential in shaping our perception of IDPs.

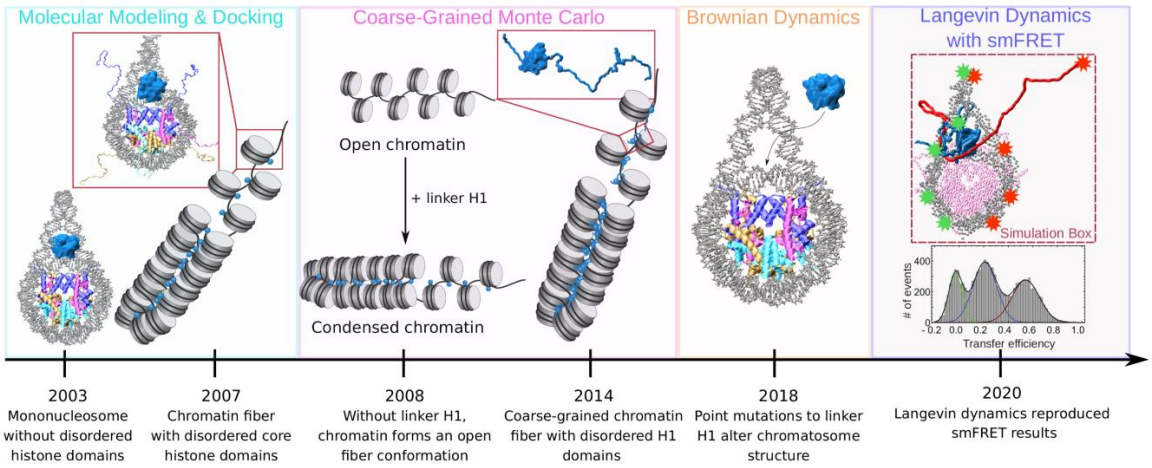



**Figure 2.3.** ***Chronological overview of computational approaches adopted to study intrinsic disorder in chromatin topology and dynamics.*** *Schematic illustration of the evolution of computational attempts used to investigate the interaction between proteins and DNA within the context of nucleosomes, chromatosomes and chromatin fibers. Early molecular modeling and docking investigated the binding of linker histone H1 to the nucleosome dyad and functioned as preliminary approaches to later attempts featuring molecular simulations on increasingly larger systems.*[152,153] *Coarse-grained Monte Carlo simulations were used to generate several possible topological arrangements of chromatin fibers with and without the linker histone H1.*[154,155] *Brownian dynamics*[156]*, which increases the amount of attainable sampling by scarifying internal motions, facilitated the understanding of how linker histone H1 diffuses towards and binds to the nucleosome dyad. Most recently, modeling and simulations featuring customized potentials, finely tuned to reproduce experimental findings, have provided a semi-quantitative overview of the disorder-mediated interactions between linker histone H1 and fully disordered chaperones involved in its displacement from the nucleosomal dyad.*[118]

Techniques that probe the behavior of individual molecules, and thus access molecular distributions, are an attractive approach to understanding disordered interactions and have been used to complement traditional ensemble methods. Single-molecule spectroscopy, usually in combination with Förster resonance energy transfer (smFRET), has emerged in recent years as an exceedingly powerful technique to study structured and unstructured proteins, *in vitro* and in living cells.[157-159] smFRET enables sensitive site-specific probing of the distance and dynamics between two or more fluorescent dyes, e.g., within a disordered region of a protein.[157] The rate of energy transfer between a donor and acceptor fluorophore is steeply dependent on the distance between them, where the useful range is typically on a convenient molecular scale of 1-10 nm. Importantly, smFRET can be applied to structurally heterogeneous systems one molecule at a time, avoiding the complication of ensemble-averaging, which can mask transient yet important molecular events. Confocal fluorescence microscopy offers a wide array of experiments that probe the thermodynamics and kinetics of biomolecular interactions through timescales covering fifteen orders of magnitude[160], as well as enabling high-resolution imaging in cells through fluorescence lifetime imaging (FLIM) and stimulated emission depletion (STED) microscopy.[161] Multi-parameter analysis of fluorescence intensity and photon timings allows quantitative investigation into molecular processes such as binding thermodynamics and kinetics, translational and intrachain

diffusion, complex stoichiometries, misfolding and aggregation.[159] With total internal reflection fluorescence (TIRF), several surface-immobilized molecules can be excited and detected simultaneously, offering higher-throughput data analysis of FRET trajectories.[159] The versatility of the method has over the years provided new insights into fundamental biological processes such as DNA maintenance and repair, signaling, translation, transcription, and molecular transport.[158,162] On the flip side of the single-molecule coin are force spectroscopy techniques, such as optical tweezers or the atomic force microscope (AFM), that allow direct tethering and manipulation of individual proteins or DNA. Force spectroscopy can probe the microscopic molecular forces involved in biomolecular interactions and has enabled a fresh view into the energetics and mechanisms of protein-nucleosome interactions.[163,164] Single-molecule methods hold great promise for understanding chromatin interactions and when combined with technologies probing ensemble biophysics[165] and genome-wide approaches, these methods can provide a comprehensive view of dynamic and disordered protein-DNA interactions.[127]

In recent years, a plethora of computational techniques have been used alongside experiments to study chromatin and chromatin binding proteins (**Figure 2.3**). By undertaking multiscale approaches, the finer molecular details of chromatin dynamics and interactions are now better understood. Early on, computational techniques were confined to molecular modeling and docking studies. All-atom molecular dynamics (MD) simulations are a gold standard technique for modeling biomolecular behavior, as they provide atomic-resolved information on the movements and interactions of molecules in their given environments[166]. By solving Newton's equations of motion for each atom, modeled as a van der Waals sphere, interatomic forces and their corresponding energies are calculated using molecular mechanics force fields describing both bonded and non-bonded interactions, either in implicit or explicit solvent conditions. In implicit solvent models, the solvent is treated as a structureless continuum, thereby reducing the number of interacting particles and degrees of freedom. In contrast to explicit descriptions, where the presence of each solvent molecule is explicitly accounted, implicit models do not include solute-solvent interactions. Although all-atom simulations provide an unparalleled level of detail, simulating nucleosomal arrays in this manner is unreasonable because of the high computational cost leading to insufficient sampling.

Due to the high computational costs associated with simulating these large and complex systems, an understanding of nucleosome and chromatin organization has started from simply creating models of single nucleosomes[167] or chromatosomes[152] that would fit experimental constraints. In particular, such studies focused on elucidating the binding mechanism of linker histone H1 to nucleosomes and were able to provide an idea, resembling that observed in electron microscopy studies[152], of how H1 and other histones shape the conformational dynamics of single and di-nucleosomes[167] as well as nucleosomal arrays composed of up to 100 nucleosomes[153], reporting on the polymorphic nature of chromatin.

A different approach involves the use of Monte Carlo (MC) simulations of coarse-grained (CG) representations. By generating conformational states according to Boltzmann probabilities, MC can be used to sample a Boltzmann distribution of configurations. Coarse-graining offers a computationally less expensive approach. In CG models, groups of atoms are embedded into beads, thereby reducing the degrees of freedoms and allowing efficient generation of conformations, without the explicit time dependence of MD simulations. Consequently, larger models, such as those encompassing entire chromatin fibers, can be

simulated yet at the expense of fine molecular details. MC simulations of CG chromatin fiber models have revealed many aspects of chromatin compaction. In particular, studies carried out using CG-MC found that H1 is required in the formation of higher order chromatin structures[155] and that, without H1, chromosomal arrays adopt an open fiber conformation[154]. Additionally, chromatin structures with highly variable nucleosome repeat lengths (NRL) produce more compact and uniform fibers[168], while fibers with a longer NRL, corresponding to a more open chromatin structure, are likely to have a higher number of binding sites for chromatin binding proteins[169].

Similar to MC, Brownian Dynamics (BD) simulations, which treat the simulated macromolecules as rigid bodies in implicit solvent, have been used to simplify the complexity of chromatosomes or chromatin. In BD simulations, the diffusion of solutes in a continuum solvent is simulated, and electrostatic interactions, which are particularly dominant in chromatin, are calculated by solving the Poisson-Boltzmann equation. BD approaches have successfully demonstrated the effects of sequence variation and PTMs on the H1-nucleosome ensemble. Due to lowering the complexity of the simulated system, docking simulations following the methodological paradigm of BD simulations, allowed a large number of PTMs to be considered, due to the increased computational efficiency associated with this technique.

Another approach to circumvent the computational constraints of all-atom MD is to coarse-grain (CG) the system. As discussed above, coarse-graining simplifies the system by reducing the degrees of freedom, making it an attractive technique for studies on chromatin fibers. In a recent study carried out by Watanabe et al., CG molecular dynamics simulations were used to create a model of an HP1α dimer bound to the histone H3 tail in a di-nucleosome complex.[170] In addition, reverse mapping of the CG structure was carried out upon the completion of the simulation to regain some of the lost atomic details. Within the context of CG models simulated by means of BD or MC, empirical potentials have added a certain level of integration with experimental data and specifically provided an experimentally-derived picture of the conformational ensembles of nucleosomes. Recently, smFRET and CG molecular simulations were tightly coupled in a complementary approach where the distance of multiple sites across protein-DNA within a nucleosome were mapped by smFRET and matched closely by the fine tuning of a single force field parameter describing van der Waals interactions between modeled beads. Integrative modeling has immense potential to deliver finer details of complex molecular systems and has already begun to uncover the physical principles governing IDP interactions.[165,171,172] We now move on to describe some recent work on disordered interactions with chromatin, highlighting new insights that biophysical methods have yielded on the role IDPs and IDRs in the nuclear environment.

## 2.4 Disordered interactions with nucleosomes

### 2.4.1 Nucleosome architectural proteins

#### Linker histone H1

Many proteins involved in generating and maintaining the overall nucleosomal architecture contain long disordered regions. H1 is involved in chromatin condensation through

stabilization of compact chromatin structures, and thus functions generally as a transcriptional repressor.[173,174] The polypeptide sequence is highly positively charged with two long disordered regions (N-terminal domain, NTD; C-terminal domain, CTD) flanking a small folded globular domain[175]. The globular domain of H1 is known to bind to the dyad axis of the nucleosome (Fig. 1), thus interacting with the nucleosomal core and both entry- and exit DNA linkers.[113,176] By binding to the nucleosome dyad, the H1 tails are free to form non-specific electrostatic interactions with linker DNA to minimize charge repulsion, thereby facilitating chromatin condensation. Therefore, although the binding mode is facilitated through the structured domain of H1, function is largely conferred through the disordered tails and an on-dyad binding mode may provide the freedom required for the H1 CTD to interact with one or both linker DNA arms. However, the conformational distributions of the disordered regions of H1 on the nucleosome have been more difficult to elucidate because of their pronounced dynamics and seeming lack of persistent structure.[177] Importantly, single point mutations on linker H1 significantly affect chromatosome structure, indicating that small changes may alter the overarching chromatin structure and, consequently, transcriptional regulation.[156] It has been suggested that H1 draws the two linker arms together, thereby reducing their mobility, and introducing a strong degree of asymmetry to the nucleosome[113]. Recent integrative studies of full-length H1 in complex with nucleosomes gave insight into the behavior of the long disordered tails.[118] The authors studied binding of human linker histone H1 to reconstituted nucleosomes using confocal single molecule spectroscopy and CG molecular simulations. Fluorescent labeling of the approximately 100 residue-long disordered CTD of H1 revealed that it becomes considerably more compact in complex with the nucleosome. This can be explained by screening of H1´s positive charges by the negatively charged nucleosomal DNA, which otherwise renders H1 highly expanded due to charge repulsion. Labeling on the terminal end of nucleosomal linker DNA arms and addition of unlabeled H1, resulted in the expected closure of the linker DNA arms in the H1-bound nucleosome (in agreement with a crystal structure of the chromatosome). Comprehensive mapping of FRET efficiencies within the H1-nucleosome complex combined with nanosecond fluorescence correlation spectroscopy (nsFCS) showed that H1 lacks persistent structure and is extremely dynamic on the nucleosome, displaying sub-μs chain reconfiguration times. Using a simple CG model describing the system in terms of non-specific short-range and electrostatic interactions, combined with existing structural information on the nucleosome and H1´s globular domain, the entire complex was simulated and the distances between the corresponding FRET pair locations back-calculated. After tuning the only free parameter in the model – the inter-bead interaction strength that was set globally for all beads – the simulation was able to capture the dynamic conformations of H1 on the nucleosome with high accuracy. Simulations were indeed in excellent agreement with the FRET-derived distances (a total of 60 FRET pairs), illustrating well how closely simulations can reconstruct experimentally determined parameters in even very complex systems.

The presence and conservation of multiple H1 variants within cells suggests that different variants may be linked to specific cellular functions. Within each variant, the structured globular domain shows the highest degree of conservation, while disordered tails are, expectedly, more variable. However, when H1 tail regions from different species are compared, a high degree of conservation is observed between orthologs. For example, human H1.4 and its mouse ortholog, H1e, share 93.5% sequence similarity[178], indicating that H1 tail regions may confer a high degree of functional selectivity in cells.

In addition to the highly basic charge in the linker H1 CTD, recent studies suggest a direct link between CTD length and chromatin affinity. FRAP experiments on human linker histones found that H1 variants with shorter CTD tails, such as H1.1 and H1.2, have rapid recovery times compared to variants with longer CTD tails, such as H1.4 and H1.5. Moreover, longer histone H1 tails were found to have two or more cyclin-dependent kinase (CDK)-dependent S/T-P-X-K phosphorylation motifs. Therefore, recovery times may be dependent on the density of lysine residues, the CTD length and the distribution of DNA-binding S/T-P-X-K motifs.[179] Because the H1 CTD directly interacts with the linker DNA, each variant will have a different effect on NRL. A higher degree of chromatin folding will likely be achieved if neutralization occurs across the chromatin fiber. Although chromatin condensation is not a direct consequence of linker histone binding, it does stabilize higher order chromatin structures to an extent that depends on the corresponding H1 variant. In general, the affinity of H1 for chromatin increases with its compacting properties. In agreement with this observation, H1.0, H1.4 and H1.5 have a longer CTD and were found to stabilize higher order chromatin.

## Post-translational modifications of linker histones

Linker histones are subject to a variety of PTMs, in both their IDRs, adding a large degree of compositional complexity to this protein family. The presence of multiple PTM sites in the H1 IDRs enables a number of regulatory mechanisms for H1 and finely regulates the affinity of each H1 variant for chromatin. The most prominent PTM for H1 is certainly phosphorylation, which occurs in a highly complex and dynamic fashion.[180] Phosphorylation mainly occurs in the CTD where S/T-P-X-K motifs (X is any amino-acid) are recognized by CDKs.[174] Although counterintuitive, H1 phosphorylation can trigger both chromatin expansion and contraction, based on the progression of the cell cycle. Such effects are likely to be a result of conformational rearrangements within H1, arising from site-specific modifications.[181] Phosphorylation levels are the lowest during the G1 phase, rise during the S phase and peak during mitosis, followed by a sharp decrease in the telophase.[115] *In vivo* studies showed that serine residues in H1.4 are generally modified during G1 and S phases, while threonine is phosphorylated in mitosis[182], outlining the cell cycle dependence of phosphorylation. CDK1 and Cyclin B are primarily responsible for H1 phosphorylation during the mitotic phase. However, recent studies suggested that several kinases phosphorylate the H1 NTD.[183-185]

The conversion of lysine into its methylated analogs (methyllysine, di-methyllysine or tri-methyllysine) is another important modification that can compete with or complement phosphorylation on a functional basis. For instance, the methylation of lysine 26 in the H1.4 NTD recruits heterochromatin protein-1 (HP1), resulting in heterochromatin formation[186], and is controlled by a phospho-switch: when H1.4 is phosphorylated at serine 17, the interaction between HP1 and methyllysine 26 is inhibited, demonstrating the importance of crosstalk between PTMs.[186] In the cell, H1 phosphorylation is CDK2 dependent and is required for progression through the S-phase. Because CDK2 colocalizes with replication sites and H1 is crucial in the formation of higher order chromatin, CDK2 recruitment to replication foci by Cdc45 may result in H1 phosphorylation and drive fork progression[187], linking H1 phosphorylation and active transcription. Methylation of the H1 NTD and CTD may elicit variant-specific cellular responses *in vivo*. For instance, the two predominant human H1 variants, H1.2 and H1.4[188], are methylated differently by the same methyltransferases.[189]

Interestingly, the methylation of H1.4 lysine 26, a highly conserved PTM in vertebrates, creates HP1 binding conditions, likely because H1.4 lysine 26 is part of an "ARKS" motif[186,189]: a conserved motif assumed to have a regulatory function in heterochromatin. Therefore, a link may be present between chromatin compaction and lysine 26 methylation.

Like phosphorylation, acetylation of the NTD leads to both heterochromatin formation and activation of transcription. Acetylation of H1.4 lysine 26 is related to the formation of facultative heterochromatin, which forms parts of the genome not shared across cell types and usually contains poorly expressed genes, which are task specific and mostly associated with cellular differentiation. Deacetylation by SIRT1, on the other hand, results in the formation of repressive heterochromatin.[190] The presence of an acetyl group on lysine 26 prevents methylation and subsequent recruitment of HP1, providing an additional level of regulation. Interestingly, an *in vivo* study using T47D cells expressing a lysine 26 to alanine H1.4 mutant, reported defects in gene regulation and cell proliferation, compared to wild-type H1.4.[191] Additionally, acetylation of H1.4 lysine 34 is also associated with transcription activation *in vivo*. In this position acetylation is, however, suggested to reduce H1-chromatin affinity and recruit TAF1; a subunit transcription factor TFIID.[192]

## Core histones

The core histones, which are the main structural support of nucleosomes, contain relatively short yet crucial IDRs when compared to the linker histone. Core histones form an octamer around which DNA is wrapped in the initial stages of chromatin condensation.[174] Each core histone shares a common histone-fold domain of three helices connected by two loop regions. To complete the octamer, each core histone homodimerizes, followed by the formation of specific H2A-H2B and H3-H4 heterodimers to create the "handshake" shaped core[193,194] (Fig. 1). In addition to the structured domains, each core histone has an intrinsically disordered, solvent-exposed N-terminal tail; with only H2A having an additional C-terminal tail.[195] Like H1, these tails are enriched with highly basic residues that form electrostatic interactions with nucleosomal DNA, linker DNA and acidic patches on neighboring nucleosomes. Such interactions are believed to stabilize the histone-DNA and nucleosome-nucleosome associations. Moreover, the dynamic nature of the tail regions makes them a target for PTMs and subsequent recruitment of histone chaperones, architectural binding proteins and chromatin remodelers.[196] The inter- and intramolecular contacts between the nucleosome and disordered core histone tails are crucial in the formation of the nucleosome core particle (NCP) and for the stabilization of higher order chromatin structures.

The intrinsically disordered tails of core histones have also been implicated in the formation of higher order chromatin structures and chromatin condensation through inter-nucleosomal interactions.[197] In the first nucleosome crystal structure published in 1997, an inter-nucleosome interaction between the N-terminal H4 tail and an acidic patch on the H2A/H2B dimer interface of a neighboring nucleosome was identified.[129] As for H1, the *in vivo* functions of core histones are complex and their ability to modulate transcription is largely dependent on their PTMs.[178] In the nucleosome, DNA accessibility is controlled by transient unwrapping from the NCP; a process that is modulated by each core histone to different degrees.[198,199] While the histone H3 tail suppresses nucleosome unwinding, the histone H4 tail enhances it.[200] Consequently, PTMs in the histone tails are especially important in their role of modulating protein-nucleosome interactions and regulating unwrapping.

Acetylation is an abundant modification in core histone tails, affecting chromatin compaction via the neutralization of positive charges.[121] Such effects have been demonstrated *in vitro*, where compaction of nucleosomal arrays required residues 14-19 in the NTD of the human histone H4[201] and the acetylation of lysine 16 prevents array compaction.[202] Moreover, acetylation is likely to reduce the electrostatic cross-talk between the DNA and the histone tails, decreasing the force required to unwrap nucleosomes[203] and increasing DNA accessibility to transcription factors and other modifying enzymes.[204] In line with this notion, histone acetylation has been shown to be strongly associated with transcription.[205] Additionally, many histone acetyltransferases interact with tri-methylated lysine 4 on H3; a modification associated with transcriptional activation.[121] Relevantly, *in vivo*, inhibition of transcription was shown to result in rapid histone deacetylation in mouse embryonic cells, indicating that much of histone acetylation occurs as a result of transcription.[206]

Phosphorylation of histones predominantly occurs in the intrinsically disordered NTD.[207] For instance, Aurora B kinase is known to phosphorylate H3 serine 10[208] and serine 28[209] during the mitotic phase *in vivo*, however, there is no evidence of both modifications being present on a single histone tail.[210] Nevertheless, phosphorylation of the histone H3 tails is likely required for cell cycle progression. Importantly, the Aurora B kinase is overexpressed in a number of human cancers[211], suggesting that phosphorylation plays a significant role in nucleosome availability to transcription.

Methylation is another predominant modification in core histone tails and for the cross-talk between different PTMs. For example, H4 arginine 3 methylation is recognized by p300; the acetyltransferase responsible for acetylation in histone H4. Methylation in this position is indeed an important PTM in gene transcription, as it is required for the subsequent acetylation that reduces electrostatic repulsion in the chromatin fibers.[212] In contrast, methylation of H3 arginine 8 is associated with gene repression, outlining the importance of site specificity in PTMs.[213]

## HP1 Proteins

Gene expression within the context of heterochromatin is facilitated by a series of important, conserved proteins called heterochromatin proteins (HP). These are fundamental units of chromatin packing that can be subdivided into families, with HP1 being the dominant family composed of three isoforms in humans – HP1α, HP1β and HP1γ –, all of which have two highly conserved structured domains; the amino-terminal chromo domain (CD) and the carboxyl chromo shadow domain (CSD). The structured domains are separated by a disordered hinge region (HR), of varying length across paralogs. Additionally, shorter intrinsically disordered extensions are present at the N- and C-termini of HP1.[214] HP1 is a major component of heterochromatin and is involved in the regulation of DNA-mediated processes including heterochromatin formation, stabilization of telomers and gene silencing in pericentric heterochromatin.[170,215]

In general terms, HPs are multivalent, structural chromatin effectors[216] that cause transcriptional repression by recognizing and binding di- or tri-methylated lysine 9 in histone H3 (H3K9me2/3) via the CD.[217], while remaining highly dynamic. Methylation of H3 provides an epigenetic mark, suitable for the hydrophobic binding pocket created by the CD. Despite the high degree of specificity between HP1 and H3K9me2/3, the binding affinity spans widely depending on the paralog.[218] Varying affinity is believed to provide a dynamic

range in which HP1 paralogs are able to elicit different cellular functions. The dynamic nature of HP1α has recently been probed using *in vitro* techniques. By employing a chemically defined assay, well suited to cellular measurements, it was found that HP1α residence time increases with H3K9me3 density, due to rapid re-binding of dissociated factors on neighboring sites. Moreover, dimeric HP1α exhibited accelerated association rates; a key feature of effector multivalency, allowing fast and efficient binding in a competitive environment.[216]

PTMs of HPs also play a fundamental role in regulating affinity. For instance, HP1α phosphorylation further strengthens its multivalency, while simultaneously reducing DNA binding, ultimately increasing HP1α residence time.[219] Phosphorylation of serine residues in the disordered NTD in mouse HP1α was found to increase CD-H3K9me2/3 affinity such that the overall affinity close to that of mouse HP1β and HP1γ.[218] Moreover, in HP1β and HP1γ the serine residues are replaced by glutamate. Such findings suggest that CD-H3K9me2/3 affinity may be partially modulated by charge differences in distal regions. Therefore, regulating charge via phosphorylation of the serine residues within the NTD in HP1α may contribute to the protein binding to H3. In turn, this interaction may impact the activity of kinases or phosphatases, increasing or decreasing binding.

Between the HP1 paralogs, both the underlying amino acid sequence and length of the HR are variable and such differences may control localization and function.[214] For instance, there are 41 and 36 residues in the HP1α and HP1β hinge regions, respectively, and both variants localize to heterochromatic regions[220,221] mediating transcriptional gene silencing.[222] Comparatively, HP1γ, where the HR has only 31 residues, localizes to euchromatin[220,221] and plays a role in transcriptional elongation and RNA processing[222]. The molecular basis for functional divergence is suggested to arise from the non-conserved residues in the HR, since the positively charged domains (KRK and KKK) are conserved across all three variants.[222] These domains are crucial for the specificity of HP1-H3K9me2/3 binding *in vitro*[223] and for intranuclear localization *in vivo.*[224] PTMs in the hinge regions have been shown to affect HP1 functionality.[214,225] Phosphorylation of serine 83 in the hinge region of HP1γ increases its interactions with Ku70, a DNA repair protein, thereby increasing its localization to euchromatin.[214] Taken together, this may suggest that modifications in the disordered HR of HP1 paralogs are able to elicit specific cellular functions.

Recent studies show that HP1 proteins play an important role in heterochromatin by interacting with histones H3 and H4 and methyltransferase enzymes.[207,226] The binding of the HP1 CD to poly-methylated H3 lysine 9 (H3K9me2/3) and H1.4K26me[186] triggers a silencing mechanism, resulting in the formation of heterochromatin.[217] Moreover, this interaction may be influenced by PTMs[227], especially those in the intrinsically disordered regions. In particular, phosphorylation of HP1α NTD poly-serine stretch 11-14, increases chromatin binding affinity by reducing tail flexibility in human and mouse cells.[218,228] Phosphorylation changes the conformation of the NTD, such that neighboring acidic residues (15E-DEE-E19) are able to interact with basic residues surrounding H3K9 (8R-Kme-STGGKAPR-K18).[229] Addition of the negatively charged residues formed upon phosphorylation results in repulsion, causing the HP1α NTD to behave as an extended intrinsically disordered region, in turn allowing the CD to dynamically bind H3K9me2/3.[229] Interestingly, in HP1β and HP1γ the residues corresponding to the poly-serine stretch of HP1α (12E-VL-E15 and 21K-VE-E24, respectively) are partially negatively charged.[218]

### 2.4.2 Intrinsically disordered proteins that interact or compete with linker histone H1

The state of chromatin compaction is tightly linked to the presence of H1. Therefore, the cell has evolved various regulatory mechanisms to actively remove H1 from nucleosomes. One such mechanism is proteins that compete with H1 for binding to the nucleosome or otherwise lead to its eviction. As mentioned above, H1 is highly disordered outside of the globular domain, a feature that is commonly shared among the diverse H1 competitors outlined here.

#### Protamines

Protamines are short (25-100 residues), highly basic, and disordered nuclear proteins[230,231] suggested to have evolved from histone H1.[232] Protamines replace core histones during the last stages of male germ terminal differentiation of spermiogenesis, where they are found to be the major packing units of DNA (Fig.4B).[233] Most mammals have only one gene coding for protamine 1 (PMR1 or P1) which is expressed in spermatids as a mature protein[234] and is responsible for chromatin condensation in sperm. However, some mammals, including humans and mice, have a second protamine, PMR2 or P2. Protamines from the protamine 2 family are longer compared to P1 and are generated by proteolytic cleavage of a precursor. DNA packed by protamines in mature sperm cells is transcriptionally inactive and forms higher order structures vital for normal sperm function.[233,235]

Protamine packaging of DNA has been studied with chemical and physical studies of both natural sperm chromatin and synthetic DNA.[236] Earlier studies demonstrated that protamines can precipitate DNA from both assembled chromatin and native chromatin extracted from calf thymus tissue in a concentration dependent manner. By analyzing the supernatant composition using gel electrophoresis after addition of protamine to a chromatin solution, H1 was discovered to be the first histone to appear in solution. However, the release of H1 was slower in native chromatin compared to reconstituted nucleosomes although somewhat affected by sample preparation, suggesting that the mechanistic picture of H1 competition and histone-protamine transition is more intricate.[237] Later work unveiled that protamines replace histones though a complex and progressive transition mechanism.[235] After meiosis in spermiogenesis, the canonical histones are replaced with testis-specific histones, and subsequently replaced by transition proteins that causes alteration in DNA structure. Many protamine molecules then bind and reorganize DNA into tightly packed structures. Thus, the conversion from histone-packed to protamine-packed chromatin is a complex interplay between different DNA-binders, regulated by PTMs including hyperacetylation of histones, in steps that are crucial for the correct progression of chromatin maturation and spermiogenesis.

The primary structure of protamines is characterized by a conserved arginine-rich core functioning as a DNA anchor[236] and cysteine-rich N- and C-termini. The overall dimensions of different protamines have been predicted by simulations to be controlled by their net charge per residue, which shifts their conformational ensembles from collapsed globule to coil-like.[238] The central arginine-rich region provides a high net positive charge that facilitates strong binding to DNA. Protamines wrap around DNA in the major groove[239] and bind with one protamine molecule per turn of DNA helix (**Figure 2.4**).[240] Even though protamines are known to be disordered when free in solution, it is challenging to probe conformational changes of the individual molecules within the context of the large chromatin structures because of the large number of protamines associating with chromatin. Potential

disorder-to-order transitions upon DNA binding are yet to be clearly demonstrated experimentally for protamines. However, many cysteines partake in multiple intra- and intermolecular disulfide bridges that provide rigidity, which is essential to stabilizing the structure of sperm cell chromatin.[241,242] Single-molecule studies demonstrated that protamines can bend DNA into loops through multiple steps[243] leading to the formation of higher order structures similar to those induced by H1 DNA packing[244], which suggests a common pathway for positively charged IDPs in chromatin condensation.

Protamines contain several conserved phosphorylation sites and have been shown to undergo various PTMs. These include phosphorylation, acetylation and methylation, which have been detected in protamines of mice sperm using peptide-based tandem mass spectrometry.[245] The data indicated that methylation, acetylation and phosphorylation do not occur at the same time on a single protamine, suggesting a complex network of PTMs affecting the epigenetic landscape of sperm cells.[245] However, the exact role of PTMs on the conformational ensembles of protamines and within the context of histone eviction in sperm remains unknown.

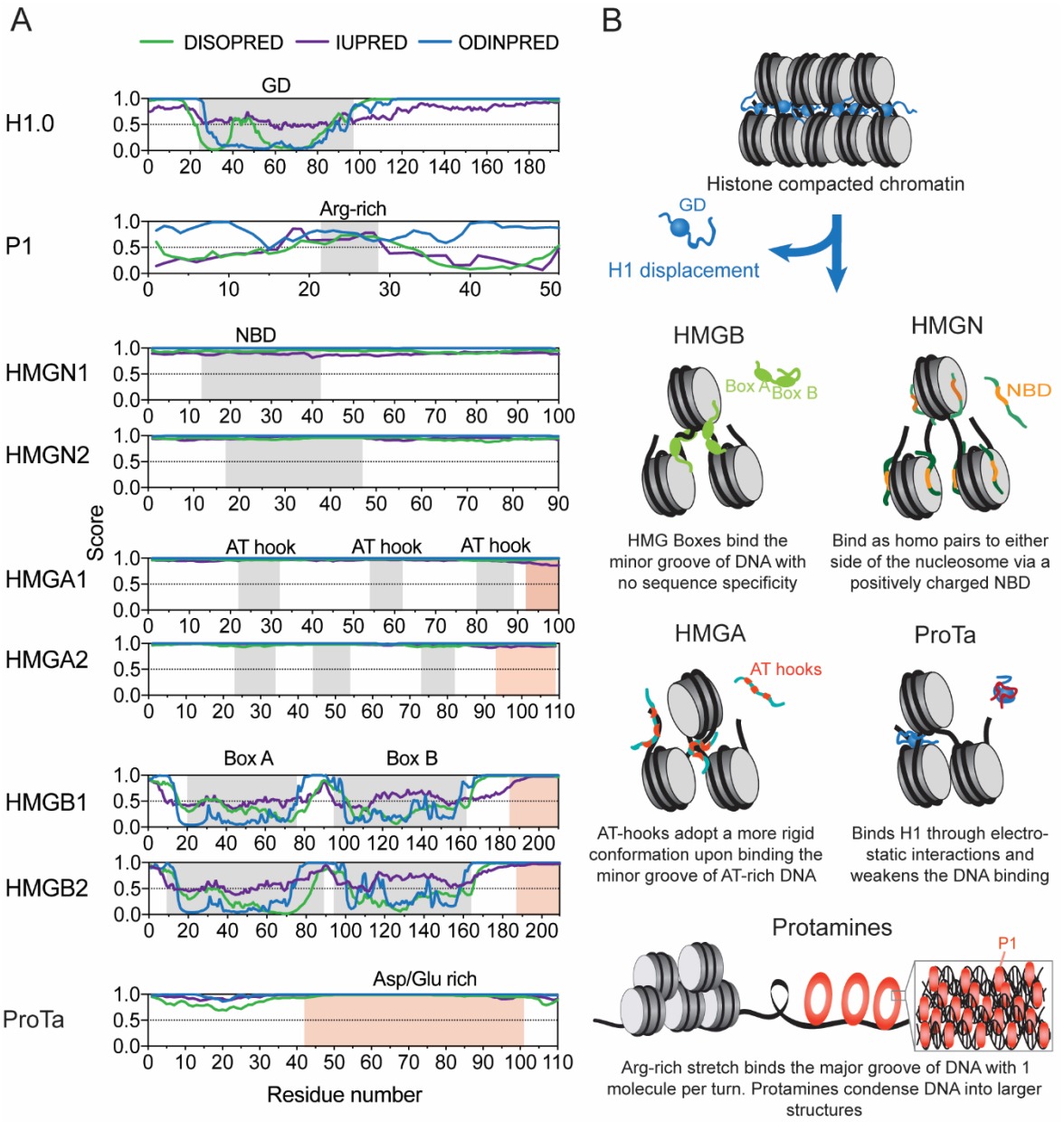



**Figure 2.4.** ***Disordered H1 competitors and their nucleosome binding modes.*** *A) Disorder predictions of H1 and competitors are shown using three predictors. Domains are highlighted in grey: Linker histone H1.0 contains a folded globular domain, HMGN1 and 2 each have an NBD, HMGA1 and 2 each contain three AT-hooks, and HMGB1 and 2 are characterized by two folded Box domains. Acidic stretches are indicated in red. B) Schematic illustration of the binding modes of H1 competitors in a nucleosomal context. The tightly packed structure of chromatin with bound H1 is remodeled by the disordered competitors, through eviction of H1. HMGB, HMGA, HMGN, ProTα and protamines all have distinct binding modes and their domains adopt different degrees of disorder essential to association with the nucleosomes.*

## HMG proteins

High-mobility group (HMG) proteins belong to a family of disordered architectural transcription factors known to interact with nucleosomes and, together with H1, were the

first nuclear proteins known to affect the structure of chromatin.[246] They were first discovered in isolated chromatin together with histones and named due to their unusually high electrophoretic mobility.[247] HMG proteins modulate local and global chromatin architecture by inducing formation of distorted DNA structures and promoting chromatin decompaction. The decompaction in turn enhances various DNA-dependent activities such as transcription, replication, and repair. HMG proteins are divided into three families– HMGA, HMGB and HMGN–, depending on their structural and functional properties, and we discuss here the role of their intrinsic disorder in nucleosome binding and H1 competition.[246,248]

***The HMGN*** subgroup has five members; HMGN1 and HMGN2 were the first to be discovered, whereas HMGN3-5 were identified later. HMGNs are fully disordered (**Figure 2.4A**) and bind the nucleosomal structure with high specificity in pairs to form complexes containing two molecules of either HMGN1 or HMGN2.[249-251] These proteins are characterized by a positively charged and conserved nucleosome binding domain (NBD), a nuclear localization signal (NLS), and an acidic C-terminal chromatin regulatory domain (CHUD) involved in modulating acetylation of histones.[248,252] HMGN proteins recognize generic nucleosome structures without specificity for DNA sequence or histones via the ~30 amino acid-long NBD.[253] The domain contains the canonical motif *RRSARLSA* which serves as an anchoring point on the nucleosome to a negatively charged patch formed by the H2A-H2B dimer surface.[254] The C-terminal domain of the HMGN protein interacts with the DNA in the two major grooves flanking the nucleosome dyad axis and is in close proximity to the N-terminal tail of histone H3.[250,255,256] HMGN also influences H3 phosphorylation and acetylation, by inducing local structural change that alters the accessibility of enzymes and thus the equilibrium of nucleosomal PTMs.[257,258] HMGNs are themselves modulated through phosphorylation[258], which decreases the affinity to chromatin and allows kinases to access histone H3.[257] HMGNs also have many lysines placed in the NBD that are acetylated, which leads to less efficient binding to nucleosomes.[259]

NMR studies using paramagnetic relaxation enhancement (PRE) experiments, which probe long-range interactions, and methyl-labeled histones in assembled nucleosomes, showed that, upon addition of HMGN, an interaction occurs between an arginine-rich region of the NBD and the folded core of the H2A/H2B dimer.[256] In addition, it was demonstrated that the several lysine residues in the C-terminal end of the NBD have an affinity for DNA non-specifically. The experimental data was then used as restraints to model the binding orientation of the NBD on the nucleosome. In the calculated structural model, the N-terminal end of the two NBDs stacked on each side of the nucleosome are predicted to bind the H2A/H2B acidic patch, while the lys-containing ends associate with DNA near the exit/entry point.[256] Due to the orientation of the NBD on the nucleosome, the disordered C-terminal tail was expected to be located where H1 associates with nucleosomes. This suggests that the chromatin decompaction function of HMGN is a result of disrupting H1 binding to DNA and the core histone tails.

Surprisingly, native gel-shift assays and cross-linking studies showed that HMGN1 binds nucleosomes already bound to H1, without interfering with the specific contacts made between the H1 globular domain and the nucleosome.[260] This observation agreed with previous MNase digestion results on mammalian chromatin, which reported nucleosomes bound to two HMGN proteins and H1.[261] This implies that binding of HMGN proteins to nucleosomes engenders a dynamic rearrangement of H1 interactions leading to modulation of chromatin structure. However, the molecular mechanism is still unclear and whether

HMGN folds upon binding to the nucleosomes or if the two bound HMGN remain fully unstructured in the complex, is yet to be determined.

***HMGB proteins (HMGB1 and HMGB2)*** have two structurally conserved DNA binding domains (DBDs,Box A and B) that each fold into three helices when in complex with DNA.[246] A disorder-to-order transition observed by NMR spectroscopy on HMG domains from Sox-proteins[262], demonstrated that HMG box domains retain a flexible structure in solution that folds upon DNA binding. In addition, HMGB proteins contain a disordered C-terminal region of ~30 amino acids (Fig 4A) enriched with acidic residues. The Box DBDs of HMGB bind with low affinity to single-stranded, linear duplex, and supercoiled DNA[263,264], but have a preference for bent and distorted DNA[265-267] which is increased upon acetylation.[268] Based on structures of closely related HMG-Box-DNA structures, the box domains distort DNA through intercalation of bulky hydrophobic amino acid residues into the DNA minor groove, resulting in bending the molecule towards its major groove (reviewed in[266]).

The acidic C-terminal region forms a flexible extended structure which was characterized by NMR spectroscopy and demonstrated narrow dispersion in $^{1}H$-$^{15}N$-HSQC typical for IDPs.[269] The flexible tail is involved in dynamic intramolecular interactions, with the highest affinity for DBD Box B. SAXS and NMR studies, including PRE measurements, suggested that the C-terminal tail promotes a more compact conformation where the two basic boxes get closer to each other.[270] Apart from intramolecular interactions of the acidic tail with the HMG-boxes, the acidic C-terminal tail of HMGB1 is also engaged in intermolecular interactions with other proteins. Notably the HMGB tail associates with histone H3, an interaction suggested to modulate the biological functions of HMGB proteins.[271] MNase digestion data suggested that HMGB protects linker DNA on one side of the NCP at the entry/exit of nucleosomes opposite to the linker histone H1 binding site.[272] As observed for other HMG proteins, early studies using chromatin fractionation experiments reported strongly enriched HMGB1 and 2 in H1-depleted fractions of salt-soluble chromatin.[273] FRAP experiments showed that HMGB also enhances H1 mobility in cells, indicating its ability to displace H1 from chromatin.[274] Studies of HMGB and H1 interactions by chemical cross linking and gel filtration experiments showed that they form a 1:1 complex. The complex persists at physiological ionic strength, where it was reported by NMR spectroscopy that H1 binds through its basic C-terminal domain to the acidic tail of HMGB1, disrupting its interaction with HMG boxes. A consequence of this interaction is enhanced DNA binding and bending by HMGB1, followed by a lowered affinity of H1 for DNA.[275] This might facilitate H1 eviction in a chromatin context and supports the data showing increased H1 mobility in cells in presence of HMBG.[274]

As outlined above, several lines of evidence support that the disordered C-terminus of HMGB1 has a crucial role in its function, including orchestrating many different interaction partners (review in[276]) and modulating chromatin structure. HMGB function is also regulated by PTMs like acetylation, phosphorylation, methylation, as well as its oxidative state as formation of a disulfide bridge in Box A leads to reduced H1 displacement from hemicatenated DNA loops.[277] Most PTMs found or predicted in HMGBs are placed in the folded box domains, but several acetylation sites have been identified in disordered stretches within NLS regions. Although acetylation within the Box domain is known to affect DNA binding affinity, more studies point towards PTMs having a large impact on controlling the nuclear localization and export of HMGBs. However, how PTMs affect the

conformational ensemble of the acidic disordered tail of HMGB and its interaction with H1 has yet to be fully defined.

The third class of highly disordered HMG chromatin binders are the ***HMGA proteins (previously named HMG-I(Y)).*** There are two genes coding for HMGA proteins, HMGA1 (and its splicing variants HMGA1a, b and c) and HMGA2, characterized by their very short DNA-binding AT-hook motifs. HMGA proteins are fully disordered in solution in absence of DNA as shown by biophysical techniques such as circular dichroism[278] and NMR spectroscopy.[279] An NMR study reported that the AT-hook DBD of HMGA transits from disordered to a well-defined crescent-shaped configuration upon binding to the minor groove of short AT-rich DNA stretches.[280] The specificity for AT-rich DNA regions was also demonstrated by a PCR-based systematic evolution of ligands by exponential enrichment (SELEX) approach, identifying nucleotide consensus sequences with two AT-rich stretches of 5-6 base pairs separated by four GC-rich base pairs.[281] EMSA studies show that HMGA binds isolated nucleosome particles with much higher affinity than to naked DNA[282,283]. It was previously suggested that HMGA1 also associates with core histones based on DNase footprinting and chemical cross-linking studies [184], possibly explaining the preference for nucleosomal DNA.

HMGA1 proteins were found to co-localize with histone H1 at AT-rich DNA stretches called scaffold attachment regions (SARs) in mammalian cells.[284] T7 polymerase assays in combination with DNA binding assays showed that HMGA can compete with H1 on SAR, and even redistribute H1 onto non-SAR DNA.[285] Purification of HMGA and H1 from HeLa cell chromatin also demonstrated that HMGA is strongly enriched in H1-depleted fractions of active chromatin.[285] In a study of chromatin condensation in neural precursor cells of mice, it was found that HMGA proteins are essential for chromatin opening in the early developmental stage. Overexpression of either HMGA1a or HMGA2 in cells increase the sensitivity to MNase digestion of chromatin from extracted nuclei, whereas depletion of HMGA mRNA led to reduction in MNase digested DNA. This clear effect of HMGA proteins on MNase digestion of extracted DNA from nuclei suggests that HMGA induce chromatin opening and accessibility.[286] The chromatin was more resistant to digestion in absence of HMGA, which supports the notion that chromatin becomes more accessible in presence of HMGA, suggesting an inhibition of H1 driven compaction. Another work observed increased H1 mobility caused by HMGA by measuring FRAP in cells expressing GFP-H1 and microinjected with purified HMGA into the cytoplasm.[274] The apparent H1 displacement in the different studies was due to competition for chromatin binding sites, since HMGA mutants incapable of binding DNA did not increase H1 mobility in a similar manner nor compete with H1 for SAR binding.[274,285] This strongly indicates that HMGA competes with H1 on chromatin resulting in destabilization of higher order chromatin structure.

Like the other HMG proteins, HMGA undergoes various PTMs which have been extensively studied. In fact, HMGA1 proteins are among the most phosphorylated proteins in the nucleus by the action of various kinases like cdc2, protein kinase C (PKC), and casein kinase II (CK2).[258] Phosphorylation of HMGA leads to considerably lower affinity towards DNA, in part because two of the main phosphorylation sites are near the positively charged AT hooks and thus disrupt their binding to the negatively charged DNA. The acidic C-terminal tail of HMGA also undergoes phosphorylations *in vivo* which can lead to a conformational change.[287] NOE measurements and 1D proton spectra of C-terminally phosphorylated HMGA indicated a more rigid structure compared to the native and free HMGA which is

fully disordered. Pull-down assays of truncated HMGA containing AT-hooks with the acidic C-terminal peptides showed a clear interaction to nucleosomes driven by electrostatics as the affinity increased with the number of phosphorylations in the C-terminal peptides. This supports the hypothesis that the phosphorylated acidic tail folds back onto positively charged clusters on the HMGA and through charge neutralization impairs binding mediated by these Arg/Lys-rich regions.

To summarize, all HMG proteins compete with H1 for chromatin binding sites (in a dose dependent fashion) although each HMG subfamily has distinct effects on the interaction of H1 with chromatin (**Figure 2.4**).[253] HMG proteins all contain disordered regions – HMGA and HMGN being completely disordered in absence of DNA – but they have different structural features and folded domains. This results in distinct modes of action in their modulation of chromatin structures, although a common feature appears to be recognition of DNA conformation. Even though the regions where the HMG proteins bind nucleosomes is known, the sequence of events leading to H1 eviction from nucleosomes is still not fully clear. It has also been suggested that the different classes of HMG proteins can weaken H1 binding cooperatively without competing with each other, hinting that they distinctly affect H1 binding to the nucleosome.[274] Overall, the HMG proteins are part of a dynamic and elaborate interaction network that leads to H1 displacement, where disorder plays a fundamental role.

### FoxA1

Forkhead box A (FoxA) transcription factor (previously called HNF-3) is part of a group of transcription factors evolutionary conserved in eukaryotes and is crucial in regulation of biological processes such as cell development, signal transduction, cell differentiation, and regeneration. FoxA1 is as a so-called pioneer transcription factor due to its ability to engage target sites on nucleosomal DNA.[288,289] FoxA1 is disordered outside of a highly conserved winged helix DBD which is structurally similar to that of histone H1.[290] The DBD contains a helix–turn–helix (HTH) motif that makes base-specific DNA contacts as well as two flanking loops (wings) that contact the phosphodiester backbone of DNA. FoxA1 is known to stably bind nucleosomes *in vitro* and *in vivo* near the nucleosome dyad[291,292] and decompact repressed chromatin compacted by H1 to make it accessible for other DNA binding factors.[293,294]

*In vitro* sequential binding experiments with purified proteins showed that FoxA1 displaces H1 prebound on assembled nucleosomes.[291] Further DNase footprinting of H1-compacted nucleosome arrays with and without FoxA1 demonstrated increased hypersensitivity in the digestion patterns and indicates that FoxA1 can open H1 compacted nucleosomes.[291,293] Truncation mutants of FoxA1 missing the 174 amino acid C-terminal domain failed to open the compacted arrays[293], underlining the importance of the disordered regions of FoxA1. Early studies also indicated that the N- and C-terminal regions of FoxA1 are crucial for binding specificity to nucleosomes over free enhancer DNA.[291] More recent work identified a short region in the C-terminus of FoxA1, conserved among FoxA pioneer factors, that interacts with core histones and contributes to chromatin opening *in vitro* (see **Figure 2.7**).[295] A single-locus study demonstrated that FoxA1 induction caused reduction of H1 occupancy at an enhancer site during retinoic acid-mediated differentiation of embryonic stem cells.[296] In later studies, an assessment of genome-wide occupancy of linker histone H1 in mouse hepatocytes showed FoxA occupancy on nucleosomes correlates with H1 displacement, whereas the FoxA deletion mutants had a striking increase in H1

disposition. All of these results indicate that FoxA binding displaces linker histones from the local chromatin, which could explain the subsequent increase in nucleosome accessibility and stimulation of transcription.

## Prothymosin α

The nuclear protein prothymosin α (ProTα) is a linker histone chaperone that modulates H1 interaction with nucleosomes. Besides affecting chromatin condensation[297] and H1 mobility in the nucleus[298], ProTα is involved in transcriptional regulation, cell proliferation, and apoptosis.[299] ProTα is fully disordered, with a highly negatively charged glutamate-rich (net charge -44) amino acid sequence and low hydrophobicity.[300,301] Borgia and co-workers used a combination of single-molecule FRET, NMR spectroscopy, and CG simulations to study the interaction between ProTα and histone H1, and showed that they form a tight complex with picomolar affinity yet remain highly disordered and dynamic in the bound state.[40] This novel interaction mode can be explained by the large opposite net charge of the two proteins which leads to complex formation through a mean-field type charge interaction without the need for defined binding sites or persistent interactions between specific individual residues. The CG simulations, which relied on a simple model involving non-specific short-range and electrostatic interactions, were able to reproduce the experimentally measured FRET efficiencies in the complex remarkably well. Later, Sottini et al. showed through an elegant set of kinetics experiments that ProTα and H1 can also form higher order but weakly interacting ternary complexes.[143] Again, integrating experiments and simulations, they showed that a second ProTα or H1 molecule can engage a preformed ProTα-H1 complex and lead to rapid exchange, keeping the system highly responsive despite the tight binding.

What is the purpose of forming such a disordered complex in the nuclear context? Heidarsson et al. addressed that question by studying the H1-ProTα interaction in the presence of reconstituted nucleosomes[118] (described also above, see section "Linker histone H1"). Kinetic experiments using immobilized and fluorescently labeled nucleosomes showed that ProTα forms a ternary complex with H1 and the nucleosome, which accelerates the dissociation of H1 by almost two orders of magnitude through a competitive substitution mechanism. Further CG simulations confirmed the dramatic increase in dissociation rate as a function of ProTα binding and provided a molecular picture of how ProTα invades the complex by dynamically and gradually sequestering the H1 C-terminal IDR. The high negative charge in ProTα thus competes with the electrostatic interactions between the linker DNA and the disordered regions of H1, which reduces the interaction strength of H1 with the nucleosome and leads to an opening of the nucleosome linkers. These results provide clues towards resolving long-standing issues on histone H1 including the nature of the structural ensemble of H1 on the nucleosome and the discrepancy between *in vivo* (minutes) and *in vitro* (hours) residence times of H1 on the nucleosome.[151,302] Through integrative modeling of these challenging molecules, the authors suggested that it is precisely the high degree of dynamic disorder on the H1 IDRs that allows chaperones like ProTα to invade the complex and accelerate the dissociation of H1 from the nucleosomes. For such unspecific, charge dominated binding between dynamic and disordered proteins, the formation of higher order complexes may commonly occur, providing additional functionality and enabling a sensitive concentration-dependent response during signaling. Formation of higher order oligomers and the dynamic exchange within them may be particularly important to achieve dissociation of strongly interacting polyelectrolytes[40], and to induce formation and regulation of phase-separated condensates.[107,303]

### 2.4.3 Chromatin remodelers and histone-modifying enzymes

Chromatin remodelers dynamically modify chromatin architecture to modulate access of the transcriptional machinery to DNA, and thus regulating gene expression.[304] Remodeling pathways are largely dependent on i) various covalent modifications of histone tails driven by ATP-independent factors[304] such as deacetylase (HDAC), methyl transferase (HMT), acetyl Transferase (HAT), ii) ATP-dependent chromatin remodeling complexes[305] which either slide, eject or restructure nucleosomes, and iii) chaperones that bind to histones and stimulate their transfer onto DNA or other proteins.[306] On the basis of their functions, chromatin remodelers can be roughly divided into two families: ATP-dependent enzymes that include imitation switch (ISWI), chromodomain helicase DNA binding (CHD), switch/sucrose non-fermentable (SWI/SNF) and INO80[305], and ATP-independent enzymes including the histone methyl/acetyl transferases, kinases, and isomerases. Despite differences in mechanisms and compositions, all ATP-dependent remodelers contain a structurally similar catalytic ATPase core which converts the chemical energy of ATP hydrolysis into conformational changes. Besides actively regulating gene expression, dynamic remodeling of chromatin imparts an epigenetic role in several key biological processes e.g., DNA replication and repair, apoptosis, and pluripotency.[307]

Chromatin remodelers have an extensive range of interacting partners. They can form multimeric complexes and interact with histones, transcription factors, nucleic acids, and various other machinery involved in the maintenance of chromatin structure.[308] Such a diverse range of interactions is difficult to explain with highly structured proteins. Predictions from amino acid sequence strongly suggest that chromatin remodelers contain substantial structural disorder[309,310], involved in forming stable complexes and transient interactions with diverse interacting partners, potentially playing a more direct functional role than acting as simple linkers.[311,312]

#### ATP-dependent chromatin remodelers

Many ATP-dependent chromatin remodeling complexes are predicted to contain IDRs.[313] These IDRs range from relatively small regions, likely functioning as linkers, all the way to the BRG1/BRM-associated factor (BAF) complex which is made up of subunits that are predicted to contain long IDRs.[310,314] A recent study looked at the predicted disorder in BAF and found that 27 of the 30 subunits that were analyzed were predicted to be highly disordered.[310] The BAF complex is among the most frequently mutated complexes in many types of cancer, many of which are located in predicted disordered regions.[315] While the function of the predicted IDRs remains largely unknown, they are likely to assist with binding to histones, nucleic acids, and transcription factors.

ATRX (alpha thalassemia/mental retardation syndrome X-linked) belongs to the SWI/SNF family of chromatin remodeling proteins, and along with Death-associated protein 6 (DAXX), forms a complex that is necessary for H3.3 depositions into pericentric, telomeric, and ribosomal repeat sequences.[316,317] ATRX has multiple functions in the chromatin landscape, acting both as a chromatin remodeler and a histone chaperone.[318] ATRX is a large protein (2492 residues) and contains two structured domains; an N-terminal PHD-like domain and a conserved Snf2 domain.[319] The remaining ~1660 residues of ATRX sequence are predicted to be structurally disordered, with over 1300 residues in a single stretch separating the two domains.[320] The partner protein DAXX, a H3.3 histone chaperone, contains a long disordered C-terminal domain (residues 418-740).[314] The involvement of the

IDRs in ATRX and DAXX for catalyzing the deposition and remodeling of H3.3 nucleosomes, remains unclear.

Chromatin accessibility complex (CHRAC) is an evolutionarily conserved nucleosome remodeling complex that catalyzes histone octamer sliding on DNA.[321] Originally purified from *Drosophila melanogaster*, CHRAC consists of ISWI (ATPase), ACF1 and two histone fold subunits, CHRAC-14 and CHRAC-16.[322] A study looking into the function of CHRAC-14 and CHRAC-16 found unstructured N- and C-terminal domains on both proteins.[323] CHRAC-14 and CHRAC-16 form a heterodimer with a fold that resembles the geometry of histone dimer H2A-H2B[323], which is predicted to create a surface for transient deposition of a segment of DNA as it is stripped from the core histone octamer. The C-terminal of both proteins is involved in DNA binding but with reciprocal effects; the C-terminal on CHRAC-14 increases DNA binding while the C-terminal on CHRAC-16 greatly decreases it but is still essential for sliding on DNA. It seems that the CHRAC-14/CHRAC-16 heterodimer enhances the catalysis of nucleosome sliding with weak and non-specific DNA binding. These findings were strikingly similar to the groups earlier work on the DNA chaperone HMGB1[324], leading the authors to speculate that CHRAC-14/CHRAC-16 heterodimer serves as a built-In DNA chaperone.

## ATP-independent chromatin remodelers

Post-translational modifications frequently occur in IDRs, as outlined above. Acetylation of the core histones enhances transcription by relaxing the condensed structure of the nucleosome, whereas deacetylation will promote chromatin condensation and transcriptional repression.[325,326] This effect is due to a charge neutralization of the acetylated lysine that weakens its interaction with the phosphate backbone of DNA. Both histone deacetylases and histone methylases are regulated by phosphorylations in predicted IDRs. Phosphorylations in HDACs 4,5,7 and 9 regulate shuttling between the nuclear and cytoplasmic compartments[327] and phosphorylations of sites flanking the nuclear localization sequence will promote chaperone protein binding and subsequent nuclear export.[328,329]

Histone methylation is a dynamic PTM central to eukaryotic transcription.[330] These modifications regulate gene expression by recruiting transcriptional cofactors that specifically recognize methylated lysine or arginine residues.[331,332] Dysregulation of histone methylation is associated with serious diseases such as cancers, developmental defects, and inflammatory bowel disease.[333,334] A recent study looked into PTMs of histone methylation enzymes in *Saccharomyces cerevisiae*, and found that phosphorylation was strongly enriched in predicted IDRs in methyltransferases while histone demethylases were phosphorylated within ordered regions.[335] Furthermore, the authors demonstrated that a phosphorylation cluster within an IDR of methyltransferase Set2p has a major effect on levels of H3K36 methylation *in vivo*. This decrease in H3K36 methylation leads to increased cryptic transcription, which can shorten the lifespan of cells.[336]

SIRT6 is an $NAD^+$-dependent histone deacetylase and is highly site-specific.[337] While early experiments, using H3 peptides[338], demonstrated that SIRT6 has an ~1,000 times slower catalytic activity then other related sirtuins, the low turnover rate did not match with recent studies using whole nucleosomes as substrates that found significantly higher catalytic rates.[339] This is likely due to interactions between the intrinsically disordered C-terminal region that has a high affinity to the nucleosome[340]; with SIRT6 tethered to the nucleosome the reaction can take place with greatly enhanced activity. Interestingly, while the SIRT6

interacts with nucleosomes in a 2:1 arrangement, only a single SIRT6 molecule can occupy the high affinity site. This arrangement may be due to the asymmetry of the two acidic patches, as observed with other chromatin remodelers that have a distinct response to each acidic patch.[341]

## Chromatin remodelers with chaperone activity

Facilitates chromatin transcription (FACT) is a histone chaperone that has a dual-role as a nucleosome remodeler and chaperone.[326,342] In gene regulation, nucleosomes must temporarily unfold and then rapidly refold after the regulatory process. FACT increases accessibility of RNA polymerase II on chromatin by unfolding the nucleosome structure (**Figure 2.5**).[343] FACT can then act as a histone chaperone that promotes nucleosome assembly by preventing some non-productive interactions between histones and DNA.[344] Both of FACT´s two subunits, SSRP1 and SPT16, contain acidic and disordered regions that are implicated in histone binding.[342,345] Unlike most other histone chaperones, FACT can bind both H2A-H2B and H3-H4 dimers simultaneously[346], with both subunits being involved in several interactions. Cryo-EM structures of FACT or SPT16 in complex with nucleosome constructs revealed that the CTD of SPT16, that includes an acidic IDR important for H2A/H2B binding, adopts a more ordered conformation when in complex with parts of the nucleosome.[347,348] Interestingly, the CTD appears to mimic DNA by compensating for the loss of histone DNA contacts (**Figure 2.5**).[348] In a follow-up study using NMR spectroscopy, it was revealed that one of the N-terminal tails of H3 adopts a different conformational ensemble when FACT is bound to the nucleosome.[349] NMR analysis of H3 tail chemical shifts indicated that it is buried in between two DNA gyres and that interaction is disrupted by the CTD of SPT16. This leads to increased solvent exposure of the tail, rendering it more susceptible to acetylation by HAT, indicating that FACT has a regulatory role in H3 acetylation. The intrinsically disordered domain (IDD) of SSRP1 has an acidic N-terminal part (AID) and a basic C-terminal part (BID). A recent study using NMR and CG molecular dynamics simulations, revealed how phosphorylation in the IDD change the intermolecular contacts between the AID and BID. These contact changes tune the affinity of SSRP1, with less phosphorylated states displaying high affinity to an intact nucleosome and highly phosphorylated states having high affinity to a deformed nucleosome, revealing an important mechanistic and regulatory role for the IDD.[311]

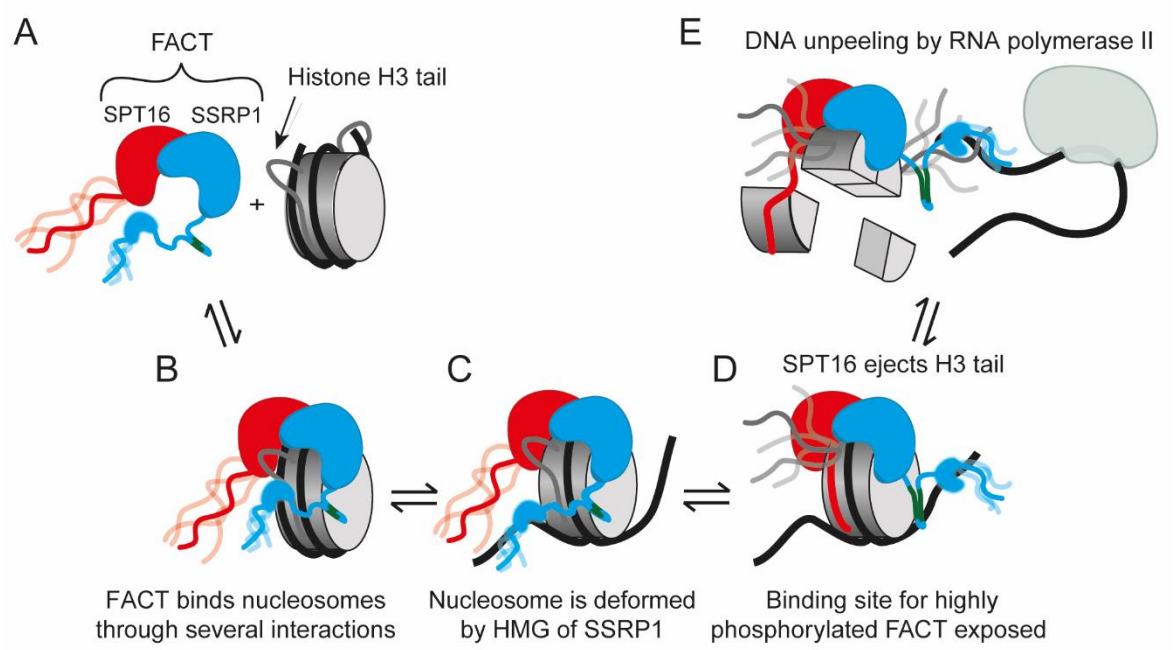



**Figure 2.5.** ***Nucleosome assembly/disassembly by the histone chaperone FACT.*** *A) A canonical nucleosome with H3 tails (for clarity only the disordered tails of histone H3 is shown) buried in DNA gyres and the two subunits of FACT, SSRP1 and SPT16. B) FACT binding to the nucleosome leads to deformation by the action of the HMG domain of the less phosphorylated (green area) state of SSRP1. C) FACT with a highly phosphorylated SSRP1 has high affinity for deformed nucleosomes and replaces less phosphorylated FACT. D) Deformation of the nucleosome exposes a binding site for the C-terminal domain of SPT16, causing increased solvent exposure of the histone H3 tail. E) DNA is peeled off the nucleosome by RNA polymerase II (or other factors). Following transcription by RNA polymerase II, FACT can reassemble the nucleosome (not shown).*

Another remodeler, decondensation factor 31 (Df31), is a fully disordered histone chaperone and an integral component of chromatin at all stages of *Drosophila melanogaster* lifecycle.[350,351] Df31 is suggested to have a role in the higher order structure of chromatin by promoting chromatin bridging in vitro.[352] Df31 binds to both histone H3 and H4 but has a higher affinity for H3.[353] Binding to H3 takes place through the intrinsically disordered H3 tail[352], making PTMs to the H3 tail a likely modulator for binding. Recently, an RNA-dependent mechanism was discovered, where Df31 tethers chromatin-associated RNA (caRNA) to chromatin, resulting in an RNA-chromatin network which is more accessible and active.[353]

We have highlighted here how structural disorder is a prominent part of chromatin remodeling complexes but for most remodelers discussed here, detailed mechanistic insights remain hidden. FACT has, however, a well-established molecular mechanism, which was revealed with a close integration of NMR experiments and coarse-grained simulations, exemplifying the strength of such approaches.

### 2.4.4 Transcription through a nucleosomal barrier with disordered proteins

The nucleosome represents a formidable barrier to transcription as the DNA sequence encoding a specific gene must become accessible to transcription factors in one way or another. The transcriptional machinery is rich with disorder and even the ribosomal assembly contains many disordered protein subunits.[354] The vast majority of transcription factors (TFs) (>85%) have long disordered linkers and transactivation domains (TADs) that flank their structured DBDs.[124,355] They bind cognate DNA sequences using predominantly their structured DBDs and may subsequently recruit other proteins to their binding site through their disordered TADs to initiate transcription. The TADs often contain hydrophobic residues (frequently aromatics) well interspersed with acidic residues, a feature that has been suggested to be important for keeping the region disordered and exposed in an active form allowing interactions with other proteins.[356] Nonetheless, the IDRs are not exclusively involved in protein-protein interactions: simulations have suggested that the affinity to DNA, cognate or non-specific, is tuned by disordered regions, especially those that have significant charges.[357] IDRs in TFs have also been linked to facilitating scanning for correct binding sites through non-specific interactions[358], and to inter-strand exchange through a monkey-bar-like mechanism.[359] In fact, recent evidence points to TFs having multiple specificity determinants encoded in their IDR sequence, helping them to identify their specific binding sites by interacting with much broader DNA regions than are recognized with only their DBD cognate sites.[360] However, in the context of our nucleosomal landscape, traditional TFs require their binding sites to be accessible for binding, i.e., within ´open´ chromatin states.

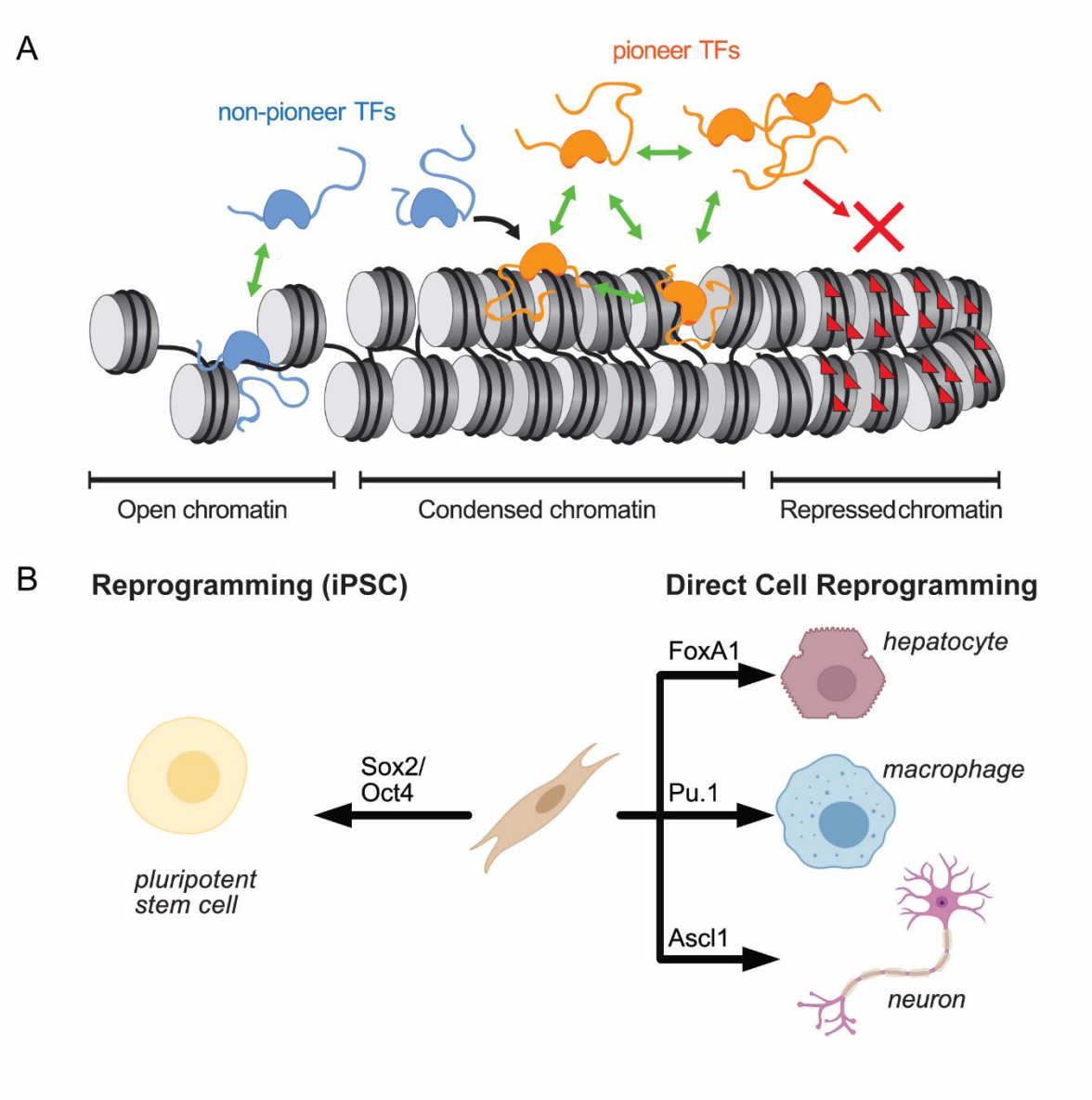



**Figure 2.6.** ***Pioneer transcription factors can invade and open condensed chromatin and initiate cell-fate changes.*** *A) Pioneer-TFs (orange) can bind to condensed chromatin regions and render it accessible to traditional TFs (blue) or other components of the transcriptional machinery. Based on Zaret & Mango.*[55] *B) Pioneer-TFs can lead to cell-fate changes, either through reprogramming with formation of induced pluripotent stem cells (iPSCs), or through direct cell reprogramming.*

## Pioneer transcription factors can alter cell fate

A unique class of TFs, called pioneer-TFs (pTFs), can bind to condensed, nucleosome-rich regions of the genome and open these previously inaccessible regions to transcription (**Figure 2.6**).[55,288] This alters the transcriptional pattern of a cell – the main determinant of its fate[55] – and can initiate cell reprogramming. Despite the ultimate change in cell fate relying on subsequent recruitment of other factors, the initial binding ability to condensed chromatin is what distinguishes pTFs from other TFs. A remarkable example of pioneer activity is the so-called Yamanaka factors; a group of four pTFs (Oct4, Sox2, Klf4, and c-

Myc) that can induce a fibroblast to revert to a pluripotent stem cell (iPSC)[361]– a process that earned the discoverers the Nobel prize in 2012. Other pTFs, such as FoxA1, Ascl1, and Pu.1, have since been shown to play key roles for inducing direct reprogramming from fibroblasts to hepatocytes, neurons, and macrophage-like cells, respectively.[288] Reprogramming cell fate has immense potential for human health, with recent reports showing extraordinary examples in regenerative medicine such as sight restoration in mice, in vitro disease modeling, and drug discovery.[362,363] However, to fully exploit the power of pTFs for cell reprogramming, a detailed and quantitative understanding of their molecular mechanism is critically needed.[288] For example, it is largely unknown whether pTFs bind to DNA that becomes spontaneously and transiently accessible on nucleosomes or whether they actively 'open' nucleosomal DNA. In other words, how pTFs can dynamically invade compacted chromatin and initiate remodeling remains unclear. Some pTFs interact with enzymes that remodel chromatin besides recruiting other TFs, and in those cases, it can be challenging to separate the actions of the two classes of proteins: are the chromatin remodelers necessary for remodeling and do the pTFs just invade chromatin to initiate binding, or can those pTFs also remodel chromatin themselves? The answers to these questions remain hidden, in part due to the highly dynamic and heterogeneous conformations of pTFs and chromatin, which render these systems notoriously difficult to assay by classical structural biology methods.

Like the vast majority of TFs, pTFs are rich in disordered linkers and TADs (**Figure 2.7**).[364] Despite their abundance in pTFs, IDRs have largely been overlooked thus far in studies of TFs, which is especially evident considering the vast number of TF DBDs in the Protein Data Bank and the total absence of 3D-structures containing entire eukaryotic TFs. Instead, intense focus has centered on the DBDs in attempts to explain pioneering activity, with impressive high-resolution structures revealing complexes between the pTF DBDs and nucleosomes.[365,366] The DBDs themselves are often disordered before binding to their cognate DNA sequence, followed by a disorder-to-order transition upon complex formation.[367] The DBDs are also often the major contributors to DNA affinity and in some cases such as for Sox2, the IDRs seemingly weaken affinity for the cognate sequence.[368] A recent computational study also implicated rotational and sliding dynamics of the DNA on the nucleosome to be important for binding of pTFs. Using CG models and simulations, Tan & Takada showed that Sox2 recognizes a certain rotational phase of its binding site which induced sliding, affecting allosterically the binding of Oct4 or another Sox2 molecule.[369] Clearly, the DBDs are critical for pTF function but what possible role do IDRs play in chromatin opening and subsequent reprogramming pathways?

## Roles of intrinsic disorder in pioneer transcription factors

Disordered regions are frequently involved in protein-protein interactions[370] and in TFs the TADs often recruit components necessary for transcription. Moreover, interplay between ordered and disordered regions is poorly understood but expected as IDRs have usually co-evolved with ordered regions and the conformational propensities of IDRs may therefore be modulated by folded domains and vice versa. It is possible that, after scanning and binding recognition sites, the DBDs act as anchors to allow the disordered regions to inflict interactions that disrupt internucleosome contacts in chromatin, leading to opening of the chromatin fiber. In that way, IDRs could be involved in actively opening chromatin, modulating oligomerization regulating pioneer activity (see below for Pu.1), or involved in recruiting chromatin remodeling enzymes, followed by opening of chromatin, and subsequent binding of other transcription factors to the exposed DNA. Ultimately, IDRs may have multiple, context-dependent roles regulated by cell-type, chromatin modifications, and

local sequence determinants. Nevertheless, the role of IDRs has been glimpsed recently for many pTFs, suggesting a function in chromatin opening and a large-scale impact on gene expression networks.

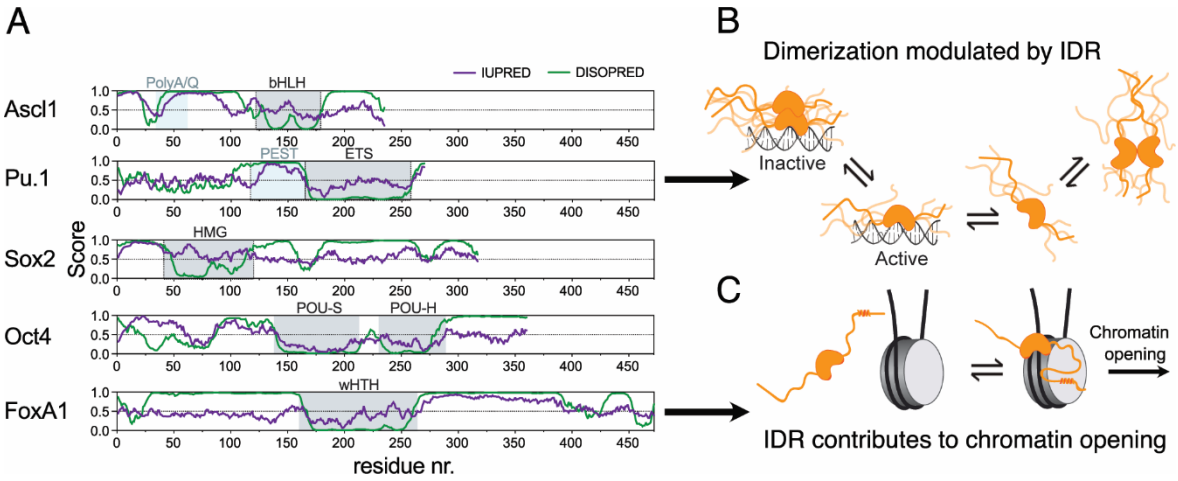



**Figure 2.7.** ***Intrinsically disordered regions in pioneer transcription factors.*** *A) Disorder predictions for five pTFs based on two different predictors. Domains are highlighted. B) The intrinsically disordered and acidic PEST domain in the pTF Pu.1 modulates the formation of a dimer on and off DNA. A 1:1 Pu.1-DNA complex activates transcription and dimerization negatively regulates the activity. The dimer in the absence of DNA is furthermore thermodynamically destabilized compared to the monomer. The reaction scheme is based on Xhani et al.*[371] *C) FoxA1 interacts with the core histones in a nucleosome through a short motif in its C-terminal IDR. This interaction contributes to chromatin opening and thus its pioneering functions. Based on results from Iwafuchi et al.*[295] *The schematics of the conformations of structured domains and IDRs in panels B) and C) are purely for illustrative purposes.*

Strong evidence of IDR involvement in chromatin opening comes from recent work from the Zaret lab, which revealed a role of IDRs in the prototypical pTFs FoxA1 and FoxA2 for modulating interactions with core histone proteins in a nucleosome[295]. Using a combination of sequence analysis, cross-linking, and mass spectrometry, the authors discovered a conserved 9 amino-acid sequence in the disordered C-terminal, which is critical for chromatin opening functions through an interaction with the core histones in a nucleosome. This short region likely forms a transient α-helix, as helix formation could be induced by addition of helix-promoting trifluoroethanol in a short peptide when monitored by circular dichroism spectroscopy. When this region was deleted, the chromatin opening ability, measured by DNase cleavage sensitivity, was severely reduced, as well as the ability to activate certain target genes. Using mouse embryos, the authors further went to show that deletion of the short a-helix led to a 60% reduction in target gene activation, severely impairing embryonic development by affecting gene expression and chromatin accessibility. Clearly, this disordered region plays a crucial role in the pioneering function of FoxA1. Beyond pioneering functions, the FoxA proteins also have heavy ties to cancer biology through their direct interaction with both the estrogen and androgen receptors[372], and FoxA1 is currently hailed as a very promising therapeutic target. The interaction of FoxA1 with both

receptors is influenced by PTMs in the disordered regions, including SUMOylation that has a negative effect on transcriptional activity and on association with the androgen receptor.[373]

The key Yamanaka factor Sox2 has a short N-terminal and a long, ~200-residue C-terminal IDR flanking an HMG-box DBD.[374] The Sox2 HMG-box cooperates with the Oct4 POU-domain, and this interaction is critical for producing iPSC and maintaining pluripotency but the efficiency of reprogramming is conferred by the extreme C-terminal IDR[375] through a currently unclear mechanism. Recent studies have shown how Sox2 and Oct4 act in concerted fashion to invoke structural changes in the core nucleosome structure ranging from subtle local distortion to fully removing DNA from one side, depending on the cognate binding site location.[365] However, the dynamic events of scanning and binding that finally lead to chromatin opening are still mostly unknown. The IDR region immediately flanking the C-terminal side of the DBD (120-160) has recently been implicated in RNA binding, even concurrently with the DBD being DNA-bound.[374] The authors went on to show that deletion of the RNA binding domain severely reduced the efficiency of iPSC generation, demonstrating a clear link between the IDR and cell reprogramming.

Pu.1 is a hematopoietic master regulator pTF that contains an N-terminal TAD, a disordered anionic PEST domain (rich in prolines, glutamic acids, serines and threonines), and a structured DBD called ETS (Erythroblast transformation specific) domain. Xhani et al. showed that Pu.1 dimerizes through its DBD and gene expression is regulated by two distinct dimeric states: a transcriptionally active 1:1 complex and an inactive ternary complex involving two Pu.1 molecules bound to a single DNA recognition site (**Figure 2.7**)[371], forming a negative feedback mechanism that the authors confirmed *in vivo*. Using NMR spectroscopy and tryptophan fluorescence experiments, the authors showed that the intrinsically disordered PEST domain reduced the binding affinity of the second Pu.1 molecule to form a ternary complex. Interestingly, however, the PEST domain also promotes homodimerization in the absence of DNA. The two dimeric forms were found to be non-equivalent, with an asymmetric DNA-bound Pu.1 dimer and a symmetric homodimer in the DNA-free state. A legion of serines in the PEST domain is phosphorylated *in vivo*, which prompted the authors to introduce phosphomimetic substitutions in that region. Indeed, the degree of negative feedback was reduced with phosphomimetic substitutions which promoted the formation of a transcriptionally active 1:1 complex with DNA. It remains to be determined whether a similar regulatory dimerization mechanism would be observed on nucleosomes but the positively charged histone tails may provide an additional interaction interface for the negative charges in the PEST domain. There may furthermore be other complicating factors, as binding of Pu.1 to nucleosomes has been reported to be context-specific, suggesting a non-classical pioneering role for Pu.1.[376]

Yet another example of a disordered pTF is the achaete-scute homolog 1 (Ascl1), which drives the conversion of fibroblasts to neurons.[377] Ascl1 is a relatively small transcription factor that has a characteristic polyA/polyQ region in the N-terminal and a basic helix-loop-helix DBD in the C-terminal. In a clever, fragment-based approach, Baronti and colleagues were able to use NMR spectroscopy to dissect the highly aggregation-prone Ascl1[378] and found an extended and dynamic structure with transient helix formation yet no persistent tertiary interactions– a classical characteristic of an IDP. Little mechanistic information is available on the interactions between Ascl1 and DNA or nucleosomes but a genome-wide analysis showed that it is one of only a handful of TFs that binds strongly to both DNA and nucleosomes albeit likely as a heterodimer.[379]

We have highlighted a subset of pTFs that have been studied by biophysical approaches but many other established pTFs are predicted to contain long IDRs.[364] Molecular biology has over the years been extraordinarily powerful at identifying the key players in transcriptional regulation networks during cell development. Yet, the link between molecular properties of pTFs, especially the role of their IDRs, and cell reprogramming is still largely missing. Integrative modeling approaches, using available structural information in concert with biophysical studies and simulations, might be a potent strategy to understand the physical principles of cell-identity pathways, leading us closer to controlling cell fate.

## 2.5 Common sequence features of disordered nucleosome-binding proteins

In the disordered interactions and their regulation reviewed above, charge emerges as a recurring theme. Charge is a principal component of chromatin and is often utilized by IDPs to elicit a specific cellular response. While the DNA backbone is highly acidic, the linker and core histone tails are highly basic, creating an electrostatic balance in the NCP.[380] Opposite charges in the DNA and histone tails have been implicated in a number of inter- and intra-nucleosomal interactions, which act to either condense or decondense chromatin. Moreover, PTMs that alter charge in the disordered histone tails have been shown to affect nucleosome stability.[381] For instance, neutralization of positive charge by acetylation or introduction of negative charge by phosphorylation of basic residues in the histone H3/H4 tail regions, weakens the histone-DNA interactions by reducing electrostatic attraction.[381] Consequently, chromatin takes on an open structure, increasing nucleosome accessibility to modifying enzymes. Charge has an especially clear role for the highly disordered H1 competitors (protamines, HMG proteins, ProTα). A common feature among these proteins may be that the unspecific nature of charge interactions and the high fraction of charges allows these proteins to interact in complexes beyond a basic 1:1 stoichiometry, exchange rapidly in a concentration-dependent manner, and keep regulatory systems highly responsive despite high affinity binding. Those molecular parameters would in turn be finely regulated by PTMs that affect charge.

In the cell, several transcription factors, chromatin remodelers and architectural proteins function in a dynamic balance, ultimately controlling gene expression. Understanding the effects of charge in IDPs that interact with chromatin and chromatin-binding proteins may provide insight into their specific cellular mechanisms. To better understand charge properties, we calculated kappa ($\kappa$) values for the IDRs of proteins discussed in this review (**Figure 2.8**). $\kappa$ is a patterning parameter used to describe strong and weak polyampholytes. A low $\kappa$ value is indicative of well distributed negative and positive charges along an intrinsically disordered domain which generate extended ensembles, where intramolecular electrostatic attractions and repulsions are counterbalanced.[126] On the contrary, a $\kappa$ value close to 1.0 indicates blocks of opposite charges that strongly interact leading to globule-like conformations with low radius of gyration.

HMG proteins compete with histone H1 to bind chromatin and thus HMG-nucleosome interactions often result in chromatin decondensation. For instance, HMGA1 competes with linker H1 on SARs of nucleosomal DNA, displacing H1 to non-SAR DNA and inhibiting chromatin compaction.[284] Interestingly, the $\kappa$ value for HMGA1 is similar to that of the CTD of linker H1.1, H1.2 and H1.5 ($0.3 < \kappa < 0.4$). Linker histone H1.5 has a long CTD tail,

containing more than two S/T-P-X-K sites, resulting in a high affinity for heterochromatin. In contrast, H1.1 and H1.2 have shorter CTD tails, with fewer S/T-P-X-K sites, and are enriched at euchromatic regions.[179] Therefore, charge distribution, in addition to net positive charge and disorder, may also impact affinity.

Like HMGA1/A2, HMGB1/B2 also contain an acidic tail and displace linker H1 from the nucleosomal dyad.[274] However, unlike HMGA1/A2, HMGB1/B2 are not completely disordered, although the disordered CTD is required for correct HMGB function.[276] HMGB1 is involved in the regulation of p53; a tumor suppressor that binds to DNA which acts by protect cells from malignant transformation.[382,383] HMGB1 has been shown to stimulate the linear DNA-p53 interaction *in vitro* and, *in vivo*, p53 activity is increased.[384] Additionally, HMGB1 and p53 have been shown to directly interact via the PXXPXP motif in the disordered NTD of p53 and HMG boxes in HMGB1. Moreover, the disordered acidic tail in HMGB1 is a direct determinant of this interaction, as it shields the positive charge in the HMG box decreasing p53-HMGB1 affinity and linking disorder to protein function.[385] In contrast to HMGB1, the interaction between H1.2 and p53 induces p53 repression in DNA damage response. Moreover, this interaction is negatively regulated by acetylation in the p53 CTD and phosphorylation in the H1.2 CTD. In both cases, PTM acts to disrupt the p53-H1.2 interactions, directly implicating charge and disorder in protein functionality.[386]

The connection between charge and disorder is prominent when considering the interplay of HMGN/N2 and H1T2, H1oo and HILS1 variants. HMGN/N2 promote chromatin decompaction by interacting with nucleosomal DNA at the major grooves flanking the dyad and competing with linker H1 for binding sites.[256] Furthermore, HMGN1/N2 has a low kappa value ($0.1 < \kappa < 0.2$), that is similar to that of the H1T2, HILS1 and H1oo CTD and characteristic of disordered proteins. Therefore, HMGN1/N2 may use its disorder to compete with these H1 variants for binding sites. Interestingly, while most H1 variants have few arginine residues, H1T2 and HILS1 have an almost equal fraction of lysine and arginine residues in the CTD. Because arginine forms stronger interactions with the DNA phosphate backbone, the testis specific variants are likely to be harder to displace. For instance, during spermatogenesis, inactivation of the gene for H1T2 leads to defects in DNA condensation and chromatin packing; effects that are not favorable in cell development.[387]

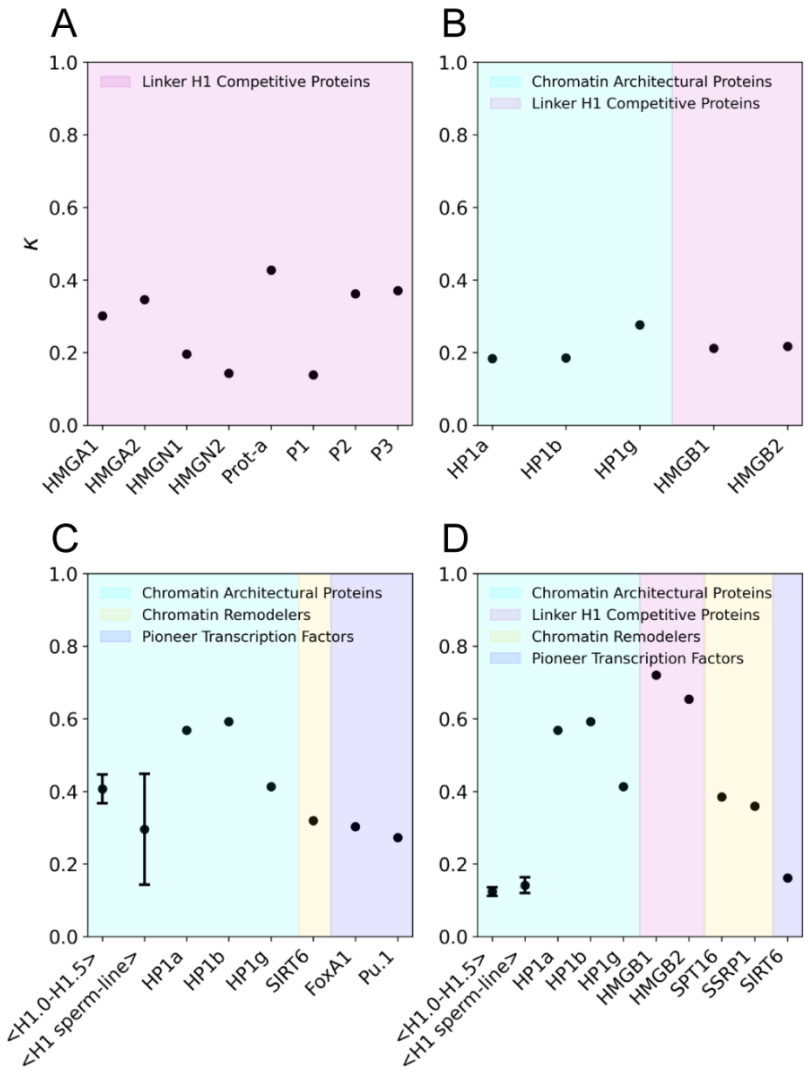



**Figure 2.8.** ***Overview of the charge distribution within the intrinsically disordered regions of nucleosome-binding proteins.*** *Understanding the role of charge in nucleosome-binding proteins may prove important to frame the functional space of different intrinsically disordered proteins within the context of transcriptional regulation. The parameter kappa (κ), which can take any value between 0 and 1 and has been formulated to link sequence properties to conformational behavior of intrinsically disordered regions*[126]*, describes the overall charge asymmetry in an amino-acid sequence. A low κ indicates more evenly distributed positive and negative charge, whereas κ increasingly close to 1 indicates blocks of positive and negative charge. The need to strongly coordinate DNA may render charge distribution an important factor to finely tune protein-DNA interactions. While for fully disordered proteins (shown in A) κ is found to vary considerably, for intrinsically disordered hinge regions linking structured domains (shown in B) κ is low and similar across all proteins. A clear difference can be noted between the κ values of the N-terminal (C) domains of linker histones that are involved in the physiological homeostasis of chromatin and those expressed by sperm-line cells, which are involved in extreme chromatin compaction and that show a lower κ. While no other trend can be clearly seen for the other proteins or for their C-terminal domains (shown in D), it is important to acknowledge that our classification is arbitrary and founded on the current understanding of the role that these proteins have within the nucleosomal landscape.*

## 2.6 Concluding remarks

The nucleus is enriched in proteins that are disordered and thus highly dynamic. These proteins play key roles in maintaining the genome and regulating its read-out. Despite decades of active research on IDPs and their well-recognized importance in ensuring the homeostasis of the nucleus, we still lack an exhaustive description of the interactions between chromatin components and IDPs, especially with respect to how they translate to biological function and regulation. New methodological paradigms are needed to tackle intrinsic disorder in the nucleus, because of both the intrinsic dynamic character and the physico-chemical properties of the interacting molecular partners, which frequently feature extremely strong electrostatics. Consequently, in recent years there has been a considerable upsurge in methodological development, especially for single-molecule techniques which can discriminate distinct conformational sub-populations and sequences of molecular events. Computational approaches that directly integrate single-molecule data and simulations, featuring customized potential energy functions tuned on the basis of experimental findings, have provided an unprecedented view of the ensemble of some key disordered interactions in the nucleus. Remarkably, simple potential energy functions that dominantly account for electrostatic contributions to binding, have been able to exhaustively reproduce experimental findings and provide a mechanistic understanding of protein-protein and protein-DNA interactions.[40,118,388] In the future, such simple customized potentials may evolve into more complex combinations of potential energy terms that might take into account, explicitly, the effects of post-translational modifications, such as methylation and acetylation, in specific sites along intrinsically disordered domains. Additionally, the modeling and parameterization of explicit ionic species, especially for coarse-grained simulations, would be a considerable advancement for a more accurate estimation of the energetics involved in the nucleosomal landscape, especially considering the primary role of ions in defining the association of strong disordered polyampholyte chains that interact with chromatin. Overall, access to integrative modeling approaches is still a challenge, as it requires strong collaborative efforts between different research groups. Nevertheless, creating synergy between experiments and simulations is key to refining our view of the disordered nuclear milieu.

**Acknowledgements**

This work was funded by the Icelandic Research Fund (Rannís grant nr. 206591-052 and 217392-051, to P.O.H.) and the Centre for eResearch (http://www.eresearch.auckland.ac.nz) at the University of Auckland for their help in facilitating this research. We apologize to the authors of other excellent work that we could not cite due to space limitations.

# 3 Chapter 3

## DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2

Sveinn Bjarnason, Jordan McIvor, Andreas Prestel, Kinga S. Demény, Jakob T. Bullerjahn, Birthe B. Kragelund, Davide Mercandante, Pétur O. Heidarsson

# About this chapter

This chapter was published as an article in Nature Communications under the same title. It centers on the pTF Sox2, specifically examining the dynamic interactions between its DBD and IDRs, and their role in regulating access to its activation domains.

# Author contribution

S.B., J.A.P.M., A.P., B.B.K., D.M. and P.O.H. designed the study; S.B. and K.S.D. prepared protein and DNA constructs; S.B performed all single-molecule and circular dichroism experiments; S.B. and A.P. performed NMR experiments; J.T.B. provided new analytical tools; J.A.P.M. and D.M. performed simulations; S.B., J.A.P.M., A.P., J.T.B., B.B.K., D.M. and P.O.H. analysed data; S.B. and P.O.H. wrote the manuscript with help from all authors.

# Keywords

Intrinsically Disordered Proteins; Intrinsically Disordered Regions; Pioneer Transcription Factors; Sox2; DNA; smFRET; Simulations and NMR.

## 3.1 Abstract

More than 1600 human transcription factors orchestrate the transcriptional machinery to control gene expression and cell fate. Their function is conveyed through intrinsically disordered regions (IDRs) containing activation or repression domains but lacking quantitative structural ensemble models prevents their mechanistic decoding. Here we

integrate single-molecule FRET and NMR spectroscopy with molecular simulations showing that DNA binding can lead to complex changes in the IDR ensemble and accessibility. The C-terminal IDR of pioneer factor Sox2 is highly disordered but its conformational dynamics are guided by weak and dynamic charge interactions with the folded DNA binding domain. Both DNA and nucleosome binding induce major rearrangements in the IDR ensemble without affecting DNA binding affinity. Remarkably, interdomain interactions are redistributed in complex with DNA leading to variable exposure of two activation domains critical for transcription. Charged intramolecular interactions allowing for dynamic redistributions may be common in transcription factors and necessary for sensitive tuning of structural ensembles.

## 3.2 Introduction

Transcription factors (TFs) consolidate information for gene expression by locating specific DNA sequences in the nucleus and recruiting cofactors to regulate transcription. Most human TFs consist of structured DNA binding domains (DBDs) and long intrinsically disordered regions (IDRs) that can harbour activation domains (ADs), and thus interaction sites for regulatory binding partners[389,390]. Whereas intense focus has been on the structured DBDs, IDRs in TFs have been understudied due to the major challenges such regions pose for traditional structural biology techniques. Consequently, there is a significant lack of accurate descriptions of IDR ensembles for all of roughly 1600 human TFs, both off- and on their DNA recognition sites. Beyond hosting the ADs important for transcriptional activation, IDRs in TFs can have many other roles such as modulating DNA binding affinity[357], contributing competence for phase separation[391], or regulating DNA binding specificity[360]. In recent years, the importance of electrostatic interactions for the conformational dynamics of IDRs has become increasingly evident[40,107,392]. Experiments and computational modelling have suggested that charged patches on folded domains modulate the dimensions of adjacent IDRs, which might have direct functional consequences[393-397], and charge modulation by posttranslational modifications (PTMs) such as phosphorylation can have a large impact on the ensemble[390,398]. However, the conformational signatures of such molecular behaviour have not been broadly established, and generally, IDR conformational dynamics and their modulation by DNA binding is poorly understood. Structural models of IDR ensembles are critical to understand the code of transcriptional regulation and to decode how PTMs affect gene regulatory networks.

In this work we address these challenges by studying the structure and dynamics of pluripotency factor Sox2, a prototypical TF, which plays a pivotal role in maintaining embryonic and neuronal stem cells[399]. Sox2 is classified as a pioneer transcription factor due to its ability to target its cognate binding sequence in condensed, nucleosome-rich DNA[52]. Sox2´s pioneer activity– along with the other so-called Yamanaka transcription factors Oct4, Klf4, and c-Myc–, has recently been applied to generate induced pluripotent stem cells (iPSCs), bringing immense potential to regenerative medicine and drug development[400]. Sox2 has 317 residues and consists of a small HMG-box DBD[401] flanked N-terminally by a short 40-residue low-complexity stretch and C-terminally by a long ~200-residue region, both of which are predicted to be disordered (N-IDR and C-IDR, respectively)( **Figure 3.1a**). Little is known about the function of the short N-IDR but there is evidence that it is important for interactions with other transcription factors[402]. The DBD is rich in positively charged residues (net charge = +13)– as commonly observed in DNA-binding proteins[390]– which

facilitate binding to the negatively charged DNA. The C-IDR is enriched in methionines, serines, glycines and prolines (~40% of total residues) and contains 18 charged residues (zero net charge) distributed throughout the sequence. The C-IDR contains two predicted ADs: AD1 (residues ~150-200), which was recently validated in a large-scale mapping of TF IDRs[403], and AD2 (residues ~250-300)[10,404-406] (**Figure 3.1a**). The two ADs are separated by a serine-rich domain (residues ~200-250), which mediates direct interaction with the TF Nanog in a process important for self-renewal of embryonic stem-cells[407]. There is evidence that the IDRs of Sox2 are neccessary for pioneering function[374] but it is unclear whether they are important only for transcriptional activation or for other functions such as chromatin binding or opening, as observed for some pioneer factors[295,408]. Indeed, the C-IDR of Sox2 has recently been found to have functions that extend beyond transcriptional activation, ranging from contributing to force exertion on DNA[409], RNA binding[368,374], and DNA scanning and target site selection[410]. However, a quantitative description of the C-IDR conformational ensemble is lacking and it is unclear how the ensemble is affected by DNA binding, and ultimately how it conveys function.

We used single-molecule Förster resonance energy transfer (smFRET) and nuclear magnetic resonance (NMR) spectroscopy, combined with molecular dynamics (MD) simulations to comprehensively map the conformational dynamics of full-length Sox2. We show that the C-IDR engages in dynamic interactions with the DBD involving its charged residues and that this constrains its dimensions in an exquisitely salt-sensitive manner. These interactions are substantially altered in complex with both DNA and nucleosomes which leads to a more extended C-IDR. We reconstruct experimentally-derived FRET values from a coarse-grained (CG) simulation and reveal the structural ensemble of free and DNA-bound Sox2. Our structural ensemble reveals a large-scale re-arrangement in the C-IDR dimensions upon DNA binding, which specifically redistributes the accessibility of the two transcriptional ADs. Considering general sequence features of TFs[390], this type of charge-driven IDR ensemble modulation is likely to be common among eukaryotic TFs where charge patterning and PTMs are expected to play an important role.

# 3.3 Results

## 3.3.1 Sox2 C-IDR is disordered and dynamic

While structures of the Sox2 DBD show that its conformations in free and DNA-bound states are highly similar[365,366], high-resolution structural information on full-length Sox2 in regions outside the DBD are currently unavailable. Structure and disorder predictions indicate that the mainly disordered C-IDR contains short polypeptide stretches with some secondary structure propensities which coincide with the ADs (**Figure 3.1a**). Indeed, far-UV circular dichroism (CD) spectra of full-length Sox2 as well as of isolated domains (N-terminal domain and DBD (N-DBD), and C-IDR) generally agree with predictions (**Figure 3.1b, Supplementary Table 3.1**). The far-UV CD spectrum of the N-DBD showed minima at 222 nm and 208 nm, suggesting the presence of mainly helices, whereas the C-IDR gave a spectrum that suggested mainly a random-coil with a large negative ellipticity minimum at 202 nm, indicating an overall lack of secondary structure.

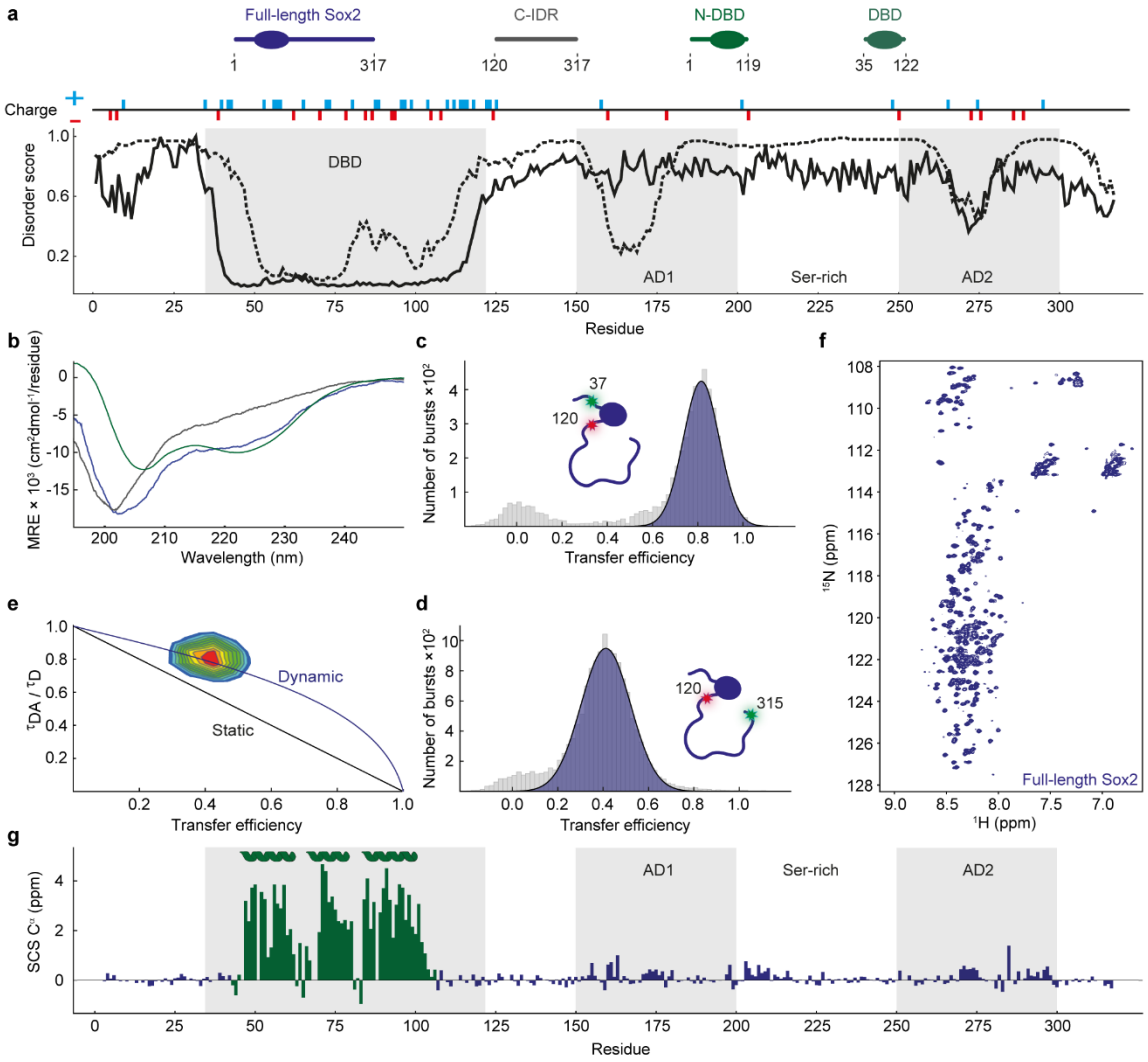



**Figure 3.1.** ***Sox2 C-IDR is disordered and dynamic. a)*** *Schematics of Sox2 illustrating the main constructs used in this study. The plot shows disorder predictions as a function of residue number, based on two different predictors (Disopred3[411] (dashed line), AlphaFold[10] normalized pLDDT (solid line)). The DBD is indicated, as are the ADs and serine-rich region (see text for details), and the locations of charged residues.* ***b)*** *Far-UV circular dichroism spectra of different Sox2 variants at 5 µM concentration; Full-length Sox2 (blue), C-IDR (grey), N-DBD (green). Spectra are averages of n=3 independent measurements.* ***c-d)*** *Single-molecule transfer efficiency histograms of Sox2 fluorescently labelled flanking the DBD (residues 37 and 120, number of molecules=5323) or probing the entire C-IDR (residues 120-315, number of molecules=14544). The small peak at E~0 originates from donor-only labelled molecules that remain after filtering (see Methods and Supplementary Figure 3.1)* ***e)*** *Fluorescence lifetime analysis of the Sox2 C-IDR. The 2D-correlation plot shows fluorescence lifetimes of the Cy3b donor ($t_{DA}$) relative to the intrinsic donor fluorescence ($t_D$). The dynamic line is based on a SAW-n polymer model. See text for details.* ***f)*** *$^{1}H^{15}N$-HSQC spectrum of full-length Sox2.* ***g)*** *$C^{\alpha}$ SCS plot of full-length Sox2 (blue). SCSs for the DBD (green) were determined for the isolated N-DBD domain. The known helix locations (UniProt P48431) are indicated, and grey shaded areas indicate the DBD and ADs.*

To quantify the dimensions and dynamics of Sox2 in more detail we turned to smFRET[162,412]. We designed cysteine mutations to specifically probe the major domains and labelled them through thiol chemistry using the fluorophore pair Cy3b and CF660R. We then used smFRET to measure mean transfer efficiency, $\langle E \rangle$, of thousands of individual and freely-diffusing molecules using a confocal fluorescence microscope. When the dyes were flanking the DBD (positions 37 and 120, **Figure 3.1c**) we measured an $\langle E \rangle$~0.8, which corresponds to an average distance between the dyes close to that expected from structural studies (PDB 6T7B), indicating that the DBD remains folded in our experiments (**Methods** and **Supplementary Table 3.2**). For probing the long C-IDR, we placed the dyes just after the DBD (position 120) and near the C-terminus (position 315), measuring a FRET efficiency $\langle E \rangle$=0.43 (**Figure 3.1d**). Given that the structure predictions and CD data indicate a mainly random coil for the C-IDR, we used a self-avoiding walk polymer model with a variable scaling exponent $\nu$ (SAW-$\nu$) to determine the root mean square distance ($R_{RMS}$) between the two dyes (**Methods**). The SAW-$\nu$ model has recently been shown to describe well the dimensions of intrinsically disordered proteins (IDPs)[413]. The $\langle E \rangle$ of the C-IDR leads to an $R_{RMS}$ of 7.5 nm and a scaling exponent $\nu$ of 0.57, which is within the range expected for an IDP[414].

To probe rapid conformational dynamics of the C-IDR, we can use relative fluorescence lifetimes to detect distance fluctuations between the two fluorophores, on a timescale between the fluorescence lifetime (ns) and the interphoton time (µs). The relative donor lifetime (the ratio between the donor lifetime in absence ($\tau_D$) and presence ($\tau_{DA}$) of an acceptor) can be shown from the Förster equation to equal to $\frac{\tau_{DA}}{\tau_D} = 1 - \langle E \rangle$ only if there is a single, effectively static distance (on the same timescale) separating the two dyes (**Figure 3.1e** and **Methods**). Conversely, if a distribution of distances is sampled due to dynamics of the polypeptide chain, the relative lifetimes cluster above the diagonal line, to an extent defined by the variance of the underlying distance distribution. For dyes probing the Sox2 C-IDR, the relative lifetimes deviate significantly from the diagonal "static" line and agree with a "dynamic" line based on the expected behaviour of a SAW polymer with a scaling exponent of 0.57, as obtained from the measured $\langle E \rangle$.

Since the FRET experiments do not report directly on potential secondary structure formation, we used NMR spectroscopy to extract residue-specific structural information on Sox2. We produced $^{15}N^{13}C$-isotope labelled full-length Sox2 and first measured a $^{1}H^{15}N$-heteronuclear single quantum coherence (HSQC) spectrum of full-length Sox2. The HSQC spectrum displayed almost the full set of expected signals from all backbone amides (**Figure 3.1f**), with little dispersion of resonances in the proton dimension, characteristic of an IDR[40]. From sets of triple resonance spectra, we could assign 275 peaks out of 290 assignable ( >95%). The peak intensities of residues in the DBD were much lower than for the disordered regions, presumably due to slow rotational tumbling, hence the assignments of the DBD NMR signals were performed for the isolated N-DBD and transferred to the spectra of full-length Sox2 (**Figure 3.2, Supplementary Figure 3.2, and Methods**). A secondary chemical shift (SCS) analysis of $C^{\alpha}$ and $C^{\beta}$ shifts revealed a general lack of secondary structures in the C-IDR with potential transient helix or turn formation in regions coinciding with the ADs (<7% helix in residue regions G150-Q175, Y200-S220, S275-S300, calculated using the shifts for the DBD as reference for 100%) in agreement with predictions, whereas we observed strong signatures for the three expected helices in the DBD (**Figure 3.1a,g**).

### 3.3.2 C-IDR dimensions are shaped by charged interactions with the DBD

The classical modular view of TFs, which assumes separate functional domains unaffected by each others' presence, has recently come into question and at the same time, interdomain synergy and context are increasingly coming into view[415,416]. Charged residues can partake in long-range interactions and play a primary role in the conformational dynamics of IDRs[40]. The fraction of charged residues in the C-IDR of Sox2 (+9,-9) classifies it as a weak polyampholyte and predicts it to adopt a collapsed state[126]. However, the DBD contains a high density of charges, with a net charge of +13 to facilitate binding with the negatively charged DNA. We therefore investigated whether interactions between the C-IDR and the neighbouring DBD might contribute to the observed dimensions of the C-IDR. We produced fluorescently labelled isolated DBD and C-IDR to compare their dimensions to that of the full-length protein using smFRET. We used a Sox2 construct with fluorophores in positions 120 and 265, which probes the majority of the C-IDR with high sensitivity ($\langle E \rangle$ =0.55, which is close to the Förster radius at $E$=0.5). We observed a significantly lower FRET efficiency for the isolated C-IDR compared to the same region within full-length Sox2 ($\langle E \rangle$ = 0.48±0.01 vs. $\langle E \rangle$ = 0.55±0.01, respectively)(**Figure 3.2a**), whereas the end-to-end distance of the DBD (fluorescently labelled in residues 37 and 120) was largely independent of context (**Figure 3.2b**). These data indicate that the C-IDR is more compact in the presence of the neighbouring N-DBD, providing strong evidence for the presence of interdomain interactions between the DBD and C-IDR.

To capture the physical basis for the interactions, we performed titration experiments by measuring FRET histograms in varying concentrations of chemical denaturants (urea or guanidinium chloride (GdmCl)) or salt (KCl). The apparent radius of gyration, $R_g$ (determined from the SAW-$\nu$ distance distribution using the measured $\langle E \rangle$ at each denaturant concentration), was plotted as a function of titrant concentration (**Figure 3.2c,d, Supplementary Figure 3.3**). In both urea and GdmCl, the C-IDR gradually expanded (increased $R_g$) with increasing concentration of denaturant for both the full-length protein and the isolated domain. We fitted the unfolding data with a weak denaturant binding model that assumes $n$-independent binding sites for denaturant molecules, which allows determination of an effective association constant, $K_a$ (**see Methods**). Interestingly, while the $K_a$ for urea, which is uncharged, is unaffected by the absence of the neighbouring N-DBD, the $K_a$ for GdmCl, which is charged, is reduced by almost 50% (**Figure 3.2e**). Since the charged GdmCl disrupts electrostatic interactions whereas urea does not, this suggests the presence of interdomain communication between the DBD and C-IDR being based predominantly on interactions between charged residues. This was further supported when we measured transfer efficiency histograms over a range of salt concentrations (**Figure 3.2f, Supplementary Figure 3.3**). Remarkably, the C-IDR dimensions in full length Sox2 were exquisitely sensitive in the physiologically relevant range of salt concentrations (100-200 mM KCl). The full-length Sox2 displayed a pronounced "roll-over", suggesting screening of charge interactions with increasing salt concentrations, but the roll-over effect was entirely absent in the isolated C-IDR. Similar observations have been reported in other proteins[417] and can be explained by polyampholyte theory[418,419]; strong interactions between oppositely charged residues cause a collapse of the chain which are subsequently screened upon addition of salt, causing the chain to expand. The chain then compacts again at higher and unphysiological salt concentrations (700-2000 mM), potentially due to an enhancement of hydrophobic interactions as observed for other charged proteins[417]. Overall, even though

the C-IDR contains relatively few charges causing it to adopt a collapsed state [126], charged interactions with the DBD sensitively control its dimensions further.

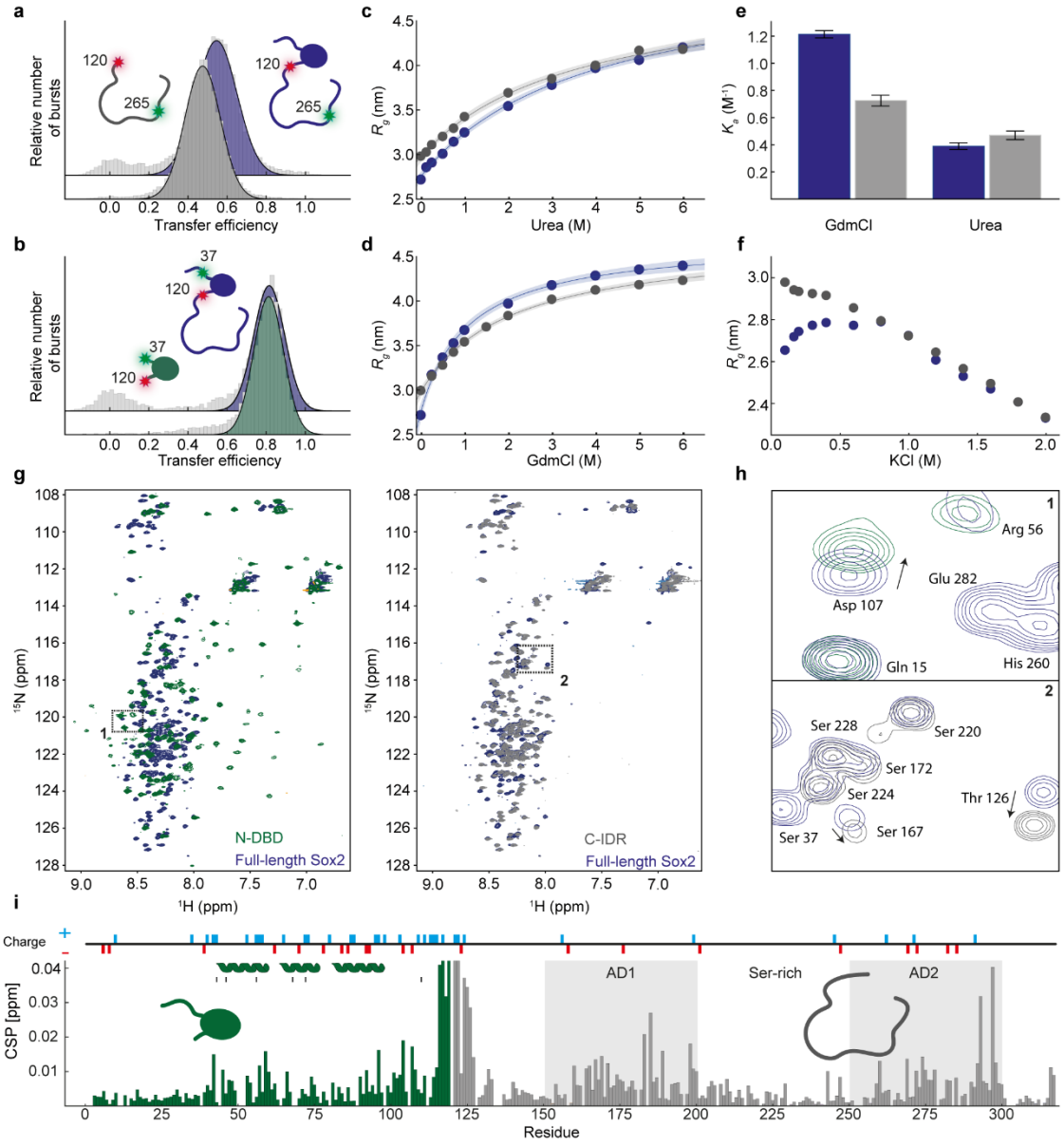



**Figure 3.2.** ***Interdomain interactions between Sox2 DBD and C-IDR. a-b)*** *Single-molecule transfer efficiency histograms of full-length Sox2 and either an isolated C-IDR, both labelled at positions 120 and 265 (a), or an isolated DBD, both labelled at positions 37 and 120 (b).* ***c-d)*** *Apparent $R_g$ (see text and Methods for details) of the fluorescently labelled C-IDR in full-length Sox2 (blue) or isolated (grey) as a function of urea (c) or GdmCl (d) concentration. The solid lines are fits to a weak denaturant binding model and the shaded areas represent 95% confidence intervals.* ***e)*** *Denaturant association constant ($K_a$), determined from fits to the data in panels c and d, for the C-IDR in full-length Sox2 (blue) and isolated (grey). Error bars are standard errors of the fit.* ***f)*** *Apparent $R_g$ of the C-IDR in full-length Sox2 (blue) or isolated (grey) as a function of KCl concentration.* ***g)*** *$^1H^{15}N$ HSQC spectra of full-length Sox2 (blue), overlayed with a spectrum of the isolated N-DBD (green,*

*left) and the isolated C-IDR (grey, right).* ***h)*** *Boxes 1 and 2 are zooms into specific regions of the HSQCs in panel g, showing overlap of some peaks and changes in position of others.* ***i)*** *CSP plot showing the chemical shift difference between full-length Sox2 and each isolated domain, N-DBD (green) and C-IDR (grey). Helix locations in the DBD are indicated where black lines denote residues important for DNA binding (PDB 6T7B). Grey shaded areas indicate ADs.*

Long-range interdomain contacts should be revealed by differences in NMR chemical shifts between the full-length protein and isolated domains. We therefore produced $^{15}$N,$^{13}$C-isotope labelled isolated N-DBD and C-IDR for chemical shift assignments using sets of triple resonance NMR spectra. For the N-DBD and C-IDR we could assign 104 peaks out of 109 (expected excluding prolines and N-terminal methionine, >95%) and all 180 observable peaks in the $^{1}$H$^{15}$N-HSQCs, respectively (**Figure 3.2g,h**). Comparing SCSs between the isolated C-IDR and the full-length protein revealed similarly lacking propensity to form secondary structure outside the DBD (**Supplementary Figure 3.4**). The spectrum of the DBD displayed dispersed peaks, indicating a well-folded domain. Importantly, the C-IDR peaks overlapped well with the peaks from the full-length Sox2 in some regions but not in others, indicating a different chemical environment due to missing interdomain interactions in the isolated constructs, in agreement with the smFRET data (**Figure 3.2a**). The regions with the largest chemical shift perturbations (CSPs) overlapped with regions of the highest charge density (**Figure 3.2i**), in the vicinity of the ADs. The N-DBD was similarly affected mostly in the folded HMG domain that contains the highest density of charge, and in the region in close proximity to the missing C-IDR, whereas the N-terminal tail was minimally perturbed. These results were re-enforced by titrating a $^{15}$N-labelled C-IDR with an unlabelled DBD and vice versa, which showed considerable CSPs around the most charge-dense regions in both domains (**Supplementary Figure 3.4**). Using the chemical shift changes of highly perturbed residues, we could estimate the dissociation constant, $K_D$, for the complex in trans to be 80 ± 4 μM (**Supplementary Figure 3.4**).

### 3.3.3 DNA and nucleosome binding expands dimensions of C-IDR

Having established the conformational dynamics and interdomain interactions in the free state of Sox2, we next asked how these might be affected by complex formation with DNA. We speculated that perturbation of electrostatic interactions across domains upon DNA binding would lead to conformational changes in the C-IDR. We first checked that Sox2 binding leads to the expected bending of DNA[365] by using fluorescently labelled oligonucleotides carrying a Sox2 binding site (TTGT) (**Supplementary Table 3.3**). At physiological salt concentrations (165 mM KCl), the free 15 bp dsDNA had a FRET efficiency $\langle E \rangle$ ~0.4 (**Figure 3.3a**). When unlabelled Sox2 was added to the solution, another population appeared at higher FRET, $\langle E \rangle$ ~0.6, indicative of the expected Sox2-mediated DNA bending. We used the areas of the resulting FRET histograms to determine the fraction of bound DNA as a function of Sox2 concentration, and thus estimated the equilibrium dissociation constant, $K_D$. We constructed and fitted binding isotherms for both full-length Sox2 and the isolated DBD, and observed that the dissociation constant was largely unaffected by the presence of the C-IDR (0.3 ± 0.1 nM for DBD vs 0.4 ± 0.2 nM for full-length Sox2), in agreement with previous results[368,374] (**Figure 3.3b, Supplementary Figure 3.5, and Supplementary Table 3.4**). This was also true for a non-specific DNA without a Sox2 binding site yet with ~10-fold higher $K_D$, also in agreement with previous results

(**Supplementary Figure 3.5**). Thus, both specific and non-specific DNA binding to the DBD was unaffected by the interdomain interaction. The dissociation constant determined using fluorescently labelled Sox2 (**Supplementary Figure 3.5**) was very similar to that obtained with labelled DNA, excluding adverse effects on binding affinity due to the fluorophores.

To detect potential changes to the C-IDR conformations when in complex with DNA, we measured single-molecule transfer efficiency histograms for Sox2 fluorescently labelled in the C-IDR and in presence of unlabelled target DNA (**Figure 3.3c**). We observed a substantial change in FRET efficiency; the C-IDR expanded considerably upon binding DNA, with FRET decreasing from 0.43 to 0.28 (**Figure 3.3c**). This is in contrast to the DBD end-to-end distance which even compacted slightly **(Supplementary Figure 3.5)**. The change in FRET corresponds to an increased $R_{RMS}$ for the C-IDR ensemble from 7.5 nm to 9.2 nm or more than 20%. Analysis of the relative lifetimes of fluorophores probing the C-IDR in complex with DNA still showed deviation from a static distance, indicating that submillisecond dynamics of the C-IDR persist on DNA **(Figure 3.3e)**. To quantify the dynamics, we performed nanosecond fluorescence correlation spectroscopy (nsFCS) experiments of Sox2 in absence and presence of DNA, probing the C-IDR dynamics (**Supplementary Figure 3.6**). Fitting the anti-correlated donor-acceptor cross-correlation functions, which decay on the timescale of interdye distance fluctuations, allowed us to determine the reconfiguration time ($t_r$) of the C-IDR (Methods). In agreement with the fluorescence lifetime analysis, $t_r$ is similar in the absence and presence of DNA (172 ns and 184 ns, respectively) whereas the isolated C-IDR reconfigures slightly faster ($t_r$ ~105 ns), presumably due to the lack of the neighbouring DBD to interact with.

Sox2 is a strong nucleosome binder, which is thought to play a role in its function as a pioneer factor. We therefore also tested whether similar conformational changes as observed for DNA would occur upon binding to nucleosomes. We reconstituted nucleosomes using the strongly positioning Widom-601 sequence with an incorporated Sox2 binding site, previously shown to be stably bound by Sox2[365] (**Figure 3.3d, Supplementary Figure 3.7, Supplementary Table 3**). We then measured transfer efficiency histograms for full-length Sox2 fluorescently labelled in the C-IDR and in the presence of unlabelled nucleosomes. The mean FRET efficiency of the C-IDR in complex with nucleosomes was very similar to the one measured in complex with a shorter DNA (**Figure 3.3c,d**), and fluorescence lifetime analysis showed slightly dampened dynamics (**Figure 3.3f**), which could indicate a weak interaction with the histone octamer. We confirmed that the DNA stays wrapped around the histone octamer during the experiment by estimating the diffusion time of Sox2 in the presence of DNA and nucleosomes, and by measuring FRET on fluorescently-labelled nucleosomes (**Supplementary Figure 3.7**)[420]. Overall, these data thus indicate that the conformational ensemble of the Sox2 C-IDR is similar in complex with DNA and nucleosomes.

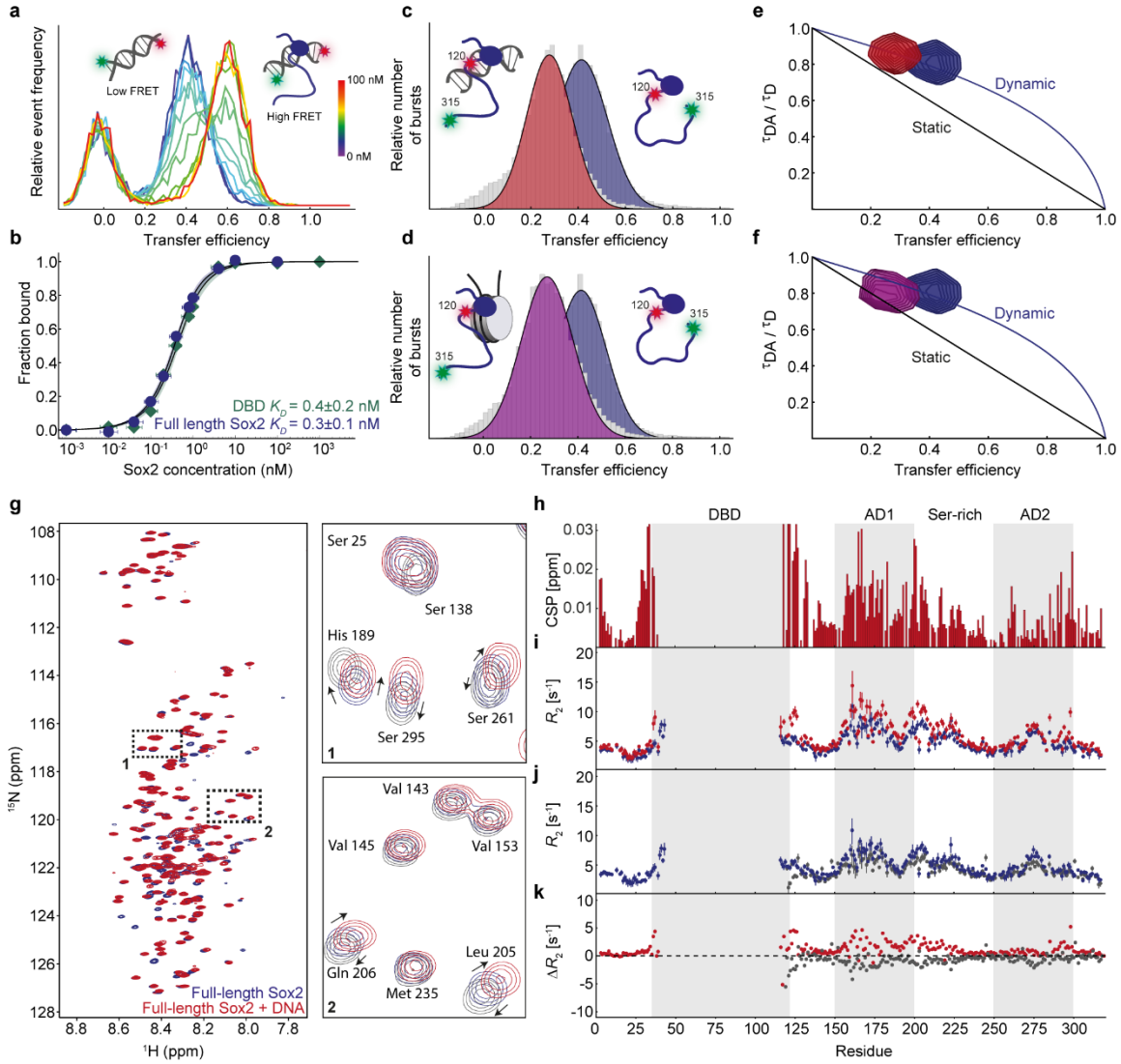



**Figure 3.3.** ***Conformational rearrangements of the Sox2 C-IDR upon binding DNA and nucleosomes. a)*** *Single-molecule transfer efficiency histograms of fluorescently labelled 15 bp DNA containing Sox2 binding site, with different concentrations of unlabelled full-length Sox2. The peak at E~0 corresponds to a population of molecules without an active acceptor.* ***b)*** *The corresponding binding isotherms with fits (solid lines) to a 1:1 binding model, for both full-length Sox2 and the isolated DBD. Error bars indicate uncertainties estimated from dilution errors.* ***c-d)*** *Single-molecule transfer efficiency histograms of full-length Sox2 fluorescently labelled in the C-IDR, in the absence (blue) and presence of c) 15 bp DNA (red) or d) 197 bp nucleosomes (purple).* ***e-f)*** *Fluorescence lifetime analysis of Sox2 in the absence (blue) and presence of e) DNA (red) or f) nucleosomes (purple).* ***g)*** *$^1H^{15}N$ HSQCs of free Sox2 (blue) and Sox2 in complex with 15 bp unlabelled DNA (red). Zoomed-in regions show resonances that are affected or unaffected by DNA binding.* ***h-k)*** *Plots of h) CSPs for Sox2 upon DNA binding and $^{15}N$-relaxation data (i-k) $R_2$ for free Sox2 (blue) and DNA-bound Sox2 (red, i), isolated C-IDR (grey, j), and the respective difference plot (C-IDR - free Sox2 (grey), DNA bound Sox2 - free Sox2 (red), k). Error bars indicate 95% confidence intervals.*

To probe DNA binding on a residue-specific level, we again used NMR spectroscopy. A $^{1}H^{15}N$-HSQC of DNA-bound full-length Sox2 showed similar low dispersion of peaks from the C-IDR but distinct chemical shift changes when compared with free Sox2 (**Figure 3.3g**), whereas peaks from the DBD were entirely absent. When we plotted the CSPs as a function of residue sequence, we observed that most of the CSPs localize to the regions we had previously observed to make contacts with the DBD (**Figure 3.3h**, **Figure 3.2i**). Importantly, many of the chemical shifts imply a different structural ensemble for the C-IDR in the DNA bound state than for the free C-IDR construct (**Figure 3.3g,h**), suggesting that it is not just a simple release of interactions with the DBD but rather a different ensemble that is populated on DNA (**Figure 3.3g, zooms**). We then measured the fast time scale dynamics of the different states using NMR. Residue-specific relaxation rates (**Figure 3.3i,j,k, Supplementary Figure 3.8**), which probe ps-ns dynamics, were generally low and globally increased slightly across the entire polypeptide chain upon DNA binding, indicating contributions due to slowed tumbling. Comparing relaxation rates between free full-length Sox2 and either DNA-bound or the isolated C-IDR showed little changes in dynamics on this timescale. Overall, the NMR data indicate that the C-IDR structural ensemble is different in complex with DNA yet it remains dynamic, in agreement with the fluorescence lifetime analysis.

### 3.3.4 Coarse-grained simulation reveals redistributed accessibility of activation domains

To reconstruct the structural ensemble of Sox2 when free and bound to DNA, we performed CG Langevin dynamics simulations. Here, every amino acid is represented by a bead mapped on the $C^{\alpha}$ atom, while the DNA is represented by three beads resembling the ribose, base, and phosphate moieties. We used an integrative approach by which simulations aim to reproduce a series of experimentally obtained FRET efficiencies (**Methods**). For this purpose, we produced a set of additional fluorescently labelled Sox2 variants, designed to comprehensively probe discrete regions of the polypeptide chain, and measured transfer efficiency histograms and fluorescence lifetimes in the absence and presence of DNA (**Figure 3.4a,b, Supplementary Figure 3.9**). This yielded a total of 22 unique intramolecular FRET efficiencies (11 for free Sox2, 11 for bound Sox2), which were then matched in the simulations by tuning a single parameter, $\varepsilon_{pp}$, as it defines the interaction strength between the beads modelling the disordered regions of the protein (**Figure 3.4c**). It is important to note that the simulation was performed at equilibrium, i.e. it was not restrained by the measured FRET efficiencies. Instead, FRET efficiencies were back-calculated from simulated distance distributions, and compared with the experiment afterwards. As in previous studies, the scalable interaction strength between beads was set to 0.4 kJ $mol^{-1}$ (0.16 kT) and gave the best match to the experimentally-derived FRET efficiencies (**Figure 3.4d,e**). This approach has previously been shown to describe well the behaviour of several disordered proteins and protein-protein complexes with and without DNA[40,143,388,417,420]. Since no published structures are available for free Sox2, we used the same structure for both free and DNA-bound Sox2[366] . However, a simulation using a thus

far unpublished NMR structure of free Sox2 DBD deposited in the PDB (PDB code 2LE4) yielded near identical results (**Fig. S10**).

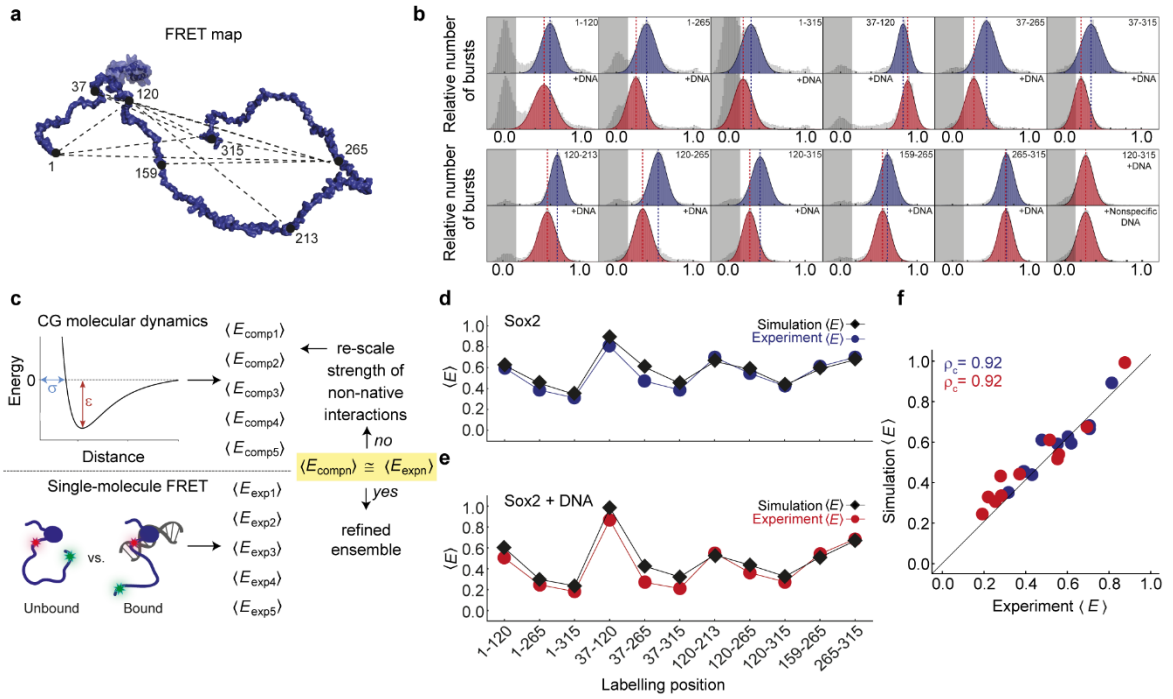



**Figure 3.4.** ***Langevin dynamics simulations reproduce FRET efficiencies from smFRET experiments. a)*** *Model of Sox2 showing FRET labelling positions that probe 11 unique intramolecular distances.* ***b)*** *Single-molecule transfer efficiency histograms of free and DNA-bound Sox2 variants, fluorescently labelled in different positions. The last panels (bottom, right) show that transfer efficiency changes for fluorescently labelled C-IDR are identical with specific and non-specific DNA.* ***c)*** *Schematic illustrating the CG computational approach. Using the Förster equation and a suitable polymer model, a series of computed FRET efficiencies (<E>) is obtained for each position labelled along the protein. The agreement between experimental and computed <E> is then refined by rescaling a single parameter ($\varepsilon_{pp}$) which uniformly defines the interaction strength between all beads in the intrinsically disordered domains, to finally obtain a refined ensemble. d-e) Comparison between computed (black) and experimentally-derived FRET efficiencies for d) free (blue) and e) DNA-bound (red) Sox2.* ***f)*** *Correlations between experimentally-derived and computed FRET efficiencies for both free (blue) and DNA-bound (red) Sox2. High correlation coefficients are obtained for both free Sox2 and Sox2 bound to DNA ($r_c$=0.92). Solid line is the identity line.*

The ensemble of both free and DNA-bound Sox2 collected from the simulated trajectories showed excellent agreement with the FRET efficiencies from experiments, yielding a concordance correlation coefficient $r_c$ of 0.92 (**Figure 3.4f**). Given that interactions between beads within intrinsically disordered stretches are set to a minimal value, our simple CG model implies that a considerable driving force for contact formation between the IDRs and DBD comes from charged residues, in agreement with the FRET and NMR data (**Figure 3.2, Figure 3.3**). We thus investigated how salt affects the dimensions of Sox2 by simulating Sox2 in its free and bound states at apparent salt concentrations ranging from 20 to 400 mM

(**Fig. S10**). The ensemble of free Sox2 expands as a function of salt concentration (**Fig. S10**) due to charge screening, but its dimensions reach a plateau at salt concentrations in proximity of the physiological range, in line with the experiments (**Figure 3.2f**). Charge screening thus has an important effect on the dimensions of Sox2.

An analysis of the collected ensembles revealed a highly dynamic C-IDR that explores a range of different conformations but to different degrees depending on whether Sox2 is in its free or DNA-bound state (**Figure 3.5a,b, Supplementary Movies 3.1 and 3.2**). In agreement with the experiments, the C-IDR dimensions are modulated by dynamic interactions with the DBD. Interestingly, an increase in salt concentration screens the interactions between the DBD and the AD1/AD2 domains with an effect that is proximity-dependent and more pronounced for AD1 (**Fig. S10**). When Sox2 binds to DNA, the C-IDR ensemble expands with more frequent excursions to extended states and thus a larger apparent $R_g$ (**Figure 3.5b**). The difference in contacts between the DBD and C-IDR for free and DNA-bound Sox2 shows that the expansion observed experimentally upon DNA binding is coincident with an increased number of contacts between the N-IDR and C-IDR, and decreased overall contacts of both IDRs with the DBD (**Figure 3.5c**). When Sox2 binds the DNA, the region experiencing the largest variation in contact space is the C-IDR AD1 (**Figure 3.5c,d,e)** directly in line with significant DNA-induced CSPs in the AD1 region (**Figure 3.3h**). Projecting the average number of contacts for each residue onto the Sox2 structure reveals an increase in proximity of regions overlapping with AD1 and DBD, when bound to DNA, but a decrease for AD2 (**Figure 3.5d**). This effect is clearer when we plot the fraction of contacts specifically between the DBD and the two ADs in the free and DNA-bound states (**Figure 3.5e**). On a residue-specific level, the difference in contacts between the C-IDR and DBD in the free and DNA bound states agrees reasonably well with the CSPs from our NMR experiments (**Supplementary Figure 3.11**). We also analyzed whether loss of C-IDR contacts with the DBD might be accompanied by formation of new contacts with the DNA but there was no enrichment in contact formation beyond a short ~20-residue stretch immediately flanking the DBD which is known to stably bind into the DNA major groove[366,368] (**Supplementary Figure 3.11**). Finally, the differential engagement of the two ADs is also re-iterated by analyzing the relaxation times of the contacts made by AD1 and AD2 with the DBD (**Supplementary Figure 3.11**). The correlation function of contact formation over time, fits better to a double exponential with distinct slow and fast components. The contact relaxation times for the two ADs are similar in the absence of DNA. However, in complex with DNA, the relaxation time for AD2 is reduced more than threefold compared to that of AD1 for which the contact lifetimes increases ($\tau_1^{AD1}/\tau_1^{AD2}$ = 0.8, $\tau_2^{AD1}/\tau_2^{AD2}$ = 0.7, for free Sox2; $\tau_1^{AD1}/\tau_1^{AD2}$ = 2.8, $\tau_2^{AD1}/\tau_2^{AD2}$ = 4.7 when bound to DNA) (**Supplementary Table 3.5**). These observations indicate that AD2 is accessible for a considerably longer time than AD1 when bound to DNA. Conversely, the serine-rich domain shows little difference in contact relaxation times before and after DNA binding (**Supplementary Figure 3.11**).

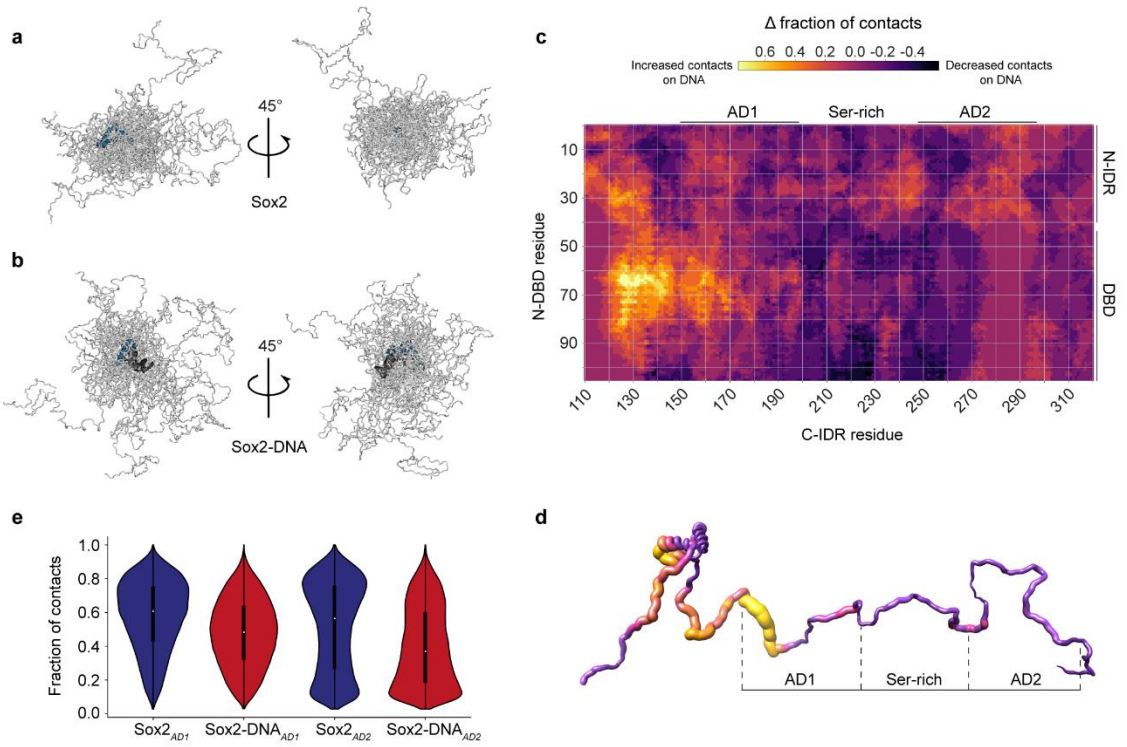



**Figure 3.5.** ***Dynamic structural model of Sox2 ensembles, free and in complex with DNA.*** ***a,b)*** *20 representative snapshots from the simulation for a) free Sox2 and b) DNA-bound Sox2. The DBD is shown in blue, C-IDR in light grey, DNA in dark grey.* ***c)*** *Difference in the fraction of intramolecular contacts in Sox2 in the unbound and DNA-bound states. Regions showing positive values report on increased contacts in the bound state, whereas regions that have negative values have decreased contacts.* ***d)*** *Difference in the fraction of intramolecular contacts in the unbound and DNA-bound states, projected on a schematic structure of Sox2. The color scale is the same as in panel c.* ***e)*** *Violin plots showing the fraction of contacts for residues within AD1 and AD2 with the DBD. The thick black lines represent the interquartile range and the white dots are the median values.*

## 3.4 Discussion

It remains a major experimental and computational challenge to determine the conformational ensembles of disordered proteins and, as in the case of TFs, to relate them to function. This challenge is thoroughly exemplified by a lack of both entries in the protein data bank and confident AlphaFold prediction of full-length TFs. In our work, we have reconstructed a detailed, experimentally-driven description of the structural ensembles for both free full-length Sox2 and Sox2 in complex with DNA. The relatively low number of charges in the C-IDR render it a weak polyampholyte[126], which is expected to populate a rather collapsed structure. However, we found that the C-IDR engages in additional dynamic but weak interactions with the DBD, driven mainly by charge interactions between the two domains. Notably, the dimensions of Sox2 are very sensitive to salt in the range corresponding to physiological concentrations; local differences in intracellular salt concentrations would be expected to further tune the accessibility of the C-IDR. This is a

noteworthy observation: even though the charges in the C-IDR are relatively sparse and well distributed, their interactions with the DBD still confer a strong effect on the overall dimension of the protein. Keeping the C-IDR in a relatively compact state in the absence of DNA may be an evolved strategy to protect against unwanted interactions, premature degradation, aggregation, or condensate formation. Interactions with the DBD may also aid in keeping an otherwise aggregation prone C-IDR[421] soluble, until the right genomic binding site or coregulator is located. We found, in agreement with others[368,374], that DNA binding affinity was unaffected by interdomain interactions; sustaining sufficiently weak interactions that maintain the advantages of a highly dynamic ensemble may be crucial to modulate the accessibility of the ADs without disturbing DNA binding. Intramolecular interactions for other nucleic acid binding proteins have been reported to influence binding affinity, mostly through interactions mediated by strongly charged but short regions[395-397]. Other TFs have recently been reported to have similar interdomain interactions, including cases such as B-MYB where a short and strongly positively charged region interacted with the DBD but with little effects on DNA binding affinity[422]. The TFs p53 and MYC/MAX have also been demonstrated to partake in intramolecular electrostatic interactions with their DBDs, to a degree dependent on their phosphorylation state[415,423,424]. For p53, interdomain interactions had no effect on binding affinity to specific DNA but led to a 5-fold affinity reduction to non-specific DNA, thus increasing specificity, a scenario not recapitulated by Sox2. Sox2 has several phosphorylation sites in the C-IDR, which may enable tuning of DNA binding affinity or specificity[69]. For example, phosphorylation of Thr116, adjacent to the DNA binding HMG box, has been shown to be necessary for recruitment to certain stem-cell dependent promoters[425]. Given the relatively few charged residues in the C-IDR of Sox2, a single PTM that affects the charge state might have a large effect on the magnitude of interdomain interactions, potentially leading to ultra-sensitivity in IDR dimensions and thus immediate shaping of the Sox2 interactome.

Our CG model shows that when using interdomain contacts and overall dimensions as an indirect proxy for accessibility, we observe changes in accessibility upon binding DNA which localize largely to regions overlapping with the ADs, harbouring many charged residues. Interestingly, part of AD1 shows decreased accessibility upon DNA binding (**Figure 3.5c,d,e**) whereas much of the remainder of the C-IDR, including AD2, has more than fourfold increased accessibility when viewed through the lens of relaxation times. Even though the precise boundaries of ADs remain to be defined, our results show variable responses of discrete C-IDR regions to DNA binding. It is likely that a combination of residue proximity to the DBD and DNA, charge number, and charge distribution[126] will dictate the exact conformational pattern for specific TFs, but deciphering the details of that code is an important future task. Addition of negative charges, e.g. in the form of phosphorylations, might be expected to enhance the interaction with the positively charged DBD and thus increase occupancy in a compact ensemble, rendering ADs more or less accessible to coregulators dependent on the sequence position of the negative charge. For example, there is a conserved positive region flanking the DBD on the C-terminal side that we observe to stably interact with the DNA in our simulations, in agreement with previous studies[368]. Phosphorylations in this region (e.g., on Thr116[426]) would be expected to decrease interactions with DNA, potentially increasing the accessibility of AD1, while a phosphorylation further downstream in the sequence (e.g., Ser251) might increase interactions with the DBD leading to decreased accessibility of AD2. Increased interactions between the C-IDR and DBD might in some cases lead to less efficient DNA binding, which could explain why Sox2 binds certain enhancers less when phosphorylated in Ser251 which

is close to AD2[427]. Nonetheless, PTM effects are complex and more intricate than simple modulation of interdomain interaction strength. Generally, our structural model of Sox2 will aid in rationalizing the effects of PTMs as well as linking them to conformational changes and cofactor binding.

The pioneer activity of Sox2 is dependent on its ability to bind to and alter the structure of nucleosomes[52,365,366]. Upon binding nucleosomes, the Sox2 C-IDR goes through similar, albeit not identical, conformational rearrangements to those that follow its binding to short DNA, suggesting that our reconstructed ensemble will also be generally populated on nucleosomes. Fluorescence lifetime analysis in the nucleosome-bound state showed that the C-IDR is slightly less dynamic on the submillisecond timescale when compared with the DNA-bound state; whether this is due to steric restrictions in the local conformational space or due to a direct interaction with the core histones is currently unknown. Nonetheless, a compelling hypothesis is that the exact nature of the Sox2 binding site, i.e. whether it is on free DNA or in different locations on a nucleosome particle, will dictate the degree of AD accessibility and the resulting interaction profile. Binding experiments with interaction partners are needed to reveal whether that is a feasible model but it would offer possibilities for specifically targeting interactions with nucleosome-bound Sox2 while excluding those that involve accessible DNA. Our structural model is a first step in that direction and creates a platform for mapping the effects of mutations, environment, and binding partners on the structural ensemble of the Sox2 IDR.

Members of the SoxB family of TFs (Sox1, Sox2, and Sox3) share general composition features in their IDRs, such as the number and position of charges, and therefore the conformational dynamics that we observe for Sox2 are likely to be closely applicable to this family (**Supplementary Figure 3.12**). Beyond the SoxB family, these types of interdomain interactions may be very common among TFs to restrain and finely tune the accessibility of ADs to varying degrees before and after they have located their binding sites. In fact, AlphaFold predictions and bioinformatics analysis support that most TFs share a similar architecture and charge profile (positively charged DBD, modest numbers of charges in IDRs)[390]. Further studies will reveal whether the accessibility tuning modulates the interaction equilibrium of TFs with coactivators within the transcriptional machinery. Finally, this type of ensemble redistribution with expansion excursions on DNA may also be linked to condensate formation, which has been suggested to be involved in transcriptional regulation, potentially rendering phase separation more likely to occur once TFs have located their DNA or nucleosome targets.

**Acknowledgements**

We thank Benjamin Schuler and Daniel Nettels for support using the data analysis toolbox Fretica. We thank Beat Fierz for technical assistance with nucleosome preparation. We thank Magnús Kristjánsson, Erna Magnúsdóttir, and Erik Holmstrom for stimulating discussions, and Matti Már, Sarah Ruidiaz, and Mahtab Hafizi for technical assistance. This work was supported by funding from the European Research Council (ERC StG, 101040601-PIONEER, to P.O.H.), the Icelandic Research Fund (grant nr. 2659105, to P.O.H.), the University of Iceland Doctoral Fund (to P.O.H.), the Max Planck Society (to J.T.B.), the University of Auckland (to J.A.P.M.), the Novo Nordisk Foundation to the Challenge centre REPIN (#NNF18OC0033926 to B.B.K.), and by cOpenNMR, an infrastructure granted from the Novo Nordisk Foundation (#NNF18OC0032996).

## 3.5 Methods

### 3.5.1 Protein expression and purification

The DNA coding for all Sox2 constructs was inserted into a modified pET24b vector. The vector contains codes for a hexahistidine small ubiquitin-like modifier ($His_6$-SUMO) tag added to the N-terminal of the constructs. Mutants were made using the QuikChange Lightning kit from Agilent using primers from Integrated DNA Technologies (IDT). All constructs were expressed in Lemo21(DE3) cells (New England BioLabs) cultured in LB-broth medium, or M9 minimal medium containing $^{15}N$-$NH_4Cl$ or $^{15}N$-$NH_4Cl$ and $^{13}C_6$-glucose. Expression was induced at $OD_{600}$ 0.5-0.7 with 0.4 mM Isopropyl β- d-1-thiogalactopyranoside (IPTG) and cells were grown for 2-3 hours at 37°C with vigorous shaking. Cells were harvested by centrifugation at 4500 × g for 15 min and resuspended in Buffer A (50 mM $NaH_2PO_4$, 300 mM NaCl, 10 mM imidazole, 6 M urea, 1 mM dithiothreitol (DTT), pH 8.0) for overnight lysis at 4°C. The soluble fraction was collected by centrifugation at 40,000 × g for 1 hour at 4°C and loaded onto a 5 ml HisTrap HP column (Cytiva) equilibrated with Buffer A. The column was washed with 10 column volumes (CV) of Buffer A and eluted with Buffer A with imidazole concentration adjusted to 500 mM. Eluted samples were dialyzed overnight against Buffer B (50 mM Tris, 150 mM NaCl, 1 mM DTT, pH 8.0), followed by ULP1 protease (made in-house) cleavage to remove the $His_6$-SUMO tag. Following cleavage, the N-DBD and DBD constructs were dialyzed against Buffer C (50 mM $NaH_2PO_4$, 6 M Urea, pH 8.0) overnight and loaded onto a 5 ml HiTrap SP Sepharose FF column (Cytiva). The SUMO tag eluted during the 10 CV wash step (Buffer C) and the proteins were eluted with Buffer C with NaCl concentration adjusted to 500 mM. Full-length Sox2 and C-IDR precipitated from solution following the removal of the $His_6$-SUMO tag, the precipitate was recovered by centrifugation and resuspended in Buffer C. All protein preparations were concentrated using Amicon Ultracentrifugal filters (Merck), reduced with DTT and purified by reversed-phase high-performance liquid chromatography (RP-HPLC) using a ZORBAX 300SB-C3 column (Agilent) with flow rate of 2.5 ml/min starting at 95% RP-HPLC solvent A (99.9% $H_20$, 0.1% trifluoroacetic acid (TFA)(Sigma) and 5% RP-HPLC solvent B (99.9% acetonitrile, 0.1% TFA) and going to 100% RP-HPLC solvent B over 95 minutes. Protein purity was analysed by SDS-PAGE, identity confirmed by mass spectrometry, and samples were lyophilized and stored at -20°C.

### 3.5.2 Protein labelling

Lyophilized proteins were resuspended in labelling buffer (0.1 M potassium phosphate, 1 M urea, pH 7.0) and labelled overnight at 4°C using Cy3B maleimide (donor) (Cytiva) (0.7:1 dye to protein ratio). The reaction was quenched using DTT and RP-HPLC was then used to remove unreacted dye, and separate unlabelled and double donor-labelled proteins. The proteins were lyophilized overnight, then resuspended in labelling buffer and labelled overnight at 4°C using CF660R maleimide (acceptor) (Sigma). The reaction was quenched using DTT and RP-HPLC was then used to remove unreacted dye, and separate donor-donor doubly labelled and acceptor-acceptor doubly labelled proteins. Donor-acceptor labelled proteins were lyophilized, resuspended in 8 M GdmCl, frozen in liquid $N_2$, and stored at -80°C.

### 3.5.3 DNA labelling

Aliquots of 5-10 nmol oligonucleotide (oligonucleotides contained a thymine modified with a C6-amino linker for the reaction with the NHS ester of the dyes) (IDT) were dissolved in 50 µl DNA labelling buffer (0.1 M sodium bicarbonate, pH 8.3) and labelled with either Cy3B NHS ester (Cytiva) or CF660R NHS ester (Sigma) in a 2:1 dye to DNA ratio. The reaction was incubated for at least two hours at room temperature, then ethanol precipitated to remove excess dye. Pellet was redissolved in 100 µl of 95% RP-HPLC solvent C (0.1 M triethylammonium acetate) and 5% RP-HPLC solvent D (acetonitrile) and separated from the unreacted dye and unlabelled oligonucleotide with RP-HPLC using a ReproSil Gold 200 C18 column (Dr. Maisch), labelled oligonucleotides were collected and lyophilized. Oligonucleotides intended for PCR amplification were resuspended in double distilled water ($ddH_2O$) to a final concentration of 2.5 µM and stored at -20°C. Oligonucleotides intended for smFRET measurements were resuspended in DNA annealing buffer (10 mM Tris, 50 mM NaCl, 1 mM EDTA, pH 7.5) and mixed with equimolar amounts of the reverse compliment oligonucleotide labelled with either Cy3B or CF660R. Sample was placed on a heating block at 95°C for 5 minutes, heating was turned off and samples allowed to cool slowly to room temperature to anneal the donor labelled and the acceptor labelled oligonucleotide strands. Labelled DNA was aliquoted, frozen in liquid $N_2$, and stored at -80°C.

### 3.5.4 Nucleosome reconstitution

PCR amplification of a pJ201 plasmid containing the 147 bp Widom sequence was used to generate DNA for nucleosome reconstitution. The amplification took place using either fluorescently labelled oligonucleotides (see DNA labelling) or unlabelled oligonucleotides (IDT). The oligonucleotides were designed to insert a Sox2 binding site (CTTTGTTATGCAAAT) and to extend the 147 bp Widom sequence by 25 bp linkers on either side. The PCR reactions were ethanol precipitated before being purified using a DNA Clean and Concentrator Kit (Zymo Research). The concentration of the DNA was determined by UV Vis. For the list of primer and DNA sequences see Supplementary table 3. To reconstitute nucleosomes 10 pmol of purified 197 bp Widom sequence containing a Sox2 binding site were used. The DNA was mixed with 1.0-1.75 molar equivalents of recombinant core histone octamer (The Histone Source) in 10 mM Tris, 0.1 mM EDTA, 2 M KCl, pH 7.5, on ice. The reaction was then transferred to a Slide-A-Lyzer MINI dialysis button (Thermo Fisher Scientific) and dialyzed against a linear gradient of 10 mM Tris, 0.1 mM EDTA, 10 mM KCl, pH 7.5 over 20 hours at 4°C. Constant volume of buffer was maintained by removing buffer at the same rate as fresh buffer with 10 mM KCl was added using a peristaltic pump. Samples were transferred to microcentrifuge tubes and centrifuged for 5 minutes at 20.000×g, 4°C to remove aggregates, supernatant was collected. Concentration was determined via absorbance at 260 nm and 0.5 pmol of the reaction was loaded on a 0.7% agarose gel and run for 90 minutes at 90 V with 0.25× Tris-borate as running buffer. Following staining with GelRed (Biotium) gels were imaged using Gel Doc EZ gel system (Bio-Rad). Only samples that contained <5% free DNA were used for measurements.

### 3.5.5 Single-molecule spectroscopy

All single molecule fluorescence experiments were conducted at 23°C using a MicroTime 200 (PicoQuant) connected to an Olympus IX73 inverted microscope. The donor dye was excited using a 520 nm diode laser (LDH-D-C-520, PicoQuant) using pulsed interleaved excitation[428](PIE) with a 640 nm diode laser (LDH-D-C-640, PicoQuant) to alternate excitation of donor and acceptor dyes with a repetition rate of 40 MHz. The laser intensities were adjusted to 40 µW at 520 nm and 20 µW at 640 nm (PM100D, Thorlabs). Excitation and emission light was focused and collected using 60× water objective (UPLSAPO60XW, Olympus). Emitted fluorescence was focused through a 100 µm pinhole before being separated first by polarization and then by donor (582/64 BrightLine HC, Semrock) and acceptor (690/70 H Bandpass, AHF) emission wavelengths, into four detection channels. Detection of photons took place using single photon avalanche diodes (SPCM-AQRG-TR, Excelitas Technologies). The arrival time of detected photons was recorded with a MultiHarp 150P time-correlated single photon counting (TCSPC) module (PicoQuant). All experiments were performed in µ-Slide sample chambers (Ibidi) at RT in TEK buffer (10mM Tris, 0.1mM EDTA, pH 7.4) with varying KCl concentrations. For photoprotection 143 mM 2-mercaptoethanol (Sigma) was added, along with 0.01% (v/v) Tween-20 (AppliChem) to reduce surface adhesion. In experiments using denaturants, the exact concentration of denaturant was determined from measurement of the solution refractive index[429].

### 3.5.6 Analysis of transfer efficiency histograms

Data for transfer efficiency histograms were collected from 50-100 pM of freely diffusing double labelled Sox2, DNA or nucleosomes. All data was analysed using the Mathematica scripting package "Fretica" (https://schuler.bioc.uzh.ch/programs/) developed by Daniel Nettels and Ben Schuler. Fluorescence bursts were first identified by combining all detected photons with less than 100 µs interphoton times. Transfer efficiencies within each fluorescence burst were calculated according to $E = n'_A/(n'_A + n'_D)$, where $n'_A$ and $n'_D$ are the number of acceptor and donor photons, respectively. The number of photons were corrected for background, direct acceptor excitation, channel crosstalk, differences in dye quantum yields and photon detection efficiencies[413]. The resulting bursts were then filtered to remove bursts where the acceptor bleaches during the transit of the molecule through the confocal volume[430], which otherwise can cause a bias towards lower FRET. Occasional fluorescence bursts with photon counts more than three times higher than the mean signal binned at 1 s, corresponding to aggregates, were removed before data analysis. The labelling stoichiometry ratio (S) was determined according to:

$$S = \frac{n_D^D + n_A^D}{n_D^D + n_A^D + n_A^A}$$

where $n_{D/A}^D$ is the number of detected donor or acceptor photons after donor excitation and $n_A^A$ is the number of detected acceptor photons after acceptor excitation. To construct the final transfer efficiency histograms, we selected bursts that have $S = 0.3\text{–}0.7$ which allowed us to filter out bursts that originate from molecules that lack an active acceptor. In some cases, a large donor-only population can cause residual donor-only bursts to remain after filtering.

To extract mean FRET efficiencies, the histograms were fitted to an appropriate number of Gaussian or logNormal distribution function, corresponding to one or more populations. Multiple transfer efficiency histograms for binding affinity analysis were fitted globally, where some parameters were shared across different measurements. For distance calculations based on the transfer efficiencies for DNA and nucleosomes the Förster equation

$$E(r) = \frac{1}{1 + r^6/R_0^6}$$

was used with $R_0 = 6.0$ nm for a Cy3B/CF660R dye pair. For double labelled proteins involving disordered segments we converted mean transfer efficiencies $\langle E \rangle$ to root-mean-square end-to-end distances $R = \sqrt{\langle r^2 \rangle}$ by numerically solving the following transcendental equation:

$$\langle E \rangle = \int_0^\infty dr\, E(r) P(r)\ .$$

Here, $P(r)$ denotes the distance probability density function of the SAW-$\nu$ model[431], given by

$$P(r) = A \frac{4\pi}{R} \left(\frac{r}{R}\right)^{2+\frac{\gamma-1}{\nu}} \exp\left(-\alpha \left(\frac{r}{R}\right)^{\frac{1}{1-\nu}}\right),$$

which is characterized by the critical exponents $\nu$ and $\gamma \approx 1.1615$. The constants $A$ and $\alpha$ are determined by requiring $P(r)$ to be normalized and to satisfy $\langle r^2 \rangle = R^2$, respectively. The dependency on $\nu$ in $P(r)$ is removed by assuming that a scaling law $R = b\, N^\nu$ must hold and substituting $\nu = \ln\left(\frac{R}{b}\right) / \ln(N)$ into the expression for $P(r)$, where $b \approx 0.55$ nm for proteins and $N$ denotes the number of monomers between the fluorescent groups. The associated radius of gyration $R_g$ can be approximated as

$$R_g \approx R \sqrt{\frac{\gamma(\gamma+1)}{2(\gamma+2\nu)(\gamma+2\nu+1)}}.$$

In denaturation experiments, the Förster radius was corrected for changes in refractive index according to[417]:

$$R_0^6(c_D) = R_{0,0}^6 \left(\frac{n R_0^6}{n(c_D)}\right)^4$$

where $n(c_D)$ denotes the refractive index of the sample at denaturant concentration $c_D$.
Fluorescence anisotropy values were determined for fluorescently labelled variants using polarization-sensitive detection in the single-molecule instrument[432], and were between 0.04 and 0.14 both for the monomeric proteins and the proteins in complex with DNA, indicating sufficiently rapid orientational averaging of the fluorophores to justify the approximation $\kappa^2 \approx 2/3$ used in Förster theory[433].

### 3.5.7 Fluorescence correlation spectroscopy

To determine the diffusion time of labelled Sox2, we performed fluorescence correlation spectroscopy by correlating the intensity fluctuations in fluorescence in an smFRET experiment according to

$$G_{ij}(\tau) = \frac{\langle \delta n_i(0) \delta n_j(\tau) \rangle}{\langle n_i \rangle^2}$$

where $i,j$=$A$, $D$ and $n_i(0)$ and $n_j(\tau)$ are fluorescence count rates for channels $i$ and $j$ at time 0 and after a lag time t, respectively, and $\delta n_{i,j} = n_{i,j} - \langle n_{i,j} \rangle$ are the corresponding deviations from the mean count rates.

Data for nsFCS[434] were collected using continuous-wave excitation at 520 nm and a ~100-pM sample of double-labelled free Sox2, DNA-bound Sox2, or isolated C-IDR. Donor and acceptor fluorescence photons from only the FRET subpopulation were used for correlations at 1 ns binning time. Photons were cross-correlated between detectors to avoid the effects of detector dead times and after-pulsing on the correlation functions. Cross-correlation curves between acceptor and donor channels were fit and analyzed as described previously[40]. Briefly, the correlation curves were fit over lag time interval from -1 µs to +1 µs using

$$g_{ij}(\tau) = a\left(1 - c_{ab}e^{\frac{-|\tau|}{\tau_{ab}}}\right)\left(1 + c_{cd}e^{\frac{-|\tau|}{\tau_{cd}}}\right)$$

where $i$ and $j$ indicate donor ($D$) or acceptor ($A$) fluorescence emission; the amplitude $a$ depends on the effective mean number of molecules in the confocal volume and on the background signal; $c_{ab}$, $\tau_{ab}$, $c_{cd}$ and $\tau_{cd}$ are the amplitudes and time constants of photon antibunching (ab) and chain dynamics (cd), respectively. $\tau_{cd}$ can be converted to the reconfiguration time of the chain, $\tau_r$, by assuming that the chain dynamics can be modeled as a diffusive process in the potential of mean force derived from the sampled inter-dye distance distribution $P(r)$[434,435] based on the SAW-ν model[413,436]. The correlation functions are displayed with a normalization to 1 at their respective values at 0.5 µs.

### 3.5.8 Fluorescence lifetime analysis

Fluorescence lifetimes were estimated from the mean donor detection times $\langle t_D \rangle$ after their respective excitation pulse. The fluorescence lifetimes were then plotted against corresponding transfer efficiencies in two-dimensional scatter plots, where $\tau_{DA}/\tau_D = \langle t_D \rangle/\tau_D$ was calculated for each burst for an intrinsic donor lifetime $\tau_D$. For a fixed distance between the donor and acceptor, the ratio $\langle \tau_{DA} \rangle/\tau_D$ must equal $1 - E$ (**Fig. 1c**, diagonal line), whereas for systems that rapidly sample a broad distance distribution this ratio significantly deviates from $1 - E$. For a rapidly fluctuating distance described by a probability density function $P(r)$ of the interdye distance $r$, the distribution of distances affects the average fluorescence lifetime $\langle \tau_{DA} \rangle$ according to

$$\frac{\langle \tau_{DA} \rangle}{\tau_D} = 1 - \langle E \rangle + \frac{\sigma^2}{1 - \langle E \rangle}.$$

Here, the variance $\sigma^2$ is given by

$$\sigma^2 = \langle E^2 \rangle - \langle E \rangle^2 = \int_0^\infty dr\, [E(r) - \langle E \rangle]^2 P(r)\,.$$

### 3.5.9 Determination of denaturant association coefficients

Association constants ($K_a$) of GdmCl and urea were determined using a weak denaturant binding model[437,438] with the form

$$E(c_D) = \frac{E_0 + \Delta E\, K_a\, c_D}{1 + K_a\, c_D}$$

where $c_D$is the denaturant concentration, with $K_a$, $\Delta E$, and $E_0$ being fit parameters.

### 3.5.10 Binding affinity measurements

Transfer efficiency histograms were recorded for either double labelled Sox2 or DNA with increasing concentration of unlabelled binding partner until the transfer efficiency remained stable. Gaussian peak functions were used to fit the histograms into two subpopulations, bound and unbound. From the relative areas of these subpopulations the fraction of bound species ($\theta$) could be quantified. To aquire the dissociation constant ($K_D$) a binding isotherm was fit using

$$\theta = \frac{c_{X,tot} + K_D + c_{Y,tot} + \sqrt{\left(c_{X,tot} + K_D + c_{Y,tot}\right)^2 - 4c_{X,tot}c_{Y,tot}}}{2c_{Y,tot}}$$

where $c_{X,tot}$ and $c_{Y,tot}$ are the total concentrations of Sox2 or DNA, depending on which molecule is kept at a constant concentration.

### 3.5.11 CD spectroscopy

Far-UV CD spectra were recorded on a Jasco J-1100. All spectra were recorded at 25°C in 25 mM $NaH_2PO_4$, 25 mM NaCl at pH 8.0 using a 1 mm cuvette. Spectra were recorded between 250 and 190 nm, data pitch was 0.1 nm, digital integration time of 0.25 s, scan speed 20 nm/min and accumulating 3 scans. Protein concentrations ranged from 2-5 µM. Identical measurements were taken of the buffer, which was then subtracted from the measurements. The ellipticity was converted to mean residual ellipticity using

$$MRE = \frac{mdeg}{10 \times L \times C \times N},$$

where $L$ is the path length in cm, $C$ is the concentration in molar, and $N$ is the number of peptide bonds.

### 3.5.12 NMR spectroscopy

All NMR spectra were recorded on a Bruker Avance Neo 800 MHz spectrometer or Avance III HD 750MHz spectrometer equipped cryogenic probe. Samples were recorded in 20 mM $NaH_2PO_4$, 50 mM NaCl, 5 mM DTT, 125 µM DSS, 5% $D_2O$ (v/v) at pH 5.5 and 15°C to minimize amide exchange. The raw free induction decays (FIDs) were transformed using NMRPipe[439] and analysed using CcpNmr software[440]. Backbone nuclei of $^{13}C$,$^{15}N$-labelled Sox2 were assigned in the unbound state (110 µM $^{13}C$-$^{15}N$-labelled Sox2 from analysis of $^{1}H^{15}N$ HSQC, HNCACB, CBCA(CO)NH, HN(CO)CA, HNCO, and HN(CA)NNH multidimensional NMR spectra (BMRB accession number 51964)). The intensity of backbone resonances from the DBD were too weak in full-length Sox2 for direct assignments but could be transferred from assignments of the isolated N-DBD (**Figure 3.2 and Fig. S1**). Secondary structure content in Sox2 was determined from secondary $C^{\alpha}$ chemical shifts using a random coil reference for intrinsically disordered proteins[441].

$T_1$ and $T_2$ $^{15}N$ relaxation times were determined from 2×2 series of $^{1}H^{15}N$ HSQC spectra with varying relaxation delays and using pulsed-field gradients for suppression of solvent resonances. The series were recorded at 800 MHz (1H), using 8 (20 ms, 60 ms, 100 ms, 200 ms, 400 ms, 600 ms, 800 ms and 1200 ms) and 8 (0 ms, 33.9 ms, 67.8 ms, 101.8 ms, 135.7 ms, 169.6 ms, 203.5 ms and 271.4 ms) different relaxation delays for $T_1$ and $T_2$, respectively. CcpNmr Analysis software[440] was used to fit the relaxation decays to single exponentials and determine relaxation times.

Binding induced weighted CSPs were measured at a protein concentration of 30 µM in absence and presence of a 1.1 fold excess unlabelled double-stranded DNA with Sox2 binding sequence (**Supplementary Table 3**). CSPs were calculated as[442]

$$CSP = \sqrt{\frac{1}{2}((\Delta\delta_H)^2 + (0.154 * \Delta\delta_N)^2}$$

The dissociation constant $K_d$ for DBD/C-IDR interactions was quantified using chemical shift perturbation analysis[443], by employing the observed chemical shift changes ($\Delta\delta_{obs}$) of $^{15}N$-labelled C-IDR upon the addition of unlabelled N-DBD. $K_d$ was calculated using:

$$CSP = \sqrt{\frac{1}{2}((\Delta\delta_H)^2 + (0.154 * \Delta\delta_N)^2}$$

where $\Delta\delta_{obs}$ is the observed chemical shift change, $\Delta\delta_{max}$ is the maximum chemical shift change, $[P]_t$ is the total C-IDR concentration, $[L]_t$ is the total N-DBD concentration[444]. The resulting chemical shift changes were fitted to the formula and $K_d$ and $\Delta\delta_{max}$ were determined.

### 3.5.13 Simulations

***Protein model.*** The all-atom starting structure for Sox2 was obtained from the electron microscopy structure of Sox2 bound to a nucleosome (PDB: 6T7B[366]). The disordered regions, not available in the starting structure, were modelled using the modeller plugin[445] embedded in UCSF Chimera[446]. Each residue of Sox2 was represented as a single bead mapped to the $C^{\alpha}$ atom of the starting full-atom structure. The simulation parameters used in the current work are identical to those outlined in Heidarsson et al[420]. We used the following potential energy function describing protein-protein interactions[40]:

$$V_{protein-protein} = \frac{1}{2}\sum_{i=1}^{N} k_b(d_i - d_i^0)^2 + \frac{1}{2}\sum_{i=1}^{N} k_\theta(\theta_i - \theta_i^0)^2 + \sum_{i=1}^{N-2}\sum_{m=1}^{4} k_{i,m}\left(1 + cos\left(n\phi_i - \delta_{i,m}\right)\right) + \sum_{i<j}\frac{q_i q_j}{4\pi\epsilon_d\epsilon_0 d_{ij}} e^{-\frac{d_{ij}}{\lambda_D}} + \sum_{(i,j)\,\in Native} \varepsilon_{ij}\left[13\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right] + \sum_{(i,j)\,\notin Native} 4\varepsilon_{pp}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right]$$

where the first three terms describe bonded while the second three non-bonded interactions. Bonds and angles (first and second terms, respectively) are treated with harmonic potentials with force constants $k_{\mathrm{b}}$, $k_\theta$ for bond lengths and equilibrium values $d_i^0$ and $\theta_i^0$ for angles. Both assignments are based on the distances and angles between the $C^{\alpha}$ atoms in the all-atom starting structure. A cosine-based dihedral potential (third term) was used to sample the behaviour of four beads linked by three bonds, with the dihedral angle parameters described by the force constant $k_{i,m}$ and a phase shift term $\delta_{i,m}$. These parameters are obtained from a sequence-specific dihedral potential, informed by structures deposited in the RCSB[447]. Electrostatic interactions are described in the fourth term using a screened Coulomb potential. While lysine and arginine are assigned a charge of +1, aspartate and glutamate are given a charge of -1 and histidine a charge of +0.5, considering that the imidazole side chain in histidine usually holds a p$K_a$ of ~6.0. The charge of all the other beads was set to 0. The Coulomb term is composed of terms pertaining to the charge of a residue ($q_i$), the dielectric constant of water ($\epsilon_{\mathrm{d}}$) set to a value of 80, the permittivity of the medium ($\epsilon_0$), and the Debye screening length ($\lambda_{\mathrm{D}}$), which is given by:

$$\lambda_{\mathrm{D}} = \frac{\epsilon_0\epsilon_{\mathrm{d}}k_{\mathrm{B}}T}{2N_{\mathrm{A}}e^2I},$$

where the Boltzmann constant and temperature are described by $k_{\mathrm{B}}$ and $T$, respectively, in addition to the Avogadro's number $N_{\mathrm{A}}$, the elementary charge $e$ and the ionic strength $I$. As such, different ionic strength values were mimicked by altering the Debye screening length.

The fifth and sixth terms collectively describe short-range attractive interactions between beads, separated by a distance $d_{ij}$. Native interactions pertain to the folded domains and are computed using a 12-10-6 pair potential, which has been successfully adopted in a Gō-model employed to investigate protein folding by Karanikolas and Brooks[447]. In this potential, $\varepsilon_{ij}$ describes the strength of the interaction calculated in accordance with a native-centric model[447], with $\sigma_{ij} = \frac{(\sigma_i + \sigma_j)}{2}$, determined based on $C^{\alpha}$-$C^{\alpha}$ distances in the crystal structure. Conversely, the interaction between residues located in the disordered regions and between disordered regions and the folded Sox2 domain, is described by a simpler 12-6 Lennard-Jones potential with $\varepsilon_{pp}$ set to a value of 0.16 $k_{\mathrm{B}}T$ (~0.4 kJ mol$^{-1}$) and $\sigma_{ij}$ to a fixed value of 0.6 nm. These values have previously been effective in giving the best agreement with experientially derived FRET efficiencies[40,420].

***DNA model.*** The CG representation used for the DNA is comprised of three beads representing the phosphate, ribose and base moieties of nucleic acids, mapped to the P, C4' and N1 atoms in the all-atom DNA structure, respectively. All phosphate beads were assigned a charge of -1, while ribose and base beads were not charged. Initially, to obtain a reliable model of the Sox2-DNA binding site, a segment of the nucleosome containing the Sox2 consensus sequence with Sox2 bound to it was taken from the electron microscopy structure with accession code 6T7B[366]. This fragment of DNA was then mutated to match the DNA sequence used in experiments. This modelling strategy ensured a lower strain between the bound Sox2 and the segment of DNA, which would have otherwise been modelled as a straight DNA segment, while Sox2 preferentially binds to curved, nucleosomal DNA.

The interactions between DNA beads are given by the following potential energy function:

$$V_{protein-DNA} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_{\mathrm{d}}\epsilon_0 d_{ij}} e^{-\frac{d_{ij}}{\lambda_D}} + \sum_{(i,j)\in\mathrm{native}} \varepsilon_{ij}\left[13\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right] + \sum_{(i,j)\in\mathrm{native}} 4\varepsilon_{pd}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right].$$

While native contacts describe interactions between DNA and Sox2 DBD, non-native contacts address interactions between DNA and the disordered tail regions of Sox2. Native contacts between bead pairs were identified from the all-atom starting structure, using a cutoff-based analysis of the crystal structure of the Sox DBD-DNA complex (PDB accession code: 6T7B[366]). If the distance between any atom of a protein residue and any atom of a nucleotide would fall below 0.5 nm, the contact was considered native and the strength of the interaction $\varepsilon_{ij}$ would be set to 2 $k_{\mathrm{B}}T$ (~5 kJ mol$^{-1}$). Otherwise the contact was considered non-native and $\varepsilon_{\mathrm{pd}}$ was set to 0.06 $k_{\mathrm{B}}T$ (~0.15 kJ mol$^{-1}$). For all contacts, $\sigma_{ij}$ was set to a value of 0.5 nm. The values of $\varepsilon_{ij}$ and $\varepsilon_{pd}$ have previously been optimized to yield the best agreement between experimental and simulated FRET efficiencies[420].

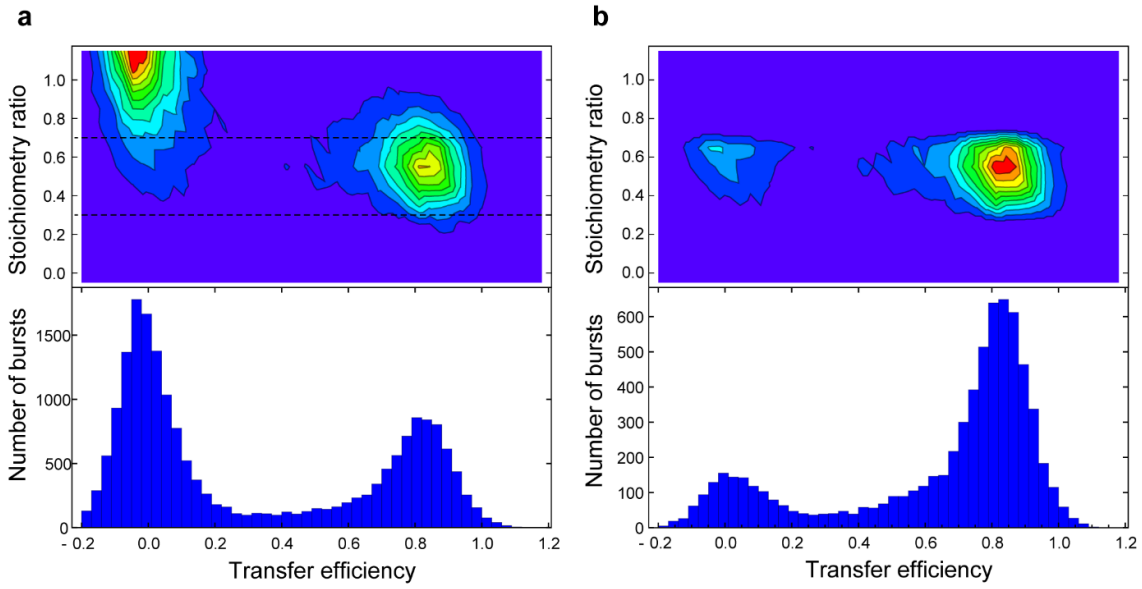
***Langevin dynamics simulations of the protein and protein-DNA complexes.*** Langevin dynamics simulations of the protein in isolation and bound to DNA were performed using GROMACS version 5.1.4[448]. Each system was placed at the centre of a cubic box measuring 30 and 120 nm$^3$ for the proteins and protein-DNA complexes, respectively. All simulations were performed using periodic boundary conditions and charge screening was obtained

considering the effect of monovalent salt at concentrations ranging from 40-800 mM, mimicked by adjusting the Debye length $\lambda_D$. After energy minimisation, each system was simulated for a total of 20 μs (4 replicates of 5 μs, with the first 1 μs of each replicate considered as equilibration time and removed). From the simulations, mean FRET efficiencies were calculated based on the distance distributions of the fluorescently labelled residues/beads, using the Förster equation. The Förster radius, $R_0$, was set to 6.0 nm as it corresponds to the $R_0$ of the Cy3b-CF660R dye pair used in experiments. All analyses were performed using tools available in the GROMACS suite, custom in-house scripts or MDAnalysis[449].

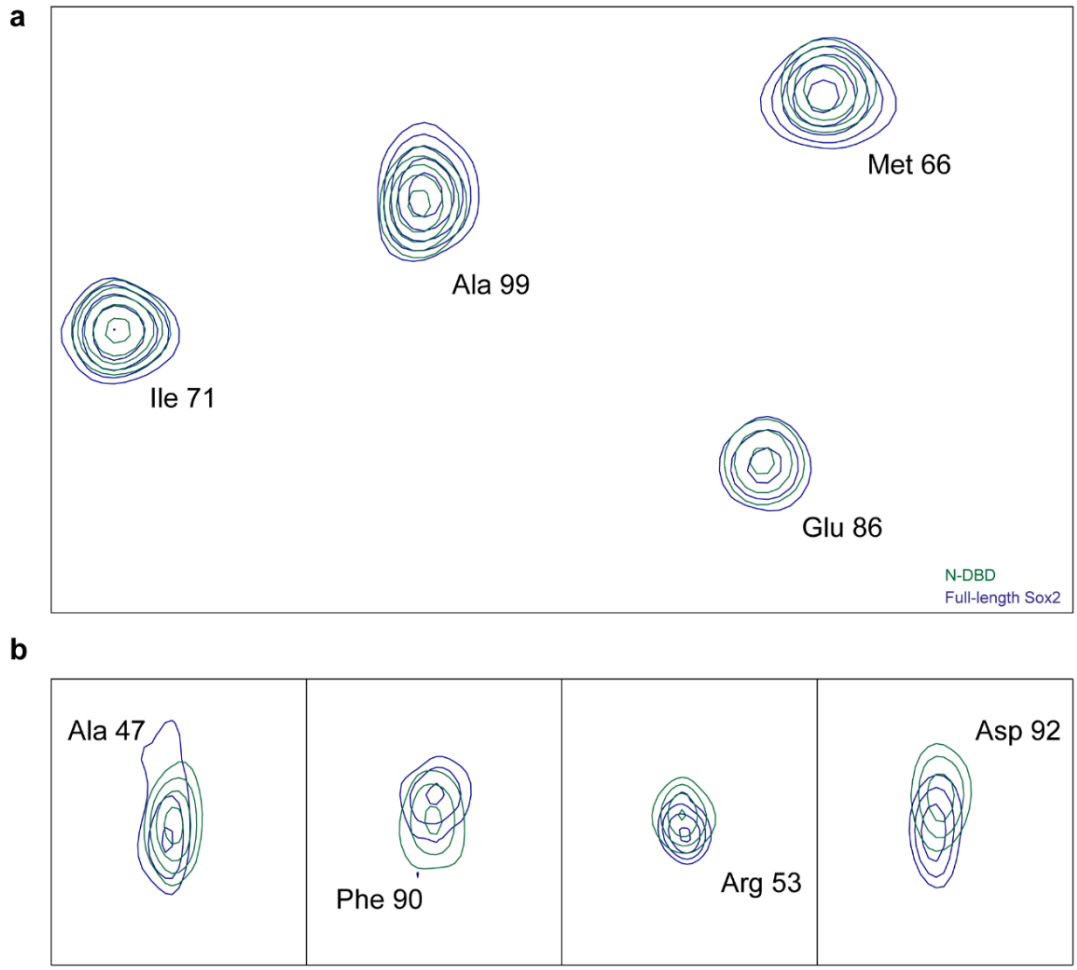
Contact lifetimes of interactions between the DBD and the AD1, ser-rich region or AD2 domains, for free and bound Sox2, were obtained from calculating the autocorrelation function of contact formation, with a contact between domains defined when the center of mass of two domains was within 1.0 nm. The interaction lifetimes were obtained by fitting the autocorrelation function using a double exponential

$$Aexp\left(-\frac{t}{\tau_1}\right) + Bexp\left(-\frac{t}{\tau_2}\right)$$

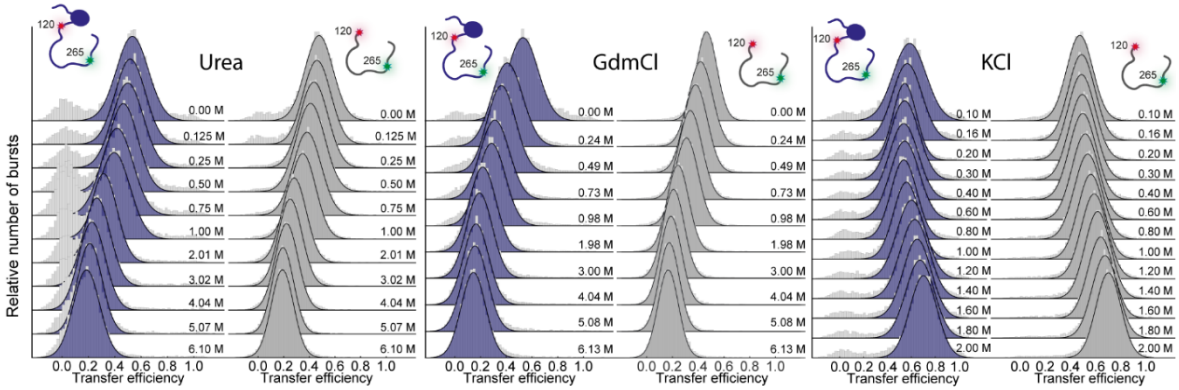
## 3.6 Supplementary information



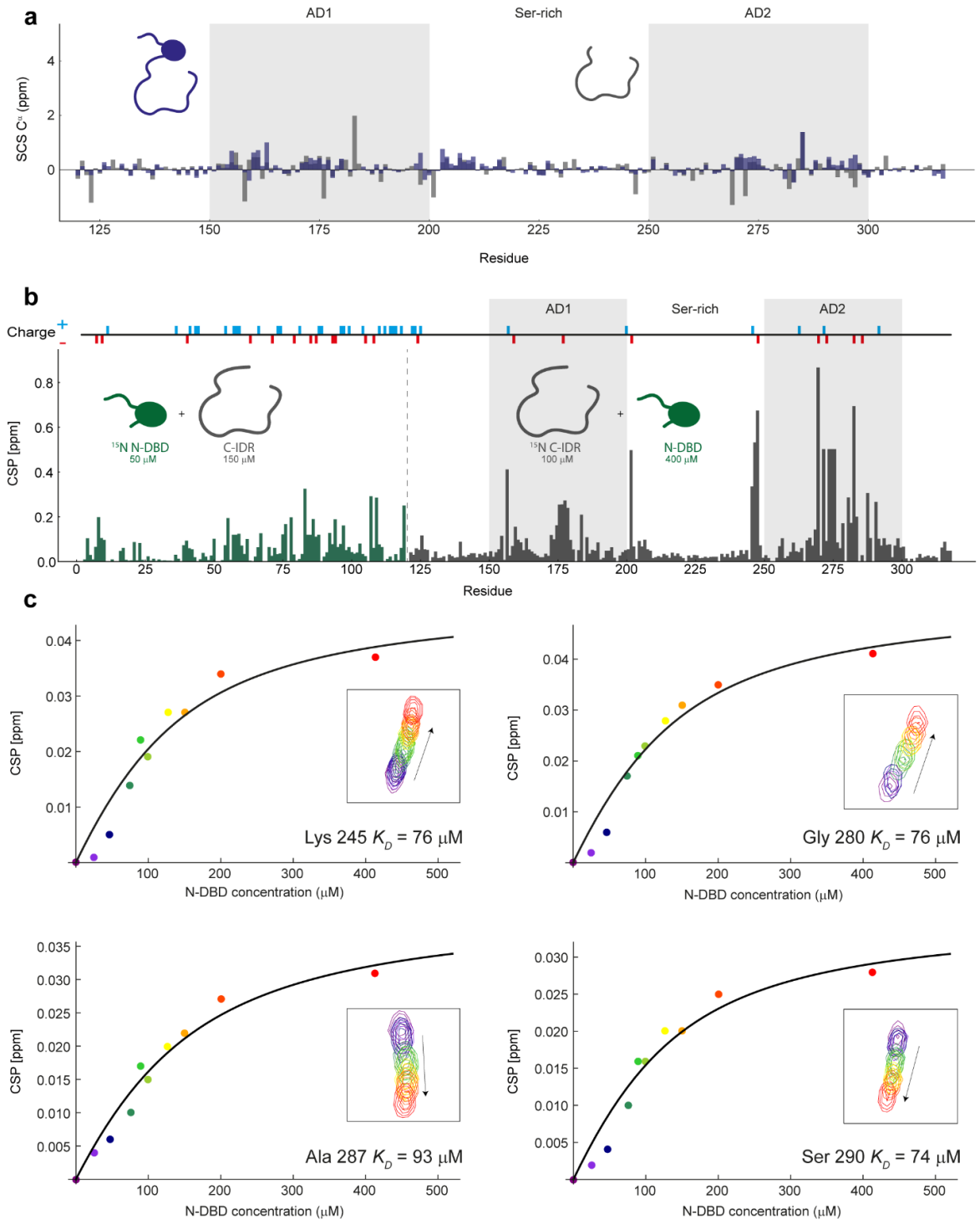

***Supplementary Figure 3.1. Stoichiometry measurements using pulsed-interleaved excitation.*** *2D-histograms of stoichiometry ratio vs. transfer efficiency from intramolecular FRET of Sox2 labelled in positions 37 and 120. The dotted lines in panel a indicate the range of stoichiometry ratios used for filtering out donor-only bursts. In this analysis, a burst originating from a molecule that has an active donor and acceptor results in a stoichiometry ratio of 0.5. In some cases, an additional residual population at a transfer efficiency close to zero but with stoichiometry ~1 can remain even after filtering (panel b) due to a large signal from molecules lacking an active acceptor dye.*




***Supplementary Figure 3.2. Peak positions and intensities of Sox2 DBD.*** *Examples of chemical shifts from $^{1}H^{15}N$ BTROSY[450] of full-length Sox2 (blue) overlapped with $^{1}H^{15}N$ BTROSY of the N-DBD (green). Most peaks from the DBD in the full-length protein generally overlap well with peaks from an isolated DBD (panel a) but many show small but specific chemical shift perturbations (panel b). This indicates that the DBD fold is generally unperturbed by the presence of the C-IDR. Both spectra were run with 640 scans at 15°C.*

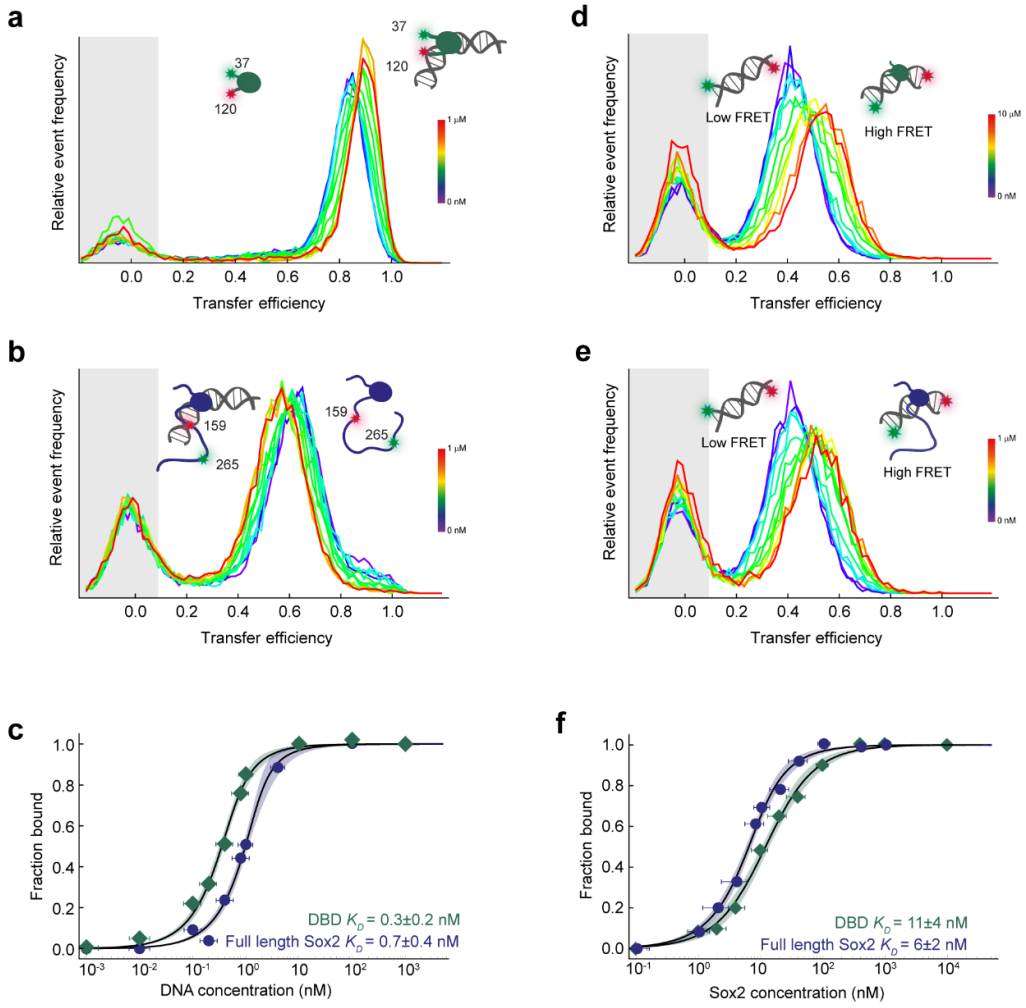



***Supplementary Figure 3.3. Dimensions of Sox2 C-IDR change in denaturants and salt.*** *Transfer efficiency histograms of full-length Sox2 (blue) or isolated C-IDR (grey) fluorescently labelled in positions 120-265, in different concentrations of urea (left), GdmCl (middle), and KCl (right).*

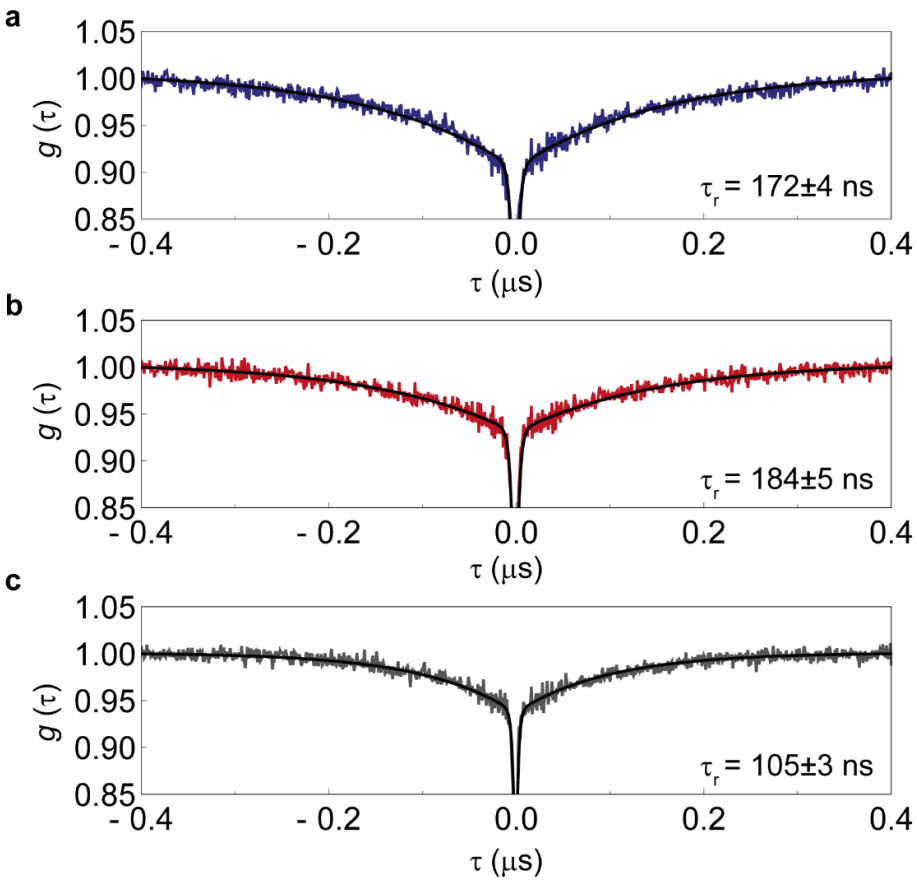



***Supplementary Figure 3.4. NMR data for isolated domains demonstrate interdomain interactions. a)*** *$C^{\alpha}$ SCS plot for residues 120-317 for full-length Sox2 and the isolated C-IDR. Secondary structure content is very similar in the two constructs and indicates general lack of structure. The main domains are indicated.* ***b)*** *CSP plot of $^{15}$N-labelled individual domains mixed with their unlabelled counterpart domain. The left side of the plot contains the combined $^{1}$H,$^{15}$N CSPs (see Methods) for the isolated $^{15}$N-labelled N-DBD (50 μM) with unlabelled C-IDR (150 μM). The right side of the plot contains the CSPs for the isolated $^{15}$N-labelled C-IDR (100 μM) with unlabelled N-DBD (400 μM).* ***c)*** *Binding isotherms using*

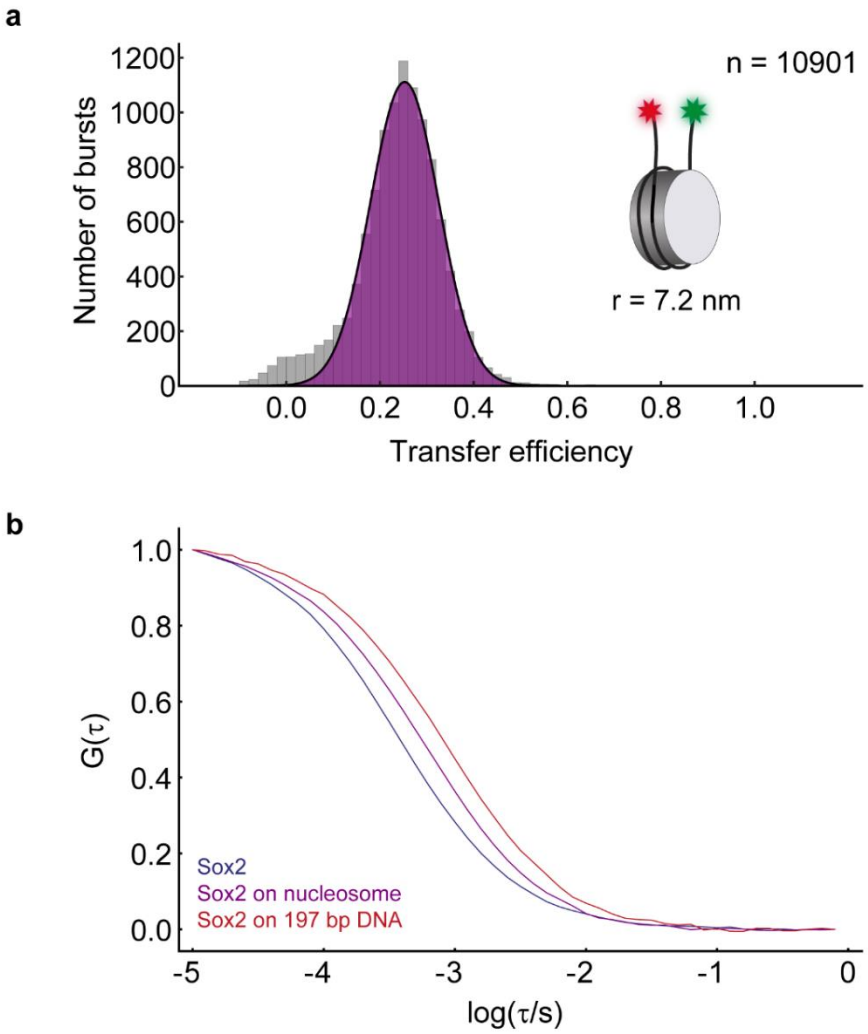
*chemical shifts of selected residues as a function of unlabelled N-DBD concentration. The inset panels show the corresponding resonance peaks with the colors matching the specific concentration point in the binding isotherm. Using these four residues and fitting to the simplest binding event of a 1:1 interaction, we determined an average dissociation constant $K_D$= 80 ± 4 μM.*



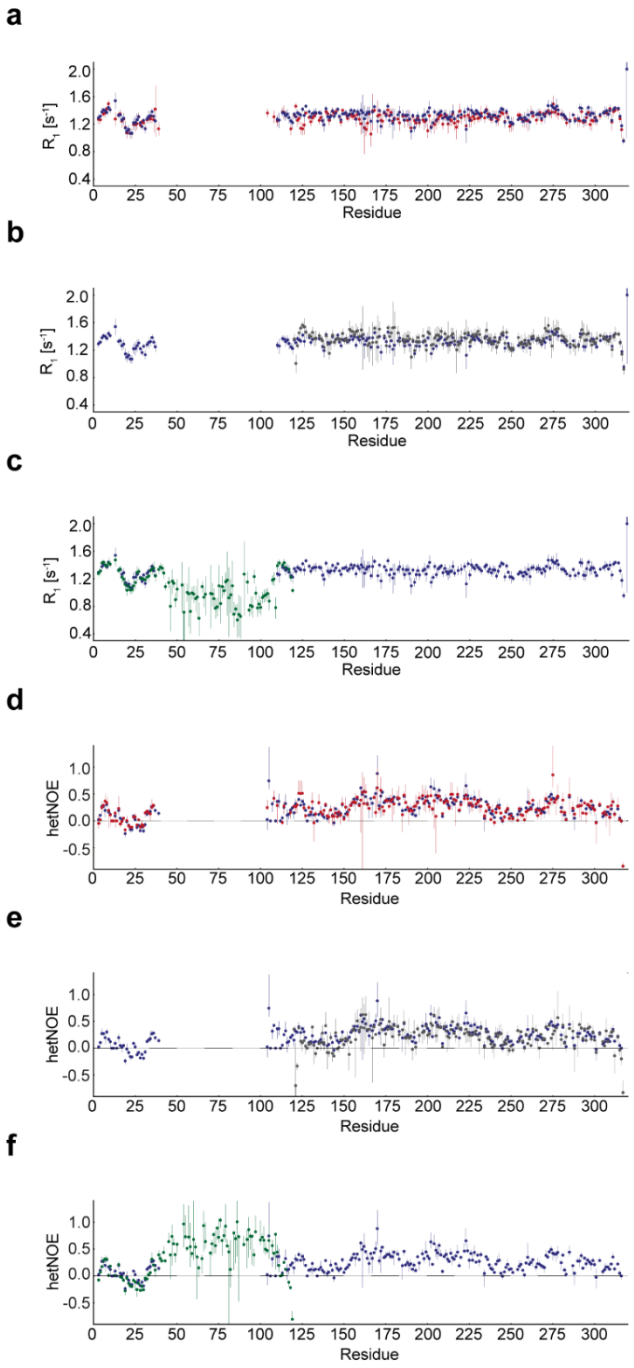

***Supplementary Figure 3.5. Binding affinity of Sox2 to specific and non-specific DNA. a-b)*** *Single-molecule transfer efficiency histograms of the a) isolated Sox2 DBD fluorescently labelled in positions 37 and 120 or b) full-length Sox2 fluorescently labelled in positions 159-265, with varying concentrations of unlabelled 30 bp specific DNA.* ***c)*** *The corresponding binding isotherms for panels a) and b).* ***d-e)*** *Single-molecule transfer efficiency histograms of 15 bp non-specific DNA fluorescently labelled at the 5´ and 3´-ends (see Supplementary Table 3) with varying concentrations of d) unlabelled isolated DBD or e) unlabelled full-length Sox2.* ***f)*** *The corresponding binding isotherms for panels d) and e). The dissociation constant for specific DNA using labelled proteins is nearly identical to the one determined with labelled DNA, excluding adverse effects from the fluorophores. Grey boxes in transfer efficiency histograms indicate donor-only populations, shaded areas in binding isotherms represent 95% confidence intervals of the fits, and error bars are from propagated dilution errors. All measurements were performed with 200 mM KCl.*



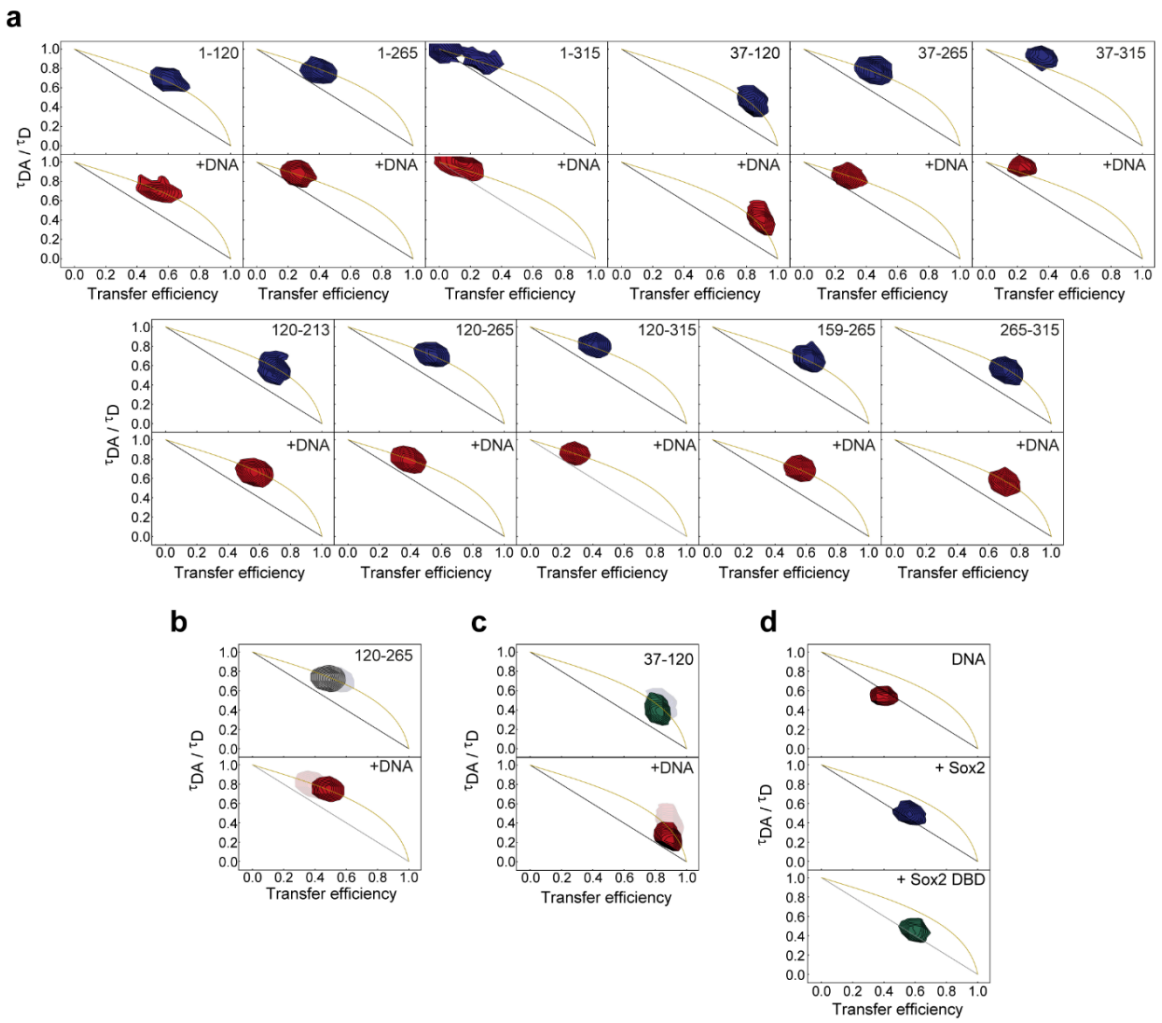

***Supplementary Figure 3.6. Rapid reconfiguration dynamics of the Sox2 C-IDR.*** *Nanosecond fluorescence correlation spectroscopy (nsFCS) of full-length Sox2 labelled in positions 120 and 315, and the isolated C-IDR labelled in positions 120 and 265. Fits of the donor-acceptor cross-correlation decays show sub-microsecond fluorescence intensity relaxation times for* ***a)*** *free Sox2 which persist in* ***b)*** *DNA-bound Sox2, and* ***c)*** *the isolated C-IDR. Errors are standard errors from the fits.*



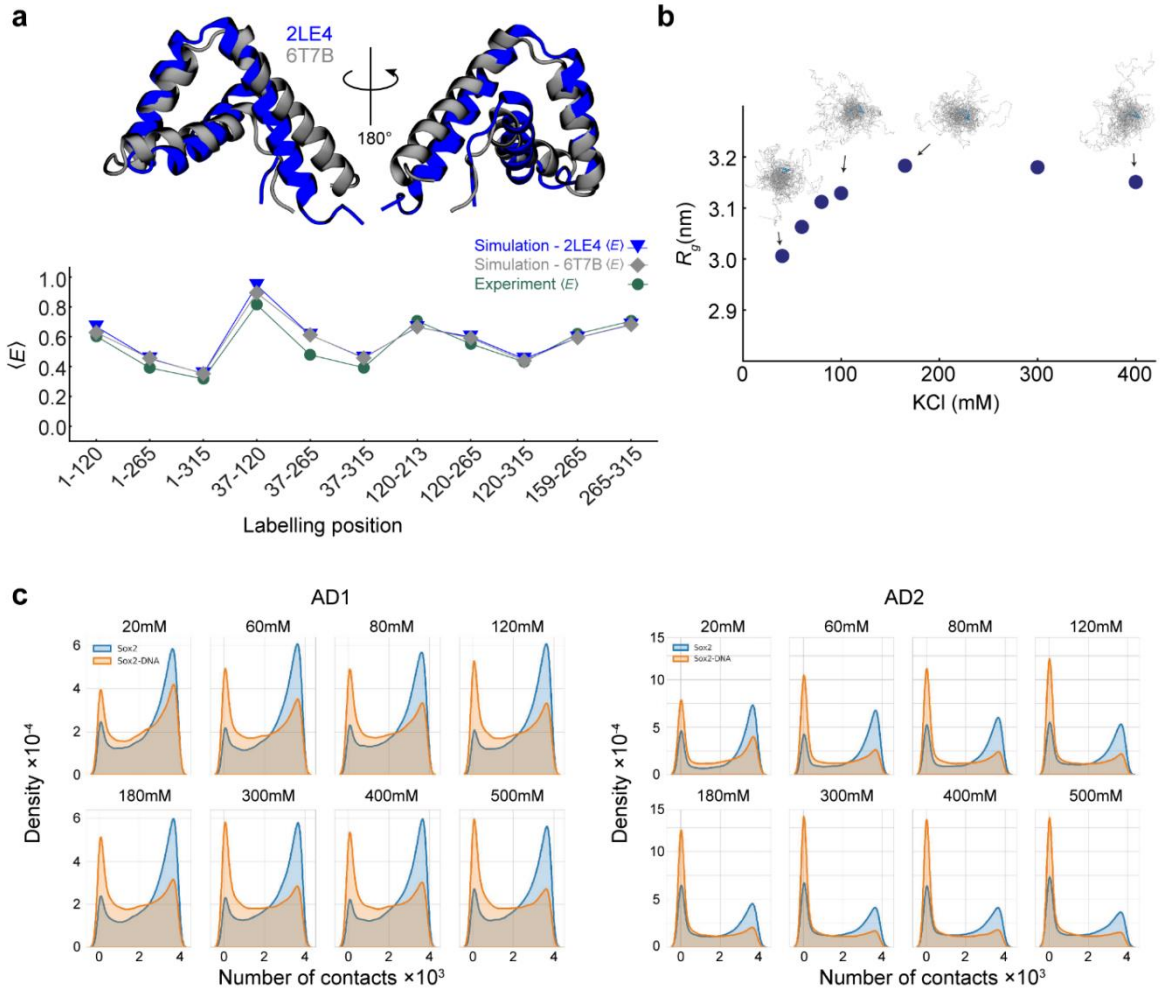

***Supplementary Figure 3.7. Single-molecule spectroscopy analysis of nucleosomes. a)*** *Single-molecule transfer efficiency histogram of 197 bp Widom 601 nucleosome fluorescently labelled at the DNA linker ends. Even at 100 pM concentrations, the nucleosome is stably wrapped as evident from the significant FRET between the fluorophores on each linker, in agreement with previous results*[420]*. The distance between the dyes is indicated, as calculated from the Förster equation.* ***b)*** *Donor-acceptor cross-correlation of fluorescently labelled Sox2, labelled in positions 120 and 315. Free Sox2 (blue) has a relatively short diffusion time through the confocal volume. In the presence of 90 nM of the same nucleosome as in panel a (unlabelled) the diffusion time is considerably increased (purple). In the presence of 197 bp 601 Widom DNA (red, without histone octamer), the diffusion time is increased even more, due to the much larger hydrodynamic radius of the DNA.*



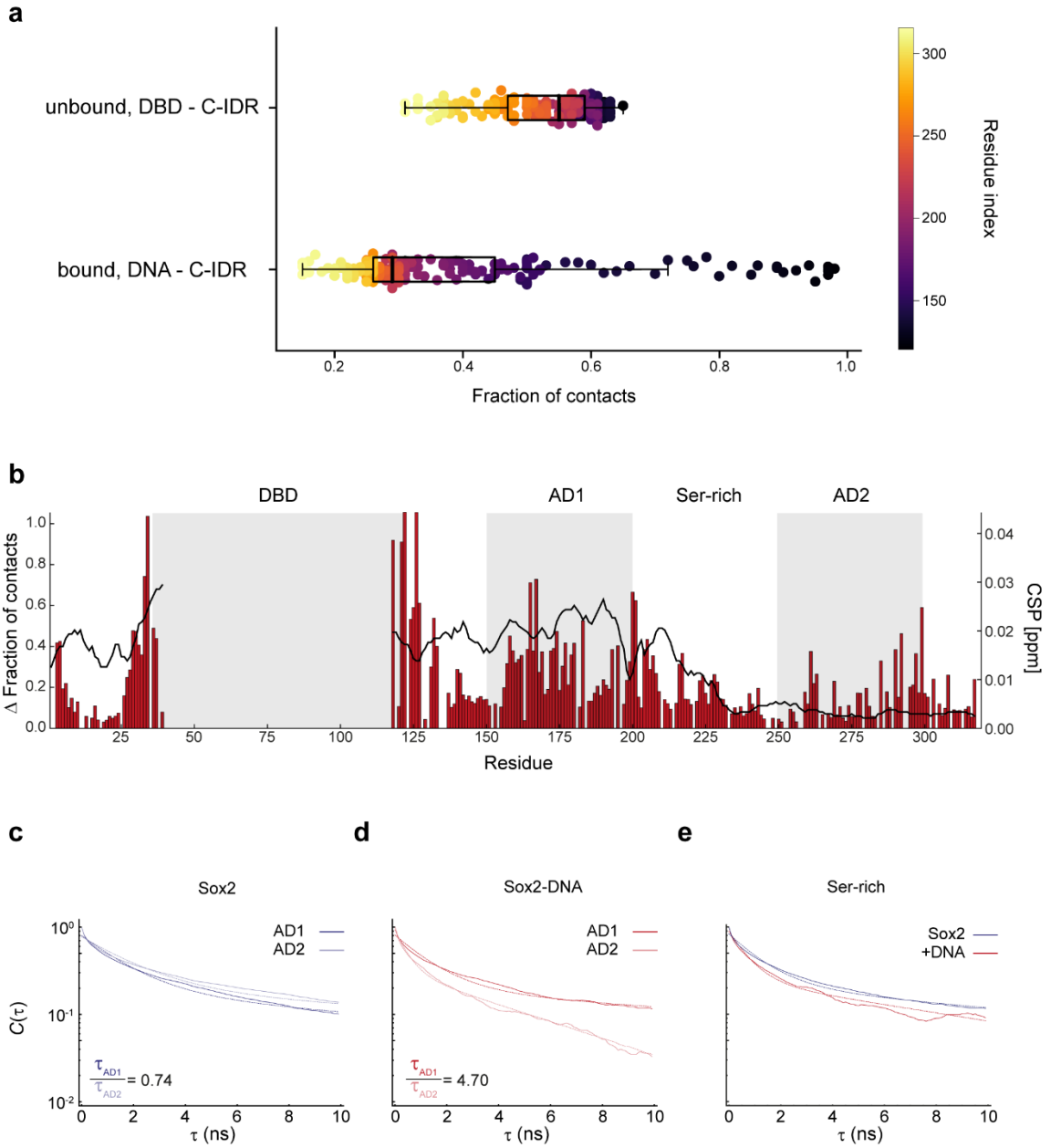

***Supplementary Figure 3.8. Relaxation data for full-length Sox2 and the isolated C-IDR, free and DNA-bound. a-c)*** *$R_1$ relaxation rates, and* ***d-f)*** *heteronuclear Overhauser effects (hetNOEs), for full-length Sox2 (blue), full-length Sox2 in complex with DNA (red), the isolated C-IDR (grey), and the isolated DBD (green). Error bars represent 95% confidence intervals.*



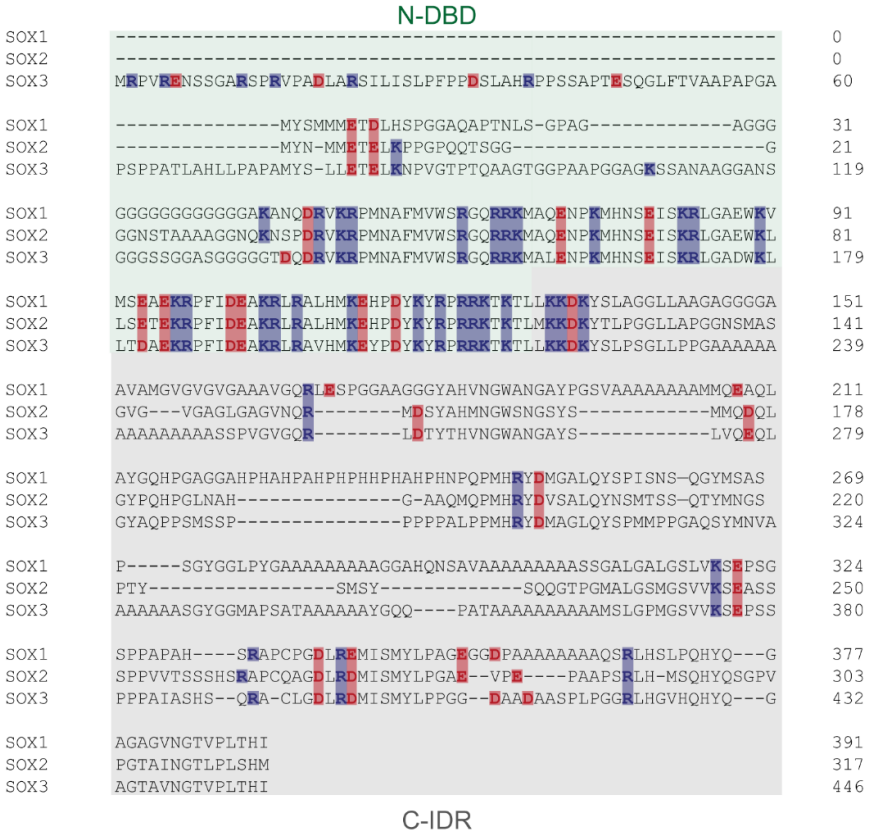

***Supplementary Figure 3.9. Fluorescence lifetime analysis of fluorescently labelled Sox2 variants and DNA. a)*** *Fluorescence lifetime analysis of all fluorescently labelled Sox2 variants, in the absence and presence of specific DNA.* ***b-c)*** *Fluorescence lifetime analysis of the isolated C-IDR (b) and DBD (c), in the absence and presence of DNA. The faint peaks represent the lifetimes of the full-length protein labelled in the same positions for comparison.* ***d)*** *Fluorescence lifetime analysis of fluorescently labelled 15 bp DNA in the free state (top), bound to full-length Sox2 (middle), and to the isolated DBD (bottom). Black lines describes the dependence of a static distance from the Förster equation, the yellow lines describe the dependence for a SAW-n distance distribution*[431].




***Supplementary Figure 3.10. Simulations under different conditions. a)*** *Agreement between FRET efficiencies from experiments and simulations using an NMR structure of the DBD (PDB ID: 2LE4). Simulations were performed exactly as those in the main text (also shown here for PDB 6T7B). The $r_c$ values of 0.89 (2LE4) and 0.92 (6T7B) show that the effects of DBD structure on the dimensions of the IDRs are minimal. An alignment of the two structures is shown (RMSD 4.4 Å).* ***b)*** *Radius of gyration, $R_g$, as a function of apparent salt concentration for the region encompassing residues 120-265 for free Sox2, calculated from the simulations. At low salt the chain is compact (low $R_g$) and then expands due to increased charge screening with increasing salt concentrations, in agreement with FRET data (Figure 3.2f). 20 overlaid snapshots from the simulations at different salt concentrations are shown.* ***c)*** *Probability distributions of the contacts between AD1 and AD2 with the DBD, for Sox2 in the free and DNA-bound states, at different salt concentrations.*




***Supplementary Figure 3.11. Simulation analysis of inter-residue contact changes upon DNA binding. a)*** *Fraction of contacts between the C-IDR and DBD for free Sox2, and for the C-IDR and DNA in the DNA-bound state, calculated from the simulations. The boxes represent the interquartile range and the thick lines within the boxes show the median values of the distributions.* ***b)*** *Plot showing the difference in fraction of contacts between the DBD and the IDRs (black line) from the simulations and the CSPs (red bars, from Figure 3.3h) from the NMR experiments, comparing free and DNA-bound Sox2.* ***c-d)*** *Contact relaxation times between the DBD and AD1 or AD2 in the absence (c) and presence (d) of DNA, calculated from simulations. Double-exponential fits (dotted lines) reveal that the contact relaxation times of both ADs are approximately equal in free Sox2 but diverge in complex with DNA, where contact relaxation times are more than four times as long for AD1 compared to AD2.* ***e)*** *Contact relaxation times of the Ser-rich domain, calculated from simulations. The contact relaxation times for the Ser-rich region are similar for free and DNA-bound Sox2.*




***Supplementary Figure 3.12. Sequence alignment of the SoxB family of TFs.*** *N-DBD and C-IDR regions are indicated with green and gray background, respectively. Charges are highlighted for clarity. Alignment generated by CLUSTAL O[451]. UniProt IDs: O00570 (Sox1), P48431 (Sox2) and P41225 (Sox3).*

***Supplementary Table 0.1.*** *Sox2 variants that were used in this study. The residues used for fluorophore labelling (after substituting for cysteine) are marked in red.*

| | |
|---|---|
| Full-length Sox2 | MYNMMETELKPPGPQQTSGGGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQRRKMAQENP<br>KMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKTKTLMKKDKYTLPGGLLAPG<br>GNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQDQLGYPQHPGLNAHGAAQMQP<br>MHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMALGSMGSVVKSEASSSPPVVTSSSHSRAP<br>CQAGDLRDMISMYLPGAEVPEPAAPSRLHMSQHYQSGPVPGTAINGTLPLSHM |
| C-IDR | MKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQDQLGYPQHP<br>GLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMALGSMGSVVKSEASSS<br>PPVVTSSSHSRAPCQAGDLRDMISMYLPGAEVPEPAAPSRLHMSQHYQSGPVPGTAINGTLPLSHM |
| N-DBD | MYNMMETELKPPGPQQTSGGGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQRRKMAQEN<br>PKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKTKTL |
| DBD | KNSPDRVKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALH<br>MKEHPDYKYRPRRKTKTLMKK |

***Supplementary Table 0.2.*** *Measured FRET efficiencies and fluorescence lifetimes of all protein variants and DNA. The average donor and acceptor lifetimes are 2.72±0.11 ns and 3.01±0.04 ns, respectively.*

| Sox2 variant | FRET efficiency (E) | Donor lifetime (ns) | Acceptor lifetime (ns) | FRET lifetime (ns) |
|---|---|---|---|---|
| 1-120 | 0.60 | 2.80 | 2.95 | 2.92 |
| 1-120 + DNA | 0.51 | 2.79 | 2.94 | 2.97 |
| 1-265 | 0.39 | 2.77 | 2.98 | 2.95 |
| 1-265 + DNA | 0.25 | 2.71 | 2.95 | 2.99 |
| 1-315 | 0.32 | 2.73 | 3.05 | 3.04 |
| 1-315 + DNA | 0.19 | 2.72 | 3.03 | 2.94 |
| 37-120 | 0.81 | 2.75 | 3.04 | 3.09 |
| 37-120 + DNA | 0.88 | 2.72 | 3.01 | 3.03 |
| 37-265 | 0.48 | 2.71 | 3.04 | 3.07 |
| 37-265 + DNA | 0.28 | 2.70 | 3.03 | 3.07 |
| 37-315 | 0.39 | 2.51 | 3.05 | 3.05 |
| 37-315 + DNA | 0.22 | 2.57 | 3.05 | 3.05 |
| 120-213 | 0.71 | 2.78 | 3.02 | 2.98 |
| 120-213 + DNA | 0.56 | 2.82 | 3.01 | 3.06 |
| 120-265 | 0.55 | 2.72 | 3.03 | 3.05 |
| 120-265 + DNA | 0.37 | 2.72 | 3.01 | 3.04 |
| 120-315 | 0.43 | 2.62 | 3.04 | 3.03 |
| 120-315 + DNA | 0.28 | 2.72 | 3.03 | 3.06 |
| 159-265 | 0.62 | 2.62 | 2.99 | 3.07 |
| 159-265 + DNA | 0.55 | 2.74 | 2.98 | 3.00 |
| 265-315 | 0.70 | 2.73 | 3.02 | 3.04 |
| 265-315 + DNA | 0.69 | 2.76 | 3.03 | 3.04 |
| C-IDR 120-265 | 0.48 | 2.62 | 3.01 | 3.08 |
| C-IDR 120-265 + DNA | 0.48 | 2.61 | 3.00 | 3.00 |
| DBD 37-120 | 0.81 | 2.46 | 3.00 | 2.98 |
| DBD 37-120 + DNA | 0.88 | 3.01 | 3.00 | 3.04 |
| Labelled 15 bp DNA | 0.40 | 2.89 | 3.11 | 3.11 |
| Labelled 15 bp DNA + Sox2 | 0.59 | 2.81 | 3.03 | 3.04 |
| Labelled 15 bp DNA + Sox2 DBD | 0.59 | 2.78 | 2.94 | 3.05 |

***Supplementary Table 0.3.*** *DNA constructs for binding experiments. The Sox2 binding site in the 601-Widom sequence is marked in red. 5AmMC6 indicates a C6-amino group for labelling with an NHS-ester fluorophore.*

| | |
|---|---|
| 15 bp specific, (+) strand | 5'-/5AmMC6/ACT CTT TGT TTG GAT-3' |
| 15 bp specific, (-) strand | 3'-TGA GAA ACA AAC CTA/5AmMC6/-5' |
| 15 bp nonspecific, (+) strand | 5'-/5AmMC6/ TCT ATC TGT GTA TGT-3' |
| 15 bp nonspecific, (-) strand | 3'- AGA TAG ACA CAT ACA/5AmMC6/-5' |
| 30 bp specific, (+) strand | 5'-ATC CCA TTA GCA TCC AAA CAA AGA GTT TTC-3' |
| 30 bp specific, (-) strand | 3'-TAG GGT AAT CGT AGG TTT GTT TCT CAA AAG-5' |
| 21 bp specific, (+) strand | 5'-GAA TAC TCT TTG TTT GGA TGC-3' |
| 21 bp specific, (-) strand | 3'-CTT ATG AGA AAC AAA CCT ACG -5' |
| '601'-Widom sequence with Sox2 binding site at SHL +6, forward strand | 5'-TCCATGGACCCTATACGCGGCCGCCCTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCG TAGACAGCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAG GGGATTACTCCCTAGTCTCCAGGCCTTTGTTATGCAAATACATCCTGTGCATGTATTGAACA GCAGTATGCCT-3' |
| '601'-Widom sequence with Sox2 binding site at SHL +6, reverse strand | 3'-AGGCATACTGCTGTTCAATACATGCACAGGATGTATTTGCATAACAAAGGCCTGGAGACTAG GGAGTAATCCCCTTGGCGGTTAAAACGCGGGGGACAGCGCGTACGTGCGTTTAAGCGGTGC TAGAGCTGTCTACGACCAATTGAGCGGCCTCGGCACCGGGATTCTCCAGGGCGGCCGCGTAT AGGGTCCATGGA-5' |
| '601'-Widom primer, forward | 5'-/5AmMC6/TCCATGGACCCTATACGCGGCCGCC-3' |
| '601'-Widom primer, reverse | 3'-AGGCATACTGCTGTTCAATACATGCACAGGATGTATTTGCATAACAAAGGCCTGGAGAC/5AmMC6/-5' |

***Supplementary Table 0.4.*** *Binding affinities of Sox2 for DNA from smFRET experiments, measured at 200 mM salt concentration. Errors are standard errors based on propagating pipetting errors.*

| **Binding reaction** | $K_D$ **(nM)** |
|---|---|
| Labelled specific DNA – Full-length Sox2 | 0.4 ± 0.2 |
| Labelled specific DNA – DBD | 0.3 ± 0.1 |
| Labelled Full-length Sox2 – 30 bp specific DNA | 0.7 ± 0.4 |
| Labelled DBD – 30 bp specific DNA | 0.3 ± 0.2 |
| Labelled non-specific DNA – Full-length Sox2 | 6 ± 2 |
| Labelled non-specific DNA – DBD | 11 ± 4 |

***Supplementary Table 0.5.*** *Contact relaxation times obtained from simulations for AD1, AD2, and Ser-rich domain, based on exponential fitting. Errors are the standard errors of the fit.*

| | τ1 (ns) | τ2 (ns) |
|---|---|---|
| AD1 | 1.58 ± 0.03 | 18.12 ± 0.33 |
| AD1 + DNA | 1.32 ± 0.02 | 21.09 ± 0.27 |
| AD2 | 1.95 ± 0.03 | 24.34 ± 0.38 |
| AD2 + DNA | 0.48 ± 0.01 | 4.49 ± 0.06 |
| Ser-rich | 1.20 ± 0.02 | 16.09 ± 0.21 |
| Ser-rich + DNA | 0.71 ± 0.01 | 19.17 ± 0.122 |

***Supplementary Movie 3.1.*** *Simulation of free Sox2.*

The free Sox2 ensemble, with the Sox2 DBD shown in blue and the disordered N-IDR and C-IDR in silver. The movie shows 400 conformations of the Sox2 ensemble, collected from the 16 μs Langevin dynamics simulations (see Methods).

***Supplementary Movie 3.2.*** *Simulation of Sox2 in complex with DNA.*

Ensemble of Sox2 bound to DNA, with the Sox2 DBD shown in blue, the disordered N-IDR and C-IDR in silver, and the DNA in dark grey. The movie shows 400 conformations of the Sox2 ensemble in complex with DNA, collected from the 16 μs Langevin dynamics simulations (see Methods).

# 4 Chapter 4

## Pioneer factor Sox2 remodels nucleosomes and displaces histone H1

Sveinn Bjarnason, Jens Nicolai Vilstrup Decker, Eliška Koutná, Kinga S. Demény, Vaclav Veverka, Pétur O. Heidarsson

# About this chapter

This chapter is a manuscript which is intended for article publication after additional experiments are conducted. It focuses on the dynamic interactions between Sox2 and nucleosomes, including the nucleosome remodeling facilitated by Sox2.

# Author contribution

S.B., J.N.V.D., E.K., V. V., and P.O.H. designed the study; S.B., K.S.D. and J.N.V.D. prepared protein and DNA constructs; S.B., K.S.D. and J.N.V.D. performed all single-molecule experiments; E.K. performed NMR experiments; S.B., J.N.V.D., K.S.D., E.K., V. V., and P.O.H. analysed data; S.B. and P.O.H. wrote the manuscript.

# Keywords

Intrinsically Disordered Proteins; Intrinsically Disordered Regions; Pioneer Transcription Factors; Sox2; Nucleosomes; Histone H1.0; *LIN28B*; Widom-601; smFRET and NMR.

## 4.1 Abstract

Through their interaction with chromatin, pioneer transcription factors not only initiate gene expression changes essential for development but also serve as key regulators in cellular reprogramming and disease processes. The mechanism by which the pioneer transcription

factor Sox2 remodels nucleosomes and its capacity to displace histone H1 remains poorly understood. Similarly, the structural configuration of Sox2's intrinsically disordered regions (IDRs) when engaged with nucleosomes, and the subsequent effects on nucleosome conformation, are not well defined. To address these knowledge gaps, we employed single molecule Förster Resonance Energy Transfer (smFRET) and Nuclear Magnetic Resonance (NMR) spectroscopy to elucidate the conformational properties of the Sox2 nucleosome complex. We explored the binding modes of Sox2, the location dependent affinity of binding sites and the resulting structural impacts on the nucleosome itself. Our findings reveal that Sox2's IDR extends significantly upon interaction with the nucleosome, and we identify a specific site within the IDR that interacts with the core histones. Interestingly, Sox2's binding affinity remains unaffected, irrespective of the location of the binding site on the nucleosomes, even as the nucleosomes are partially opened upon Sox2 binding. Moreover, our data not only confirms Sox2's capability to displace linker histone H1 while retaining a high affinity, highlighting a critical mechanism in chromatin remodeling, but also positions this activity within the wider context of pioneer transcription factors strategies for chromatin engagement. Identifying an interaction site within Sox2's IDRs marks a step forward in our grasp of pioneer transcription factor mechanism, offering promising leads for development strategies aimed at fine-tuning Sox2's function in cellular reprogramming and oncogenesis.

## 4.2 Introduction

Pioneer transcription factors (pTFs) play a vital role in the regulation of gene expression, possessing the unique ability to recognize specific DNA sequences within nucleosomes, thereby initiating chromatin remodeling that renders the DNA accessible to other transcription factors[52]. The chromatin opening capability is essential for orchestrating complex gene expression programs during development, differentiation, and cellular reprogramming[50,452]. Recent advances have shed light on the many roles of pTFs in development[453], cell reprogramming[454] and cancer progression[455]. However, the molecular intricacies of how pTFs remodel chromatin, particularly through their interactions with nucleosomes and displacement of histone H1[456], remain elusive. While most pTFs are characterized by their structured DNA binding domains (DBDs), which have been extensively studied and modeled[457-459], the true complexity of these proteins lies in their long IDRs. Despite the critical role of IDRs in the functionality of pTFs[460-462], detailed structural insights into these regions, both in their free state and when interacting with nucleosomes, remain scarce. This lack of information on IDR ensembles poses a significant challenge, as these regions are essential for the diverse function of pTFs, including Sox2.

The pioneering activity of Sox2 has been instrumental in the generation of induced pluripotent stem cells (iPSCs)[463,464], pivotal for a new era in regenerative medicine[465,466] and pharmaceutical development[466]. Beyond its critical function in cellular reprogramming, Sox2 also plays a complex role in cancer development[74,77]. Sox2 is composed of 317 amino acids, featuring an HMG-box DBD, flanked by IDRs N-terminally (40 residues) and C-terminally (200 residues)[467] (**Figure 4.1a**). Although it is established that Sox2's IDRs are crucial for its reprogramming function[374], the underlying mechanisms remain elusive. The IDRs could be responsible for interactions with other transcription activators or direct involvement in chromatin binding and opening.

Part of the challenge in elucidating these mechanisms stems from the limitations of traditional structural biology techniques in studying the complex nature of pTFs. Addressing these gaps, our study aims to dissect the complex interactions between the pTF Sox2 and nucleosomes, focusing particularly on understanding how Sox2's IDRs contribute to its ability to bind to and remodel nucleosomal DNA[53,366]. Specifically, we seek to characterize the dynamic conformations adopted by Sox2 when in complex with nucleosomes, using state-of-the-art smFRET and NMR spectroscopy. Additionally, we assess the impact of Sox2 binding on the structure of nucleosomes, particularly those bound by linker histone H1, thereby clarifying the role of Sox2 in chromatin accessibility and gene regulation.

We show that the C-terminal IDR (C-IDR) of Sox2 interacts with the core histones, identifying the specific residues in the C-IDR involved in this interaction. Next, we determined Sox2's binding affinity to Widom-601 nucleosomes using several Sox2 binding site locations. Finally, we demonstrated Sox2's ability to bind and remodel H1-bound nucleosomes, employing both the Widom-601 sequence and the native-like *LIN28B* sequence (**Figure 4.1b,c**). These insights advance our understanding of the interplay between pTFs like Sox2 and the chromatin landscape. By delineating the specific mechanisms through which Sox2 interacts with and influences nucleosome structure, this work opens new pathways for exploring therapeutic interventions in conditions where Sox2's regulatory functions are altered.

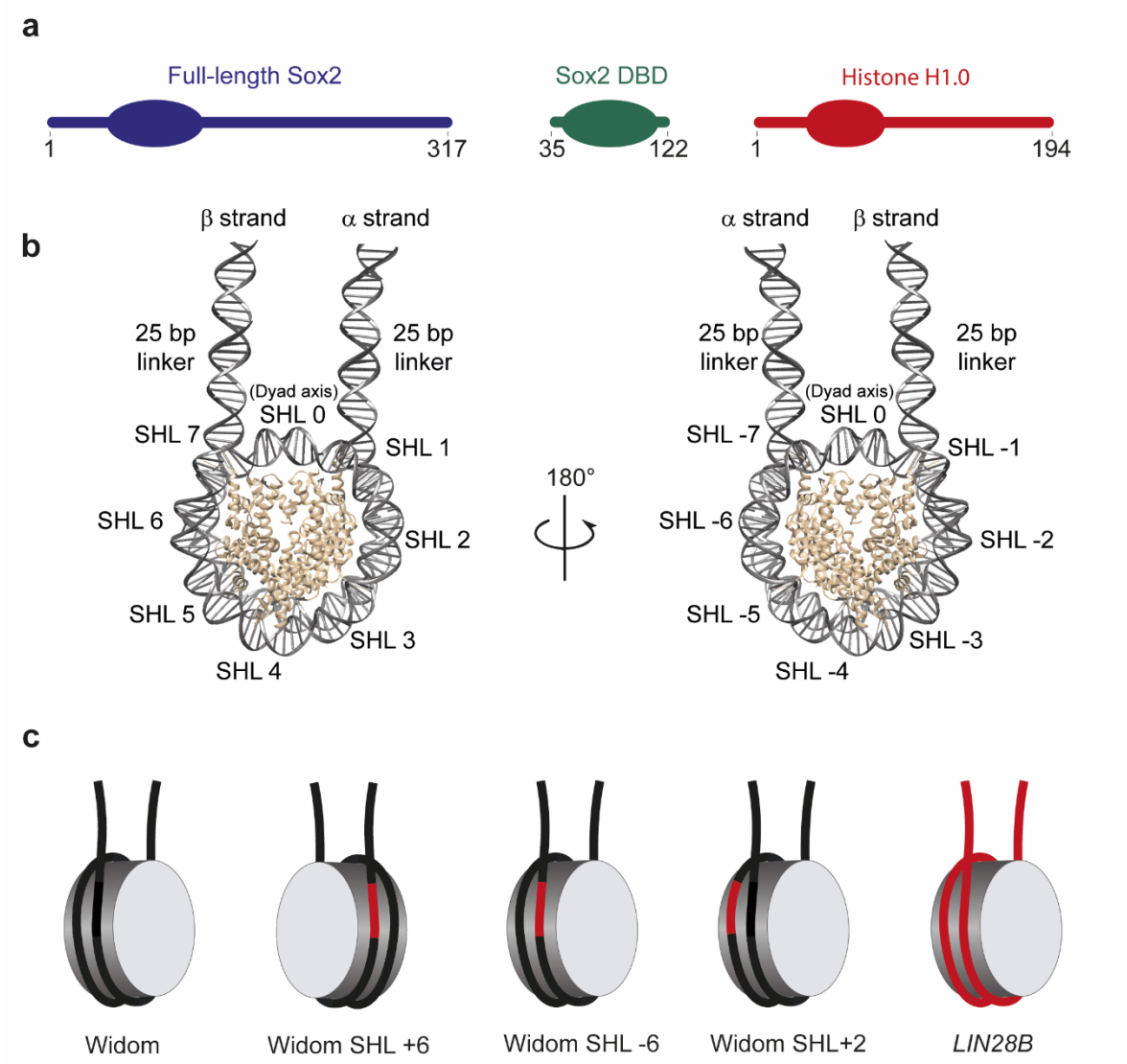



***Figure 4.1. Schematic of Sox2 and different nucleosome variants. a)*** *Schematics of Sox2 constructs and histone H1.0.* ***b)*** *Schematics of a 197 bp nucleosome, with SHL labelled (modified from PDB:5NL0, for illustrative purposes only). Each SHL site indicates where the major groove of the DNA faces the histone surface.* ***c)*** *Schematics of the nucleosomes used in this study, with red indicating the location of an Sox2 binding site. LIN28B is colored red as it contains several Sox2 binding sites.*

## 4.3 Results

### 4.3.1 Sox2 binds nucleosomes with high affinity independent of binding site location

Recently, structures of Sox2 DBD bound to nucleosomes revealed, at least in part, Sox2 remodels nucleosomes[468,469]. However, the impact of Sox2's IDRs on its binding capabilities is currently unclear. Furthermore, previous work has not been clear on the binding site location preference of Sox2, with some studies reporting a link between binding site and

affinity and others showing binding site independence from affinity[468,470]. To answer these questions, we reconstituted fluorescently labelled nucleosomes using the strongly positioning Widom-601 sequence[471] with an incorporated Sox2 binding site at superhelical locations (SHL) -6, +6 and +2 (**Figure 4.1c**). The fluorophores were placed in position α11/β11 at the ends of the nucleosomal linkers (**Figure 4.2**). The free nucleosomes had a FRET value, E, of 0.56 (corresponding to an average interdye distance of 5.7 nm, see methods for details), indicating that the Sox2 recognition sequence does not interfere with stable nucleosome structure (**Supplementary Figure 4.2**). Adding increasing concentrations of unlabelled Sox2 (up to 1 μM) led to a change in transfer efficiency from 0.56 to 0.41 (6.4 nm) for the α11/β11 nucleosomes. As was observed with the recent Cryo-EM structures[468,469], Sox2 remodels the nucleosomal structure, by inducing a bend in the DNA on its binding site that peels that section away from the core histone octamer. We then used the areas of the resulting FRET histograms to determine the fraction of bound nucleosomes as a function of Sox2 concentration, using the area corresponding to unbound nucleosomes as a baseline for a 0.0 fraction bound. This analysis allowed us to estimate the equilibrium dissociation constant, $K_D$.

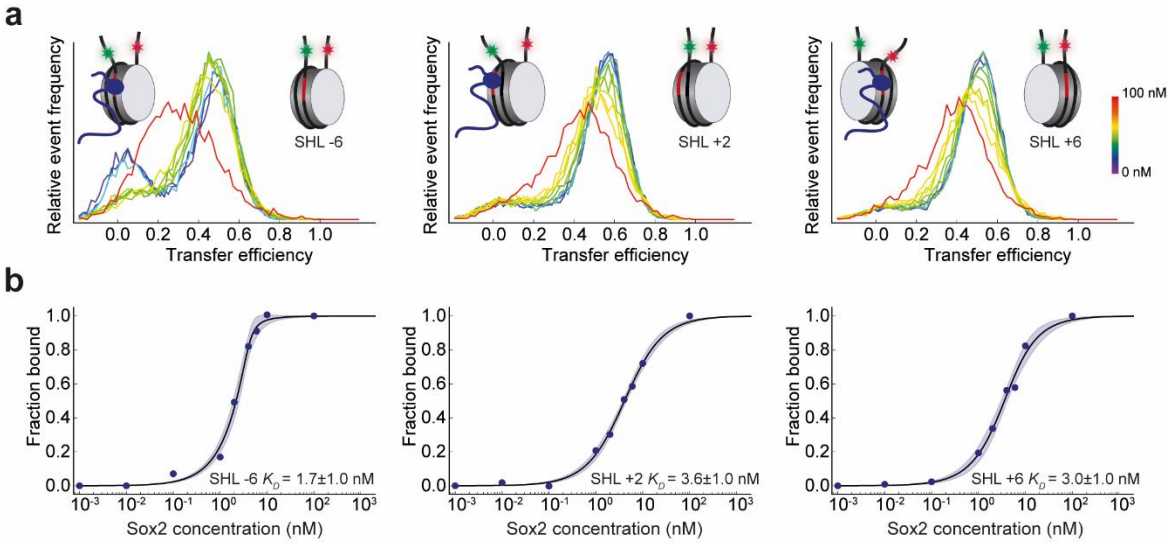



***Figure 4.2. Binding isotherms using fluorescently labelled nucleosomes. a)*** *Single-molecule transfer efficiency histograms of fluorescently labelled nucleosomes, with Sox2 binding sites in SHL-6 (left), SHL+2 (middle) and SHL+6 (right).* ***b)*** *The corresponding binding isotherms with fits to a 1:1 binding model.*

While measuring binding affinity of Sox2 to nucleosomes, an interesting observation was made. The data indicates that the high binding affinity (average of 2.8 nM, 200 mM ionic strength) of Sox2 to its target site remains consistent regardless of the location of these sites within the nucleosome. Furthermore, measurements made using the isolated DBD of Sox2 (**Supplementary Figure 4.1**) suggests that the binding affinity remains largely the same in the absence of Sox2's IDRs (2.2±0.8 nM). This suggests that the interaction of Sox2's DBD to nucleosomal DNA is inherently robust and that Sox2's IDRs do not have a large effect on binding affinity for mononucleosomes.

### 4.3.2 Direct interaction between Sox2 C-IDR and nucleosomes

Our findings so far seem to indicate that there would be no interaction between Sox2's IDRs and the nucleosomes. But by focusing solely on binding affinities, weaker interactions could be overlooked. Indeed, IDRs often take part in weak, multivalent interactions[30,472]. Given the strong binding affinity of Sox2, it is highly probable that any such weak interactions would be missed. To observe potential changes to the C-IDR conformation when bound to nucleosomes, we used smFRET to measure transfer efficiency histograms for Sox2 fluorescently labelled in the C-IDR (positions 120 and 315) when bound to either nucleosomes or free nucleosomal DNA (SHL+2) (**Figure 4.3**). Interestingly, the histograms revealed a similar decrease in FRET when bound to DNA and nucleosomes.

To delve deeper into the dynamics of Sox2 on the nucleosome, we probed the rapid conformational fluctuations of the C-IDR by measuring relative fluorescence lifetimes, which allowed us to detect distance fluctuations between the two fluorophores on a timescale bridging the fluorescence lifetime (nanoseconds) and the interphoton time (microseconds). The dynamic nature of these regions appears to be considerably dampened, indicated by deviation from the dynamic line, when bound to nucleosomes compared to when interacting with the same DNA sequence in the free form (**Figure 4.3b**). This dampening effect suggests a specific interaction between the IDRs of Sox2 and the nucleosomal components, presumably the core histones. To further pinpoint the interaction area, we used Sox2 constructs fluorescently labelled in discrete regions across the C-IDR. Narrowing the distance between the labels, we created constructs labelled at positions 120-265, then 159-265, and finally 265-315 (**Figure 4.3c**). The construct labelled at positions 120-265 exhibited dampened dynamics, similar to the behaviour of the 120-315 construct. However, a significant dampening effect was observed in the 159-265 construct, suggesting it is in close proximity to the interaction site. In contrast, the 265-315 construct displayed no difference in dynamics whether bound to DNA or nucleosomes, indicating that the interaction sites are likely located between residues 159 and 265.

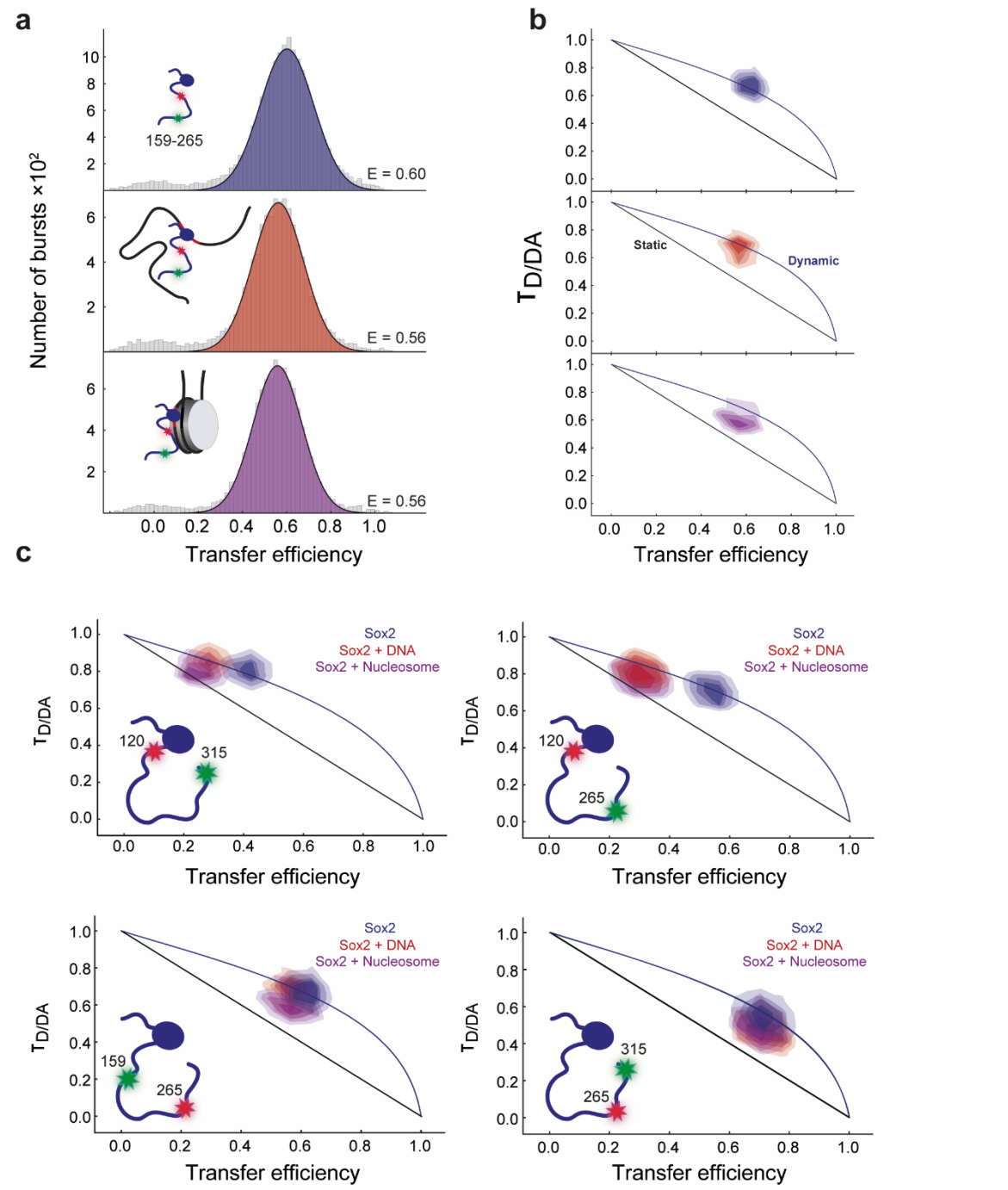



***Figure 4.3. Sox2 off and on SHL+2 DNA and nucleosomes. a)*** *Single-molecule transfer efficiency histograms of full-length Sox2 fluorescently labelled on residues 159 and 265, in the absence (blue) and presence of 197 bp DNA (red) or 197 bp nucleosomes (purple).* ***b)*** *Fluorescence lifetime analysis of Sox2 fluorescently labelled on residues 159 and 265 in the absence (blue) and presence of 197 bp DNA (red) or 197 bp nucleosomes (purple).* ***c)*** *Fluorescence lifetime analysis of Sox2 fluorescently labelled according to scheme in the absence (blue) and presence of 197 bp DNA (red) or 197 bp nucleosomes (purple).*

Having narrowed down the interaction site on Sox2 to the C-IDR region proximal to the DBD, we turned to NMR due to its unparalleled resolution on the residue level[473]. NMR allows us to probe which amino acids within Sox2 are responsible for its interaction with the nucleosome. A $^{1}H^{15}N$- heteronuclear single quantum coherence (HSQC) of full length Sox2 bound to nucleosomes (SHL+2) (**Figure 4.4a**) shows a similar narrow distribution of peaks like the unbound Sox2. Similar to when Sox2 is bound to DNA there are distinct chemical shift changes visible[467], but unlike when Sox2 is bound to DNA some resonance peaks from the C-IDR have substantially reduced intensity or fully disappeared from the spectrum. To further quantify these changes we turned to the HNCO spectra, as backbone chemical shifts are more sensitive to local conformational states.[474] We observe a distinct dip in peak intensities between residues 160-170 (**Figure 4.4c,d**), overlapping with the interaction region detected in our smFRET lifetime analysis.

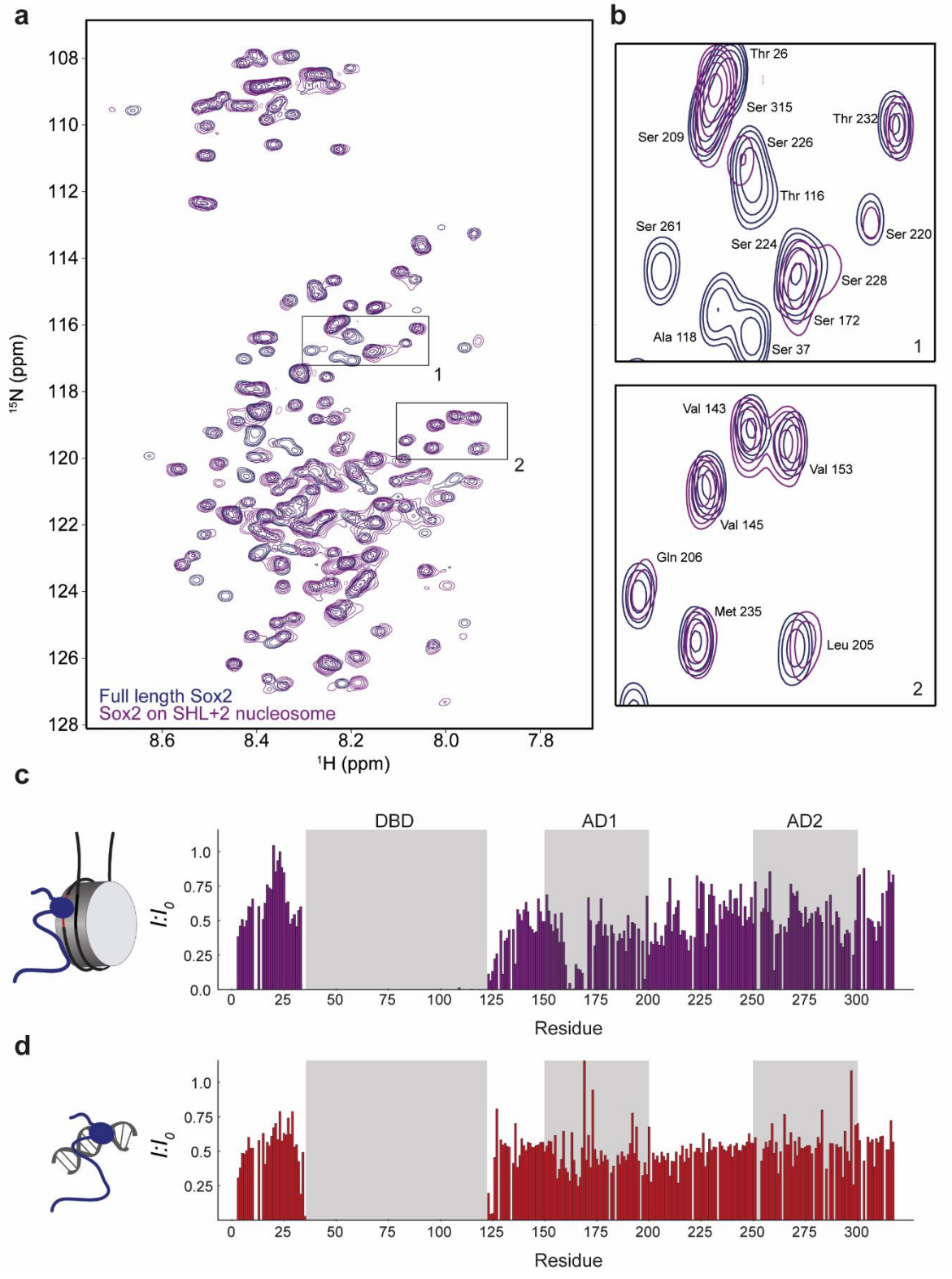



***Figure 4.4. A short activation domain interacts with the histone octamer. a)*** *$^{1}H^{15}N$ HSQCs of free Sox2 (blue) and Sox2 in complex with 197 bp unlabeled nucleosome (purple).* ***b)*** *Zoomed-in regions show resonances that are affected or unaffected by nucleosome binding.* ***c)*** *The change in intensity of NMR peaks from HNCO spectra when bound to 197 bp nucleosomes.* ***d)*** *The change in intensity of NMR peaks from HSQC spectra when bound to 15 bp DNA.*

We performed a sequence alignment for Sox2 from several organisms as well as Sox1, Sox3 and Sox17. We find that the 160-170 residues are well conserved within the SoxB family (**Supplementary Table 4.1**). The fact that the tryptophan is completely conserved across all compared organisms is noteworthy, especially since IDRs are typically depleted of tryptophans[475]. Interestingly, the conservation is not as pronounced with Sox17, indicating variability in this sequence across different Sox families. Additionally, the interaction site coincides with an activation domain[403] (AD1) that is predicted to contain some secondary structure[467].

### 4.3.3 Sox2 opens chromatosomes

Nucleosomes that are further compacted by the binding of linker histone H1 are called chromatosomes[476]. Formation of the chromatosome renders the DNA of nucleosomes even more inaccessible to transcription factors and other transcriptional machinery[477]. However, given Sox2's low nanomolar affinity to nucleosomes, we investigated whether Sox2 can bind and open chromatosomes. Using the same nucleosomes labelled in a11/b11 (SHL+6), we added histone H1 which shifts the transfer efficiency value to 0.8 (4.76 nm), corresponding to a nucleosome conformation with collapsed linkers as previously demonstrated[88] (**Figure 4.5a**). When Sox2 is added, a new and structurally heterogeneous ensemble of populations arises concomitantly with a decrease in the H1 bound population. This change suggests either a displacement of H1 or a conformational alteration of H1 within the chromatosome. However, we cannot conclusively determine from this data alone whether H1 remains bound following chromatosome opening. Upon further addition of Sox2, the H1 bound nucleosome population reaches undetectable levels and a single low-FRET population is observed (**Figure 4.5b**). Similarly to the free nucleosomes, we determined the apparent $K_D$ by following the reduction of the area of the H1-bound population (**Figure 4.5c,d**).

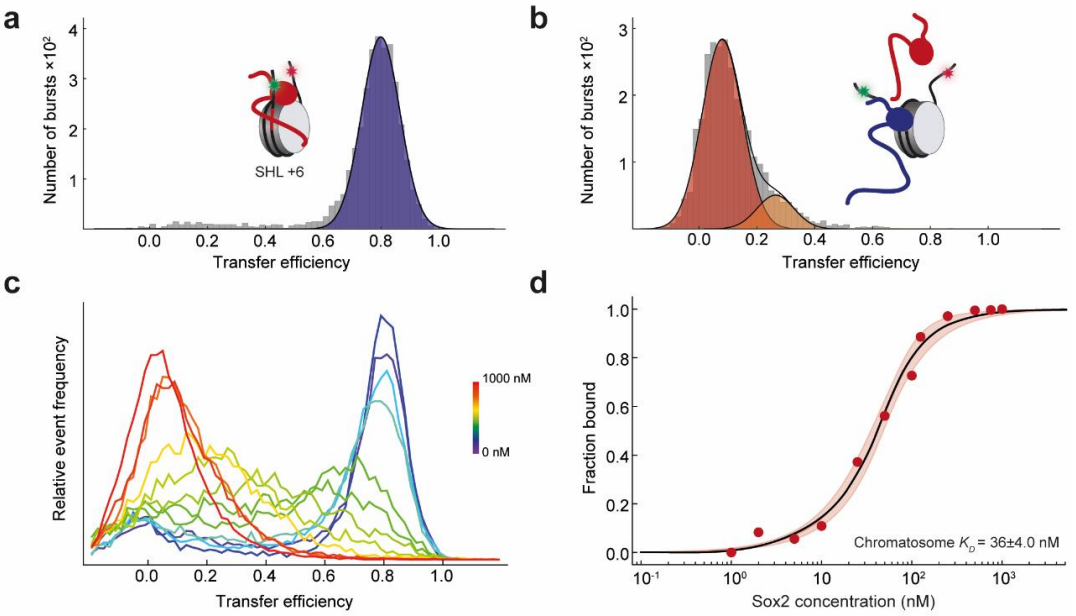



***Figure 4.5. Sox2 remodels both free and histone H1-bound nucleosomes a)*** *Single-molecule transfer efficiency histogram of 197 bp nucleosomes bound by histone H1.* ***b)***

*Single-molecule transfer efficiency histogram of 197 bp nucleosomes remodeled by Sox2.* ***c)*** *Single-molecule transfer efficiency histograms of fluorescently labelled nucleosomes bound by histone H1.* ***d)*** *The corresponding binding isotherm with a fit to a 1:1 binding model.*

We observed a significant decrease in binding affinity for the chromatosomes (36 ± 4 nM) compared to the nucleosomes (3 ± 1 nM), clearly demonstrating that linker histone H1 interferes with Sox2 chromatin remodeling.

### 4.3.4 Sox2 binding to nucleosomes reconstituted with native enhancer DNA

Native sequence nucleosomes, which are nucleosomes reconstituted with endogenous sequences extracted directly from the genome, often display a degree of inherent instability that can be critical for understanding the nuances of chromatin dynamics[478,479]. Unlike the widely used Widom 601 positioning sequence that forms highly stable nucleosomes for in vitro studies, native sequences can form nucleosomes with a range of stabilities, reflecting the diverse and dynamic nature of the genomic landscape. Of special interest to Sox2 are the nucleosomes containing the native sequence of *LIN28B*[480,481], a gene implicated in developmental pathways and stem cell rejuvenation. Upon reconstituting nucleosomes using the *LIN28B* sequence (**Supplementary Table 4.1**) we encountered a new challenge as the LIN28B nucleosomes were significantly less stable than the Widom nucleosomes. Specifically, these nucleosomes would disassemble when diluted down to smFRET concentrations (50-100 pM), resulting in a transfer efficiency of ~0, characteristic of free 197 bp DNA. Attempting to stabilize them, we incubated the LIN28B nucleosomes with linker histone H1 before diluting them to smFRET conditions. This resulted in a folded chromatosome (**Figure 4.6a**). Interestingly, the resulting transfer efficiency was significantly lower compared to the Widom chromatosome, 0.56 (5.7 nm) vs 0.8 (4.8 nm) suggesting a more open conformation. This compares well with previous computational findings that the LIN28B nucleosomes are less stable than those reconstituted with the Widom-601 sequence[482], but to the best of our knowledge that has never been demonstrated for a LIN28B chromatosome. Upon the addition of Sox2 (200 pM), there is a noticeable shift in transfer efficiency indicative of Sox2 induced remodelling, resulting in a single low FRET population (**Figure 4.6b**) corresponding to a more open conformation.

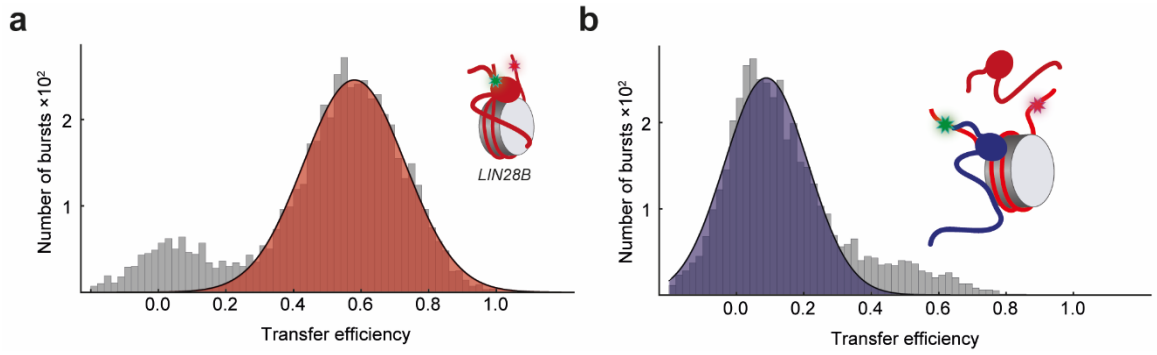



***Figure 4.6. Sox2 remodels nucleosomes from native DNA sequences. a)*** *Single-molecule transfer efficiency histogram of 197 bp LIN28B nucleosomes bound by histone H1.* ***b)*** *Single-molecule transfer efficiency histogram of 197 bp LIN28B nucleosomes remodeled by Sox2.*

## 4.4 Discussion

In this study, we have extended our investigation into the multifaceted role of Sox2, a pioneer transcription factor crucial for stem cell maintenance and differentiation. We present insights into the structural nuances of Sox2's interactions with chromatin and reveal that the conformational dynamics of Sox2, particularly within its C-IDR, exhibit notable changes when bound to nucleosomes.

We demonstrated Sox2's ability to remodel nucleosomes reconstituted using the Widom-601 sequence, incorporating the Sox2 binding sequence in various positions. Curiously, we observed that Sox2's binding affinity remained consistent regardless of the location of these sites, challenging previous research that the binding site's location could influence the strength of Sox2 nucleosome interactions[470]. A recent study reported Sox2 binding affinities ranging from 2 nM to 41 nM across different binding site locations[470]. While the sequence and the length of the DNA varied from our study, the core Sox2 binding motif "TTGT" remained consistent. One explanation for this discrepancy might be that the selection of binding site locations guided by Cryo-EM structures[53,366], inadvertently favoured high affinity positions. Alternatively, there may be an unidentified Sox2 binding site within the Widom-601 sequence that Sox2 has a higher affinity for then the introduced binding sites. We identified specific amino acids within Sox2 that are directly involved in nucleosome interactions, with their location in AD1 suggesting a functional involvement. Residues 160-170 are YAHMNGWSNGS, with the tryptophan, an aromatic residue commonly found in $\pi - \pi$ interactions[483], standing out as a likely key player of the interaction. While the conservation of this tryptophan within Sox2 across different organism, as well as in Sox1 and Sox3, indicates an important role for this aromatic residue, it is noteworthy that other regions of Sox2 sequence are also well conserved. This suggests that these regions could be conserved because of additional roles. Despite the interactions between the C-IDR and nucleosomes, we discovered that the overall binding affinity of Sox2 to nucleosomes remains the same in the absence of its C-IDR.

Furthermore, we demonstrate Sox2's capability to engage with and open chromatin structures bound by histone H1, using both the canonical Widom-601 nucleosome positioning sequence and the native-like Lin28b sequence. This capability highlights Sox2's role in facilitating chromatin accessibility, a key step in transcriptional initiation and gene regulation.

Building on these findings, we initially observe that the histone H3 tails, which protrude from the nucleosome at specific SHLs, contain an aromatic residue, tyrosine. These tails are dynamic, adopting multiple configurations that can directly interact with the nucleosome core and influence DNA breathing dynamics and nucleosome stability.[484] Based on these observations, we propose two models for the mechanism by which Sox2 interacts with core histones. First, we suggest that the region of Sox2's C-IDR, particularly the conserved tryptophan, may engage in π-π stacking interactions with the tyrosine in the H3 tail. This interaction could potentially lead to a reduction in nucleosome stability, similar to the effect observed in nucleosomes with a truncated, tail-less H3 (**Figure 4.7**).[485] Another possible model considers Sox2's potential disruption of histone H3's role in chromatin architecture. The H3 tail predominantly engages in intra-nucleosome interactions within loosely packed chromatin but shifts to mostly inter-nucleosome interactions in condensed chromatin.[486] Sox2's interference with these H3 mediated inter-nucleosome interactions could play a role

in chromatin decondensation. However, it is important to note that the involvement of π-π interactions is hypothesized based on the aromatic nature of these residues and that further experimental investigations are required to substantiate these models.

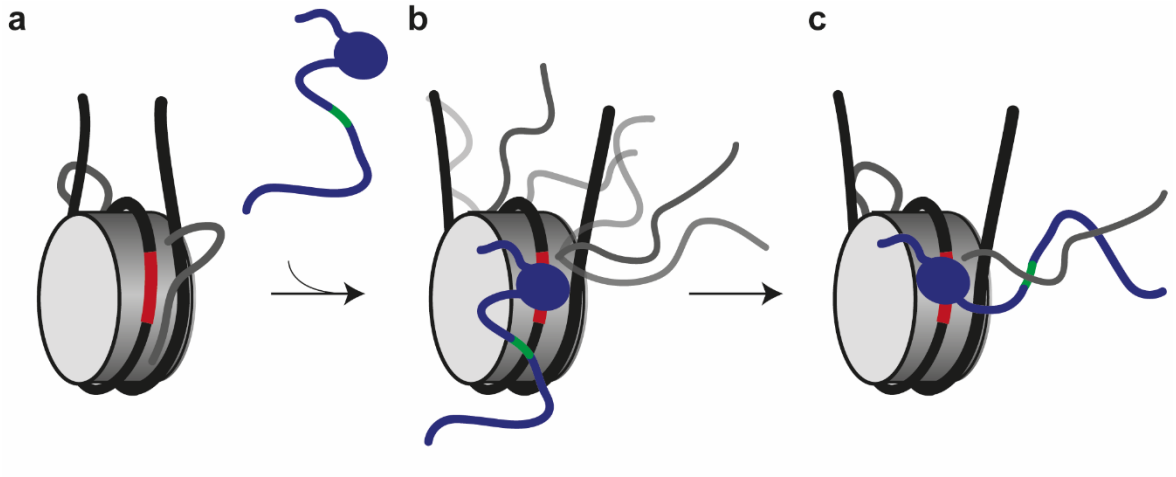



***Figure 4.7. Sequential Binding model of Sox2 on SHL+2 nucleosome. a)*** *Illustration of a nucleosome with a Sox2 binding site at SHL+2. Residues 160-170 on Sox2 are highlighted in green. For clarity, only the disordered tails of histone H3 (gray) are shown.* ***b)*** *Sox2 binds to the nucleosome, destabilizing it, leading to the release of the histone H3 tails.* ***c)*** *Sox2 seizes the disengaged histone H3 tail preventing reengagement, further destabilizing the nucleosome.*

The interplay between Sox2 and chromatin, particularly its pioneering activity, has been a subject of intense study[487]. Prior research has primarily focused on the DBDs of pTFs[468-470], highlighting their role in recognizing specific DNA sequences within nucleosomes. Our findings contribute to this body of work by offering insights into the interactions facilitated by the IDRs of Sox2, a feature less explored in the literature. Given the importance of IDRs in pTFs[488], our findings hint at the potential significance of Sox2's C-IDR interaction with nucleosomes, although the precise impact of this interaction remains unknown. Moreover, our observation that the binding affinity of Sox2 for nucleosomes remains unchanged despite the location of the specific binding site challenges previous work that has found that the binding depends strongly on the location[470]. In addition, our demonstration of Sox2's ability to interact with and open histone H1 bound nucleosomes underlines its role in chromatin dynamics.

We acknowledge certain constraints, notably the predominant use of the robust Widom nucleosome sequence, a limitation somewhat mitigated by the use of the native-like Lin28b sequence. Our experiments focused on mononucleosomes, which, while informative, do not fully capture the complexity of the natural chromatin environment. Additionally, the observed direct interaction between Sox2's C-IDR and nucleosomes, may not necessarily extend to other pioneer transcription factors, constraining the broader impact of the study. Further studies will reveal the implications of modifying or eliminating the interaction domain within Sox2's C-IDR on its functional dynamics and chromatin interactions. This work sets a foundation for future investigations into the molecular mechanisms of pioneer

transcription factor behaviour at the molecular level, that could lead to the development of novel therapeutic strategies targeting chromatin remodelling.

# 4.5 Methods

## 4.5.1 Protein expression and purification

The constructs for all Sox2 variants were expressed using a pET24b vector including a hexahistidine small ubiquitin-like modifier (His6-SUMO) tag at the N-terminus. Site-directed mutagenesis was performed using the QuikChange Lightning kit (Agilent), with primers obtained from Integrated DNA Technologies (IDT). Protein expression was carried out in Lemo21(DE3) cells (New England BioLabs) in either LB-broth or M9 minimal medium supplemented with $^{15}N$-$NH_4Cl$ and $^{13}C6$-glucose for isotopic labelling. Expression was induced at OD600 of 0.5-0.7 using 0.4 mM Isopropyl β- d-1-thiogalactopyranoside (IPTG), followed by a 2–3-hour expression period at 37°C. Cells were then harvested by centrifugation at 4500 × g for 15 min and lysed in Buffer A (50 mM NaH2PO4, 300 mM NaCl, 10 mM imidazole, 6 M urea, 1 mM dithiothreitol (DTT), pH 8.0) overnight at 4°C. Lysate was cleared by centrifugation, with the supernatant purified using an HisTrap HP column (Cytiva). Sample was eluted using high imidazole buffer A, and samples were dialysed against buffer B (50 mM Tris, 150 mM NaCl, 1 mM DTT, pH 8.0) for His6-SUMO tag cleavage by ULP1 protease (produced in-house). Post-cleavage, Sox2 DBD was further purified with HiTrap SP Sepharose FF column (Cytiva), while the full length Sox2 precipitated out of solution and was collected by centrifugation. For the final step all proteins were purified using RP-HPLC on a ZORBAX 300SB-C3 column (Agilent), for a more detailed description of the purification see our previous work[467]. The final protein purity was confirmed by SDS-PAGE before lyophilization and storage at -20°C.

## 4.5.2 Protein labelling

Lyophilized proteins were dissolved in labelling buffer (0.1 M potassium phosphate, 1 M urea, pH 7.0) for overnight labelling at 4°C with Cy3B maleimide (donor)(Cytiva) at a 0.7:1 dye to protein ratio. Post labelling, DTT was added to quench the reaction, followed by RP-HPLC to eliminate excess dye and to isolate singly and doubly donor-labelled proteins. After lyophilization, proteins were again dissolved in the same buffer and subjected to overnight labelling with CF660R maleimide (acceptor)(Sigma). As with the donor labelling process, the reaction was quenched with DTT, and RP-HPLC was employed to remove unbound dye and to separate proteins with double donor or double acceptor labels. Finally, proteins labelled with both donor and acceptor were lyophilized, dissolved in 8 M GdmCl, flash frozen in liquid nitrogen, and stored at -80°C for future use.

## 4.5.3 DNA labelling

Oligonucleotide aliquots (5–10 nmol), featuring a thymine modified with a C6-amino linker for NHS ester dye conjugation (IDT), were dissolved in 50 µl of DNA labeling buffer (0.1 M sodium bicarbonate, pH 8.3). These were then labelled with Cy3B NHS ester (Cytiva) or

CF660R NHS ester (Sigma) using a 2:1 dye-to-DNA ratio. The labelling reaction proceeded for a minimum of two hours at ambient temperature before being ethanol-precipitated to discard unbound dye. The resultant pellet was resolubilized in 100 µl of 95% RP-HPLC solvent C (0.1 M triethylammonium acetate) and 5% RP-HPLC solvent D (acetonitrile), then purified using RP-HPLC on a ReproSil Gold 200 C18 column (Dr. Maisch) to separate labelled oligonucleotides from unreacted dye and unlabelled strands. The purified, labelled oligonucleotides were lyophilized, reconstituted in double-distilled water ($ddH_2O$) to achieve a concentration of 2.5 µM, and stored at -20°C for subsequent PCR amplification.

### 4.5.4 Nucleosome reconstitution

DNA for nucleosome assembly for smFRET experiments was obtained by PCR amplification of a pJ201 plasmid containing either the 147 bp Widom sequence including a Sox2 binding site (CTTTGTTATGCAAAT) and 25 bp linkers on both ends or the Lin28b sequence. This amplification utilized either fluorescently labelled oligonucleotides (as described in DNA labelling) or unlabelled ones from IDT. Post amplification, the PCR products were purified using ethanol precipitation followed by a DNA Clean and Concentrator Kit (Zymo Research), with DNA concentrations determined via UV-Vis spectroscopy. For primers and sequences see Supplementary Table 4.2. DNA for nucleosome assembly for NMR experiments was obtained via restriction digest of a high copy number plasmid containing the 197 bp Widom sequence with a Sox2 binding site at SHL+2. For reconstitution, 10 pmol (labelled DNA) or 100 pmol (unlabelled DNA) of the 197 bp Widom sequence (with and without Sox2 binding site modifications) or the Lin28b was mixed with 1.0-1.75 molar equivalents of recombinant histone octamer (The Histone Source) in a buffer of 10 mM Tris, 0.1 mM EDTA, and 2 M KCl (pH 7.5), and incubated on ice. The mixture was then dialyzed against a decreasing KCl gradient (to 10 mM) over 20 hours at 4°C using Slide-A-Lyzer MINI dialysis buttons (Thermo Fisher Scientific), with buffer exchange facilitated by a peristaltic pump to maintain a constant volume. Post dialysis, the solution was centrifuged at 20.000 × g for 5 minutes at 4°C to remove any aggregates. The nucleosome concentration and purity were assessed by measuring absorbance at 260 nm. A 0.5 pmol aliquot of each sample was analyzed on a 0.7% agarose gel at 90V for 90 minutes, using 0.25× Tris-borate as the running buffer. Gels were stained with GelRed (Biotium) and visualized with the Gel Doc EZ system (Bio-Rad). Only samples with less than 5% free DNA were used for subsequent experiments.

### 4.5.5 Single-molecule spectroscopy

All experiments were carried out at ambient temperatures utilizing a MicroTime 200 system (PicoQuant) integrated with an Olympus IX73 inverted microscope. Excitation of the donor dye was achieved with a 520 nm diode laser (LDH-D-C-520, PicoQuant), and a 640 nm diode laser (LDH-D-C-640, PicoQuant) was employed for the acceptor dye, utilizing Pulsed Interleaved Excitation (PIE) at a 40 MHz repetition rate. Laser outputs were set to 40 µW for 520 nm and 20 µW for 640 nm (measured with PM100D, Thorlabs). A 60× water-immersion objective (UPLSAPO60XW, Olympus) was used for focusing the excitation light and collecting emitted fluorescence, which was then directed through a 100 µm pinhole. Subsequent separation of fluorescence signals into four channels was achieved first by polarization and then by wavelength, using specific filters for donor (582/64 BrightLine HC, Semrock) and acceptor (690/70 H Bandpass, AHF) emissions. Photon detection was performed with single photon avalanche diodes (SPCM-AQRG-TR, Excelitas

Technologies), and photon arrival times were captured by a MultiHarp 150 P TCSPC module (PicoQuant). Measurements took place in µ-Slide chambers (Ibidi), using TEK buffer (10 mM Tris, 0.1 mM EDTA, pH 7.4) with added 2-mercaptoethanol (143 mM, Sigma) for photostability and 0.01% Tween-20 (v/v, AppliChem) to minimize surface adhesion.

### 4.5.6 Analysis of transfer efficiency histograms

Data for transfer efficiency histograms were collected from 50 to 100 pM concentrations of double labelled freely diffusing Sox2 or nucleosomes. Analysis was performed using the "Fretica" Mathematica scripting package (https://schuler.bioc.uzh.ch/programs/), created by Daniel Nettels and Ben Schuler. Fluorescence bursts were identified by aggregating photons with interphoton durations under 100 µs. Transfer efficiencies for each burst were computed as $E = n'_A/(n'_A + n'_D)$, with $n'_A$ and $n'_D$ denoting the counts of acceptor and donor photons, respectively, adjusted for background noise, direct acceptor excitation, cross-channel interference, and disparities in dye quantum yields and photon detection efficiencies. Bursts affected by acceptor bleaching during transit through the confocal volume were excluded to avoid underestimating FRET. Bursts significantly exceeding the average photon counts, indicative of aggregates, were also removed. The stoichiometry ratio (S) for labelling was calculated using:

$$S = \frac{n_D^D + n_A^D}{n_D^D + n_A^D + n_A^A}$$

where $n_{D/A}^D$ is the count of detected donor or acceptor photons post donor excitation and $n_A^A$ is the count of acceptor photons post acceptor excitation. To create the final transfer efficiency histograms, we selected bursts that have S= 0.3-0.7, this selection helped isolate bursts from incorrectly labelled molecules, discarding those without an active acceptor. Despite filtering, a large donor only population can cause residual donor only bursts to remain. To calculate average FRET efficiencies, we applied Gaussian or lognormal distribution fits to the histograms, identifying distinct populations(s) within. For comprehensive binding affinity analysis, multiple histograms were fitted globally, allowing certain parameters to be consistent across different measurements. Distance metrics from FRET efficiency data for nucleosomes utilized the Förster resonance energy transfer equation:

$$E(r) = \frac{1}{1 + r^6/R_0^6}$$

where $R_0 = 6.0$ nm for the Cy3B/CF660R dye pair. For double labelled Sox2 involving disordered regions we converted mean transfer efficiencies $\langle E \rangle$ to root-mean-square end-to end distances $R = \sqrt{\langle r^2 \rangle}$ by numerically solving the following transcendental equation:

$$\langle E \rangle = \int_0^\infty dr\, E(r) P(r)\ .$$

Here, $P(r)$ denotes the distance probability density function of the SAW-$\nu$ model:

$$P(r) = A\frac{4\pi}{R}\left(\frac{r}{R}\right)^{2+(\gamma-1)/\nu} \exp\left(-\alpha\left(\frac{r}{R}\right)^{1/(1-\nu)}\right),$$

which is characterized by the critical exponents, $\nu$ and $\gamma \approx 1.1615$. The constants $A$ and $\alpha$ are determined by forcing $P(r)$ to be normalized and to satisfy $\langle r^2\rangle = R^2$. By adopting a scaling law $R = b\,N^{\nu}$, we eliminate the $\nu$ dependency from $P(r)$, substituting it with the expression $\nu = \ln\left(\frac{R}{b}\right)/\ln(N)$ where $b \approx 0.55$ corresponding an average value of approximately 0.55 nm for proteins and $N$ indicates the number of monomer units separating the fluorescent dyes. From this, we can infer the corresponding radius of gyration $R_g$, which is estimated using:

$$R_g \approx R\sqrt{\frac{\gamma(\gamma+1)}{2(\gamma+2\nu)(\gamma+2\nu+1)}}.$$

### 4.5.7 Fluorescence correlation spectroscopy

To calculate the diffusion time of labelled Sox2, we performed fluorescence correlation spectroscopy by correlating the intensity fluctuations in fluorescence in an smFRET experiment according to:

$$G_{ij}(\tau) = \frac{\langle \delta n_i(0)\delta n_j(\tau)\rangle}{\langle n_i\rangle^2}$$

where where i,j=A, D and $n_i(0)$ and $n_j(\tau)$ are fluorescence count rates for channels i and j at time 0 and after a lag time t, respectively, and $\delta n_{i,j} = n_{i,j} - \langle n_{i,j}\rangle$ are the corresponding deviations from the mean count rates.

### 4.5.8 Fluorescence lifetime analysis

Fluorescence lifetimes were derived from the average detection times of the donor $\langle t_D\rangle$ following its excitation pulse. These lifetimes were subsequently mapped against their corresponding transfer efficiencies on two-dimensional scatter plots. Here, $\tau_{DA}/\tau_D = \langle t_{\mathrm{D}}\rangle/\tau_D$ was determined for each fluorescence burst given an inherent donor lifetime $\tau_D$. In scenarios where donor-acceptor distance remains constant, the ratio $\langle\tau_{DA}\rangle/\tau_D$ must match $1-E$. However, in systems exploring a wide range of distances, governed by a probability density function P(r) for the inter dye distance r, the mean fluorescence lifetime $\langle\tau_{DA}\rangle$ is influenced by the distribution of distances as according to:

$$\frac{\langle\tau_{DA}\rangle}{\tau_D} = 1 - \langle E\rangle + \frac{\sigma^2}{1-\langle E\rangle}.$$

Here, the variance $\sigma^2$ is given by:

$$\sigma^2 = \langle E^2\rangle - \langle E\rangle^2 = \int_0^{\infty} dr\,[E(r) - \langle E\rangle]^2 P(r)\,.$$

### 4.5.9 Binding affinity measurements

Transfer efficiency histograms for doubly labelled Sox2 or nucleosomes were collected as the concentration of the unlabelled counterpart was incrementally increased until a stable transfer efficiency was observed. The data were segmented into two distinct subpopulations, representing bound and unbound states, using Gaussian peak fitting. The proportion of the bound fraction ($\theta$) was then determined based on the relative area under these peaks. To derive the dissociation constant ($K_D$), we applied a binding isotherm model to the data, fitting it with the following equation:

$$\theta = \frac{c_{X,tot} + K_D + c_{Y,tot} + \sqrt{\left(c_{X,tot} + K_D + c_{Y,tot}\right)^2 - 4c_{X,tot}c_{Y,tot}}}{2c_{Y,tot}}$$

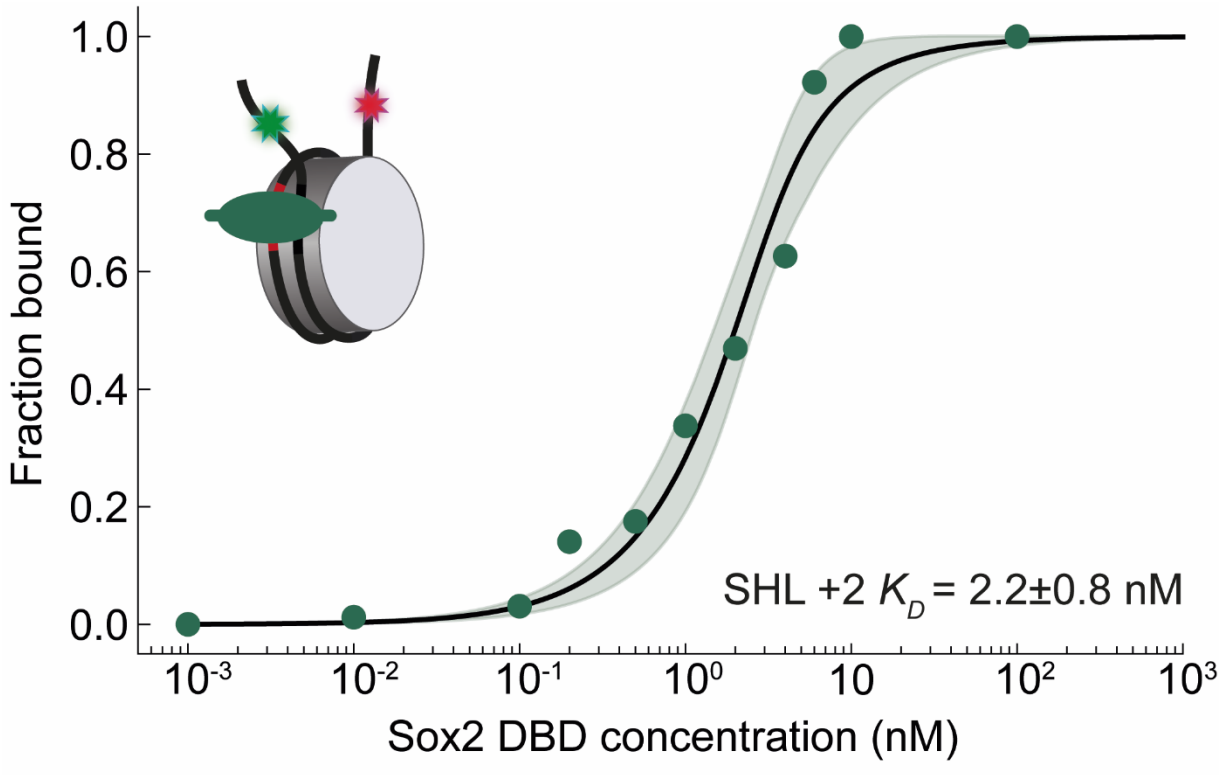
where $c_{x,tot}$ and $c_{Y,tot}$ are the total concentration of Sox2 and nucleosomes, respectively, with one kept constant throughout the experiment.

### 4.5.10 NMR spectroscopy

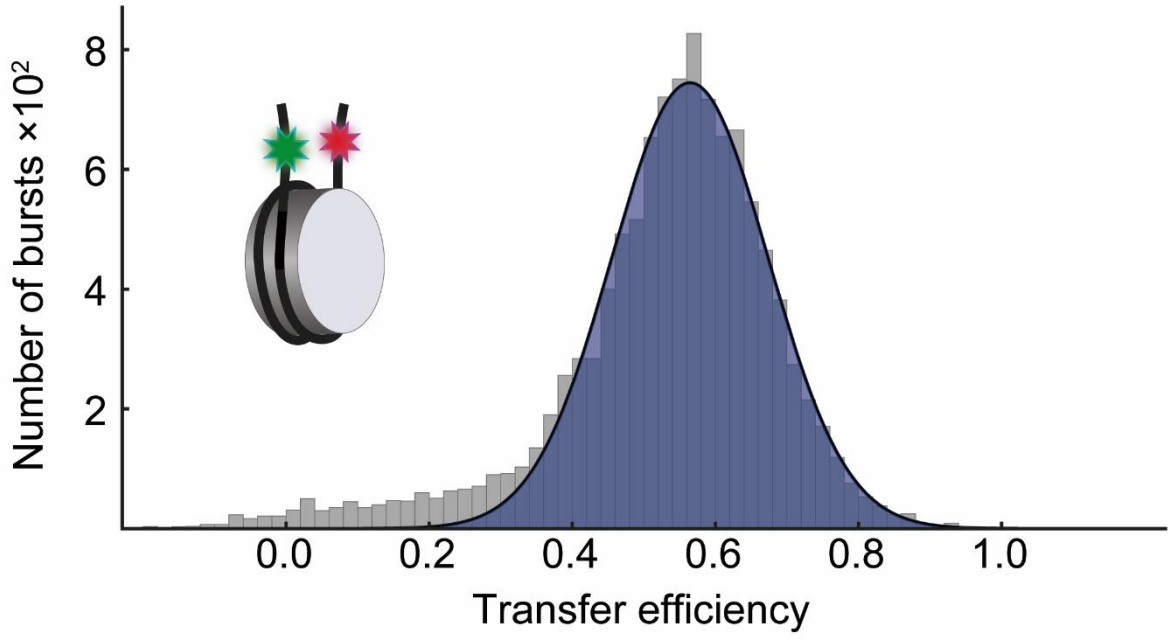
All NMR data were acquired using a Bruker Avance spectrometer operating at 850MHz, equipped with a triple resonance ($^{15}N/^{13}C/^{1}H$) cryo-probe. Experiments were performed in a buffer containing 20 mM $NaH_2PO_4$, 50 mM NaCl, 2 mM tris(2-carboxyethyl)phosphine (TCEP), 125 µM DSS, and 5% $D_2O$ (v/v), adjusted to pH 6.0 and maintained at 15°C to reduce amide proton exchange. Backbone assignments for $^{13}C$,$^{15}N$ labelled Sox2 in its free state (90 µM concentration) were derived from $^{1}H$-$^{15}N$ HSQC, HNCACB, CBCA(CO)NH, HN(CO)CA, HNCO, and HN(CA)NNH spectra. Chemical shift perturbations (CSPs) due to binding were assessed at 90 µM of Sox2, comparing conditions without and with a 1.1 molar excess of the 197 bp Widom nucleosome with a Sox2 binding site (SHL+2) (Supplementary Table 4.2 for details). CSP calculations were performed as follows:

$$CSP = \sqrt{\frac{1}{2}((\Delta\delta_H)^2 + (0.154 * \Delta\delta_N)^2)}$$

## 4.6 Supplementary information

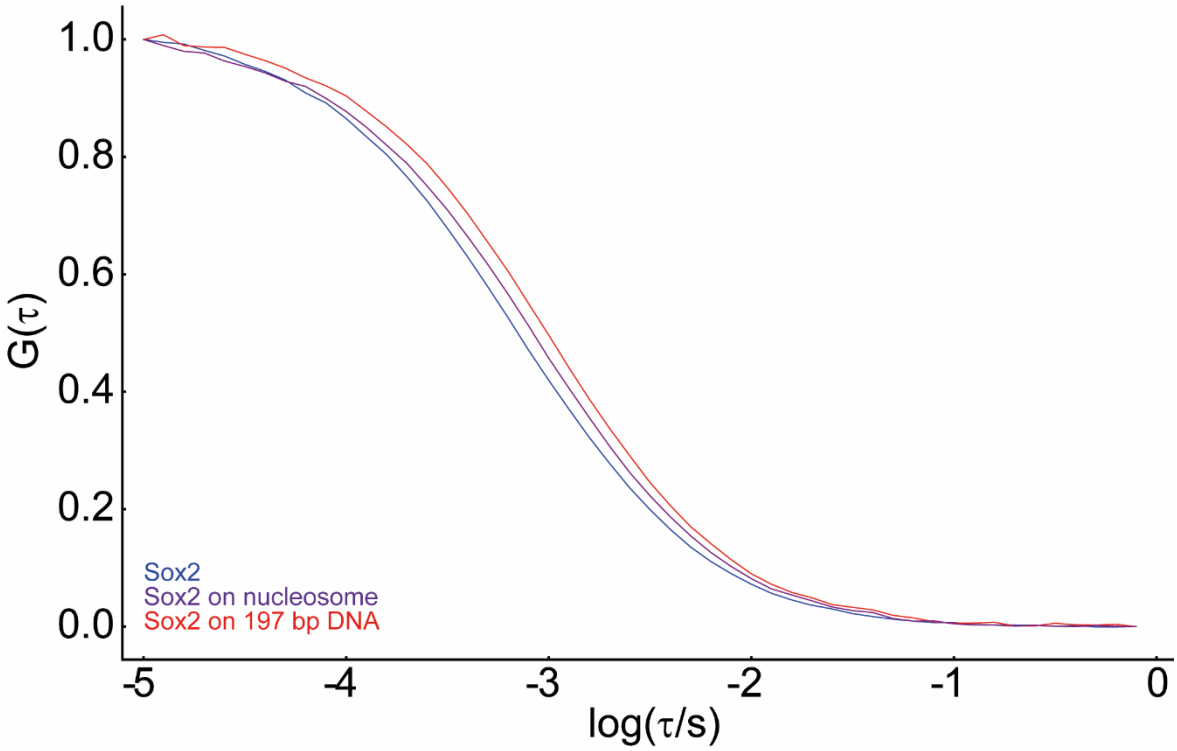



***Supplementary Figure 4.1. Binding affinity of Sox2 DBD to SHL+2 nucleosomes.*** *A binding isotherm for Sox2 DBD to SHL+2 197 bp nucleosomes with a fit to a 1:1 binding model.*

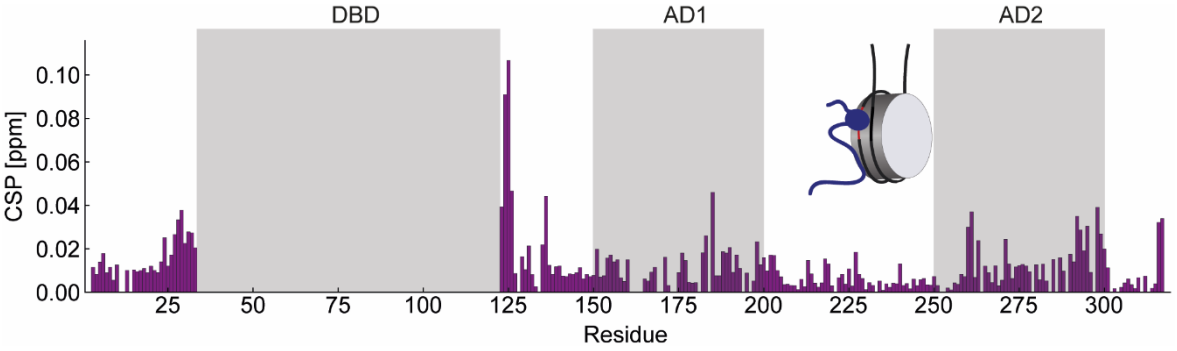



***Supplementary Figure 4.2. Single-molecule analysis of 197 bp Widom nucleosomes.*** *Single-molecule transfer efficiency histogram of 197 bp Widom nucleosomes.*


***Supplementary Figure 4.3. Sequence alignment of Sox 1,2,3 and Sox17.*** *N-DBD and C-IDR regions are colored green and gray, respectively. Interaction site (residues 160-170, Sox2) is highlighted. Alignment generated by CLUSTAL O*[451].




***Supplementary Figure 4.4. Donor-acceptor cross-correlation of fluorescently labelled Sox2.*** *Sox2 labelled at residues 159 and 265 diffuses through the confocal volume relatively quickly when unbound (blue). When bound to SHL+2 nucleosomes (100 nM) it diffuses slower (purple). When bound to the 197 bp DNA the diffusion is further reduced due to the larger hydrodynamic radius of the DNA.*




***Supplementary Figure 4.5. Chemical shift perturbation of Sox2 bound to SHL+2 nucleosome.*** *CSP plot showing the chemical shift difference in HNCO spectra between free Sox2 and Sox2 bound to SHL+2 nucleosome.*

***Supplementary Table 4.1. Sox2 sequence alignment.*** *Residues 160-170 (Human) are highlighted.*

```
Zebrafish            MYNMMETELKPPAPQPNTGGTGNT-------NSSGNNQKNSPDRIKRPMNAFMVWSRGQR  53
Western clawed frog  MYNMMETDLKPPAPQQASGGNSNSG--------SNNQSKNSPDRVKRPMNAFMVWSRGQR  52
African clawed frog  MYSMMETELKPPAPQQPSGGNSNSA--------SNNQNKNSPDRVKRPMNAFMVWSRGQR  52
Chicken              MYNMMETELKPPAPQQTSGGGTGNS------NSAANNQKNSPDRVKRPMNAFMVWSRGQR  54
Mouse                MYNMMETELKPPGPQQASGGGG--GGGNATAAATGGNQKNSPDRVKRPMNAFMVWSRGQR  58
African elephant     MYNMMETELKPPGPQQTSGGG---GGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQR  57
Human                MYNMMETELKPPGPQQTSGG----GGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQR  56
Sheep                MYNMMETELKPPGPQQTSGGGG-GGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQR  59
Goat                 MYNMMETELKPPGPQQTSGGGG-GGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQR  59
Lion                 MYNMMETELKPPGPQQTSGGGGGGGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQR  60
Chimpanzee           MYNMMETELKPPGPQQTSGGG--GGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQR  58

Zebrafish            RKMAQENPKMHNSEISKRLGAEWKLLSESEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  113
Western clawed frog  RKMAQENPKMHNSEISKRLGAEWKLLSEAEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  112
African clawed frog  RKMAQENPKMHNSEISKRLGAEWKLLSEAEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  112
Chicken              RKMAQENPKMHNSEISKRLGAEWKLLSEAEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  114
Mouse                RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  118
African elephant     RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKKHPDYKYRPRRKT  117
Human                RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  116
Sheep                RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  119
Goat                 RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  119
Lion                 RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  120
Chimpanzee           RKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKT  118
```

*Continued*

```
Zebrafish            KTLMKKDKYTLPGGLLAPGGNGMGAGVGVGAGLGAGVNQRMDSYAHMNGWTNGGYGMMQE  173
Western clawed frog  KTLMKKDKYTLPGGLLAPGANPMTSGV--GASLGAGVNQRMDTYAHMNGWTNGGYGMMQE  170
African clawed frog  KTLMKKDKYTLPGGLLAPGANAMTSGV--GGSLGAGVNQRMDTYAHMNGWTNGGYGMMQE  170
Chicken              KTLMKKDKYTLPGGLLAPGTNTMTTGVGVGATLGAGVNQRMDSYAHMNGWTNGGYGMMQE  174
Mouse                KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGGGLNQRMDSYAHMNGWSNGSYSMMQE  178
African elephant     KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYGVMQD  177
Human                KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQD  176
Sheep                KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQD  179
Goat                 KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQD  179
Lion                 KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQD  180
Chimpanzee           KTLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQD  178


Zebrafish            QLGYPQHPSLNAHNTAQMQPMHRYDMSALQYNSMTNSQTYMNGSPTYSMSYSQQSTPGMT  233
Western clawed frog  QLGYPQHPGLSAHNAPQMQPMHRYDVSALQYNSMSSSQTYMNGSPTYSMSYSQQGAPGMS  230
African clawed frog  QLGYPQHPGLNAHNAPQMQPMHRYDVSALQYNSMSSSQTYMNGSPTYSMSYSQQGAPGMS  230
Chicken              QLGYPQHPGLNAHNAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  234
Mouse                QLGYPQHPGLNAHGAAQMQPMHRYVVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  238
African elephant     QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYTQQATPGMA  237
Human                QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  236
Sheep                QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  239
Goat                 QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  239
Lion                 QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  240
Chimpanzee           QLGYPQHPGLNAHGAAQMQPMHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMA  238
```

*Continued*

```
Zebrafish            LGSMGSVVKSESSSSPPVVTSSSHSRAGQCQTGDLRDMISMYLPGAEVQDQSAQSRLHMS  293
Western clawed frog  LGSMGSVVKSESSSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAQSRLHMS  289
African clawed frog  LGSMGSVVKSESSSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPESAAQSRLHMS  289
Chicken              LGSMGSVVKTESSSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  293
Mouse                LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMA  297
African elephant     LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  296
Human                LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  295
Sheep                LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  298
Goat                 LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  298
Lion                 LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  299
Chimpanzee           LGSMGSVVKSEASSSPPVVTSSSHSRA-PCQAGDLRDMISMYLPGAEVPEPAAPSRLHMS  297

Zebrafish            QHYQSAPVPGTTINGTIPLSHM                                       315
Western clawed frog  QHYQSASVAGTAINGTLPLSHM                                       311
African clawed frog  QHYQSASVAGTGINGTLPLSHM                                       311
Chicken              QHYQSAPVPGTAINGTLPLSHM                                       315
Mouse                QHYQSGPVPGTAKYGTLPLSHM                                       319
African elephant     QHYQSGPVPGTAINGTLPLSHM                                       318
Human                QHYQSGPVPGTAINGTLPLSHM                                       317
Sheep                QHYQSGPVPGTAINGTLPLSHM                                       320
Goat                 QHYQSGPVPGTAINGTLPLSHM                                       320
Lion                 QHYQSGPVPGTAINGTLPLSHM                                       321
Chimpanzee           QHYQSGPVPGTAINGTLPLSHM                                       319
```

***Supplementary Table 4.2. DNA constructs for binding experiments.*** *The Sox2 binding site in the 601-Widom sequence is marked in red.*

| | |
|---|---|
| '601'-Widom sequence 197 bp | 5'-TCCATGGACCCTATACGCGGCCGCCCTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACA GCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAG TCTCCAGGCACGTGTCAGATATATACATCCTGTGCATGTATTGAACAGCAGTATGCCT-3' |
| '601'-Widom sequence with Sox2 binding site at SHL +6, forward strand | 5'-TCCATGGACCCTATACGCGGCCGCCCTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCG TAGACAGCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAG GGGATTACTCCCTAGTCTCCAGGCCTTTGTTATGCAAATACATCCTGTGCATGTATTGAACA GCAGTATGCCT-3' |
| '601'-Widom sequence with Sox2 binding site at SHL -6, forward strand | 5'-TCCATGGACCCTATACGCGGCCGCCCTGGAGACTTTGTTATGCAAATCCGCTCAATTGGTCGTAGACAG CTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTC TCCAGGCACGTGTCAGATATATACATCCTGTGCATGTATTGAACAGCAGTATGCCT |
| '601'-Widom sequence with Sox2 binding site at SHL +2, forward strand | 5'-TCCATGGACCCTATACGCGGCCGCCCTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACA GCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTCTTTGTTATGCAAATTACTCCCTAGT CTCCAGGCACGTGTCAGATATATACATCCTGTGCATGTATTGAACAGCAGTATGCCT-3' |
| *LIN28B* sequence with 25 bp linkers | ACCCTATACGCGGCCGCCAGTGGTATTAACATATCCTCAGTGGTGAGTATTAACATGGAACTTACTCCAAC AATACAGATGCTGAATAAATGTAGTCTAAGTGAAGGAAGAAGGAAAGGTGGGAGCTGCCATCACTCAGA ATTGTCCAGCAGGGATTGTGCAAGCTTGTGAATAAAGACAGCATGTATTGAACAGCA |

***Supplementary Table 4.3. Sox2 variants and Histone H1.0 used in this study.*** *The residues used for fluorophore labelling (after substituting for cysteine) are marked in red.*

| | |
|---|---|
| Full-length Sox2 | MYNMMETELKPPGPQQTSGGGGGNSTAAAAGGNQKNSPDRVKRPMNAFMVWSRGQRRKMAQENP KMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKTKTLMKKDKYTLPGGLLAPG GNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQDQLGYPQHPGLNAHGAAQMQP MHRYDVSALQYNSMTSSQTYMNGSPTYSMSYSQQGTPGMALGSMGSVVKSEASSSPPVVTSSSHSRAP CQAGDLRDMISMYLPGAEVPEPAAPSRLHMSQHYQSGPVPGTAINGTLPLSHM |
| Sox2 DBD | KNSPDRVKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALH MKEHPDYKYRPRRKTKTLMKK |
| Histone H1.0 | MTENSTSAPAAKPKRAKASKKSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKVGENADSQIKLS IKRLVTTGVLKQTKGVGASGSFRLAKSDEPKKSVAFKKTKKEIKKVATPKKASKPKKAASKAPTKKPKATPVK KAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVKPKAKSSAKRAGKKK |

***Supplementary Table 4.4. Sequence of disordered tails of the 4 core histones.*** *Aromatic residues are marked in red.*

| | |
|---|---|
| H2A | MSGRGKQGGKARAKAK |
| H2B | MPEPSKSAPAPKKGSKKAITKAQKKDGKKRKRSRKES |
| H3 | MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYR |
| H4 | MSGRGKGGKGLGKGGAKRHRKVLRD |

# 5 Conclusions and Perspectives

By addressing the challenges of characterizing disordered proteins, particularly TFs, our work has significantly advanced the understanding of Sox2, a key TF in development and pluripotency. In our study we have reconstructed the structural ensembles for Sox2, both in its free form and when complexed with DNA. This was made possible by integrating single-molecule and ensemble techniques with computational modelling.

In **Chapter 2** we reviewed the literature concerning IDPs in the nucleus, with a focus on disordered interactions with and within nucleosomes. Despite significant research on IDPs and their role in nuclear homeostasis, the nature of their interactions with chromatin and the biological implications of these interactions are yet to be fully elucidated. We highlighted recent advancements in the study of various classes of nuclear proteins, acknowledging the challenges posed by their dynamic nature. To address these challenges, we discussed integration of data from single-molecule techniques with computational modelling. Furthermore, we emphasised the importance of collaborative efforts that merge experimental and computational research to unravel these dynamic systems.

Our findings in **Chapter 3** challenge the predicted weak polyampholyte behaviour of Sox2's C-IDR, expected to form a collapsed structure. Contrary to this expectation, the C-IDR exhibits dynamic interactions with the DBD, mostly driven by charge interactions. These findings highlight the sensitivity of Sox2's dimensions to physiological salt concentrations, suggesting local intracellular variations could fine tune the C-IDR accessibility. The interaction between the C-IDR and the DBD not only influences the overall dimensions of Sox2 but also regulates accessibility to its activation domains. Notably, DNA binding affinity remains unaffected by these interdomain interactions, underscoring the importance of maintaining a dynamic ensemble for functional versatility.

Our work in **Chapter 4** focused on Sox2's interactions with nucleosomes, with a particular focus on the conformational changes within its C-IDR when bound to nucleosomes. We identified amino acids, notably including an aromatic tryptophan, that interact directly with the core histones. These amino acids appear to be well conserved across Sox2 and other members of the SoxB family, suggesting a functional role for them. Our findings indicate that Sox2 maintains its nucleosome binding affinity regardless of the presence of the C-IDR and we demonstrate Sox2's remodelling ability by opening histone H1-bound structures.

This detailed insight into the structural ensemble of Sox2 not only enhances our understanding of its function at a molecular level but could also shed light on its broader biological implications, such as cell reprogramming, regenerative medicine, and cancer biology. The dynamic interactions and structural flexibility of Sox2 are likely pivotal in modulating its binding to specific genomic loci and interaction partners. Our findings pave the way for further research into the various roles Sox2 serves. The structural ensemble provides a framework to explore the effects of mutations, environmental factors, and binding partners on Sox2. Gaining this understanding is important for guiding the development of targeted therapies or mutations to influence the functioning of Sox2. The structural characterization represents progress in protein chemistry and contributes to a broader

understanding of the mechanisms behind cell reprogramming, tissue regeneration, and cancer therapies.

While the detailed biophysical characterization and structural ensembles of TFs like Sox2 provide invaluable insights into the function of Sox2, they are only one of the many methods being used to study Sox2. It is important to recognize that other methods, such as directed evolution, have also demonstrated some success in improving the function of Sox2.[489] One study explored three residues in the DBD, which by randomizing led to 8,000 variants that were screened for reprogramming ability.[489] Several variants showed an increase in reprogramming ability. However, these methods depend on structural data to identify interaction sites or on molecular biology techniques to pinpoint mutational hotspots, due to the vast number of variants generated even with minor changes. For instance, randomizing merely four residues can result in as many as 160,000 variants, a formidable number for functional characterization.

Looking ahead, the future directions of this project will attempt to study Sox2 in a context that more closely mirrors its natural cellular environment. A key area of focus will be exploring the effects of post-translational modifications (PTMs) on the structural ensemble of Sox2 and how these modifications influence its binding affinities to DNA and nucleosomes. PTMs such as phosphorylation, methylation and acetylation can dramatically alter the behaviour of proteins[490-492], and in the case of Sox2, these modifications could provide some insights into its functional dynamics. Of special interest are the phosphorylation of serine 37 and threonine 116 located on either end of the structured DBD, with both of them being associated with an increase in transcriptional activity.[69] Another significant avenue of research will involve studying Sox2's involvement in phase separation. Given the growing evidence linking phase separation with transcriptional regulation[493], understanding how Sox2 participates in these biomolecular condensates could reveal Sox2's behaviour when transcriptionally active. A recent groundbreaking study showcased the effectiveness of smFRET and simulations in characterizing IDPs within biomolecular condensates[494], revealing that IDPs can maintain their dynamics even when part of these condensates. Exploring Sox2's role in phase separation, therefore, represents a promising research direction. Lastly, the specifics of how Sox2 interacts with partners like NANOG and Oct4, and the impact of these interactions on Sox2's dynamics and structural conformations, are still to be determined. Employing techniques such as intermolecular FRET and NMR could shed light on the residue specific contacts Sox2 makes, whether in its free form or when bound to DNA/nucleosomes, and how these contacts influence Sox2's dynamic behaviour. By continuing to leverage advanced techniques and integrated modelling, we aim to keep unravelling the mystery of Sox2's function and regulation at a molecular level.

# References


1 Christine, A. O., Annabel, E. T. & Janet, M. T. From protein structure to function. *Current Opinion in Structural Biology* **9**, 374-382 (1999).

2 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

3 Uversky, V. N. Introduction to Intrinsically Disordered Proteins (IDPs). *Chemical Reviews* **114**, 6557-6560 (2014).

4 Cast, K. *et al.* in *Biochemistry* Vol. 34 13211-13218 (1995).

5 David, A. M. *et al.* Increased Aggregation Tendency of Alpha-Synuclein in a Fully Disordered Protein Complex. *Journal of Molecular Biology* **431**, 2581-2598 (2019).

6 Wang, X. *et al.* Dynamic Autoinhibition of the HMGB1 Protein via Electrostatic Fuzzy Interactions of Intrinsically Disordered Regions. *Journal of Molecular Biology* **433**, 167122 (2021).

7 Clinckemalie, L., Vanderschueren, D., Boonen, S. & Claessens, F. The hinge region in androgen receptor control. *Molecular and Cellular Endocrinology* **358**, 1-8 (2012).

8 Ma, B., Tsai, C.-J., Haliloğlu, T. & Nussinov, R. Dynamic Allostery: Linkers Are Not Merely Flexible. *Structure* **19**, 907-917 (2011).

9 van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **114**, 6589-6631 (2014).

10 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

11 Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters* **579**, 3346-3354 (2005).

12 Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* (2023).

13 Shammas, S. L. Mechanistic roles of protein disorder within transcription. *Current Opinion in Structural Biology* **42**, 155-161 (2017).

14 Zhang, X., Bai, X.-c. & Chen, Z. J. Structures and Mechanisms in the cGAS-STING Innate Immunity Pathway. *Immunity* **53**, 43-53 (2020).

15 Cuylen, S. *et al.* Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature* **535**, 308-312 (2016).

16 Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16**, 18-29 (2015).

17 Dyson, H. J. & Wright, P. E. Equilibrium NMR studies of unfolded and partially folded proteins. *Nat Struct Biol* **5 Suppl**, 499-503 (1998).

18 Borgia, A. *et al.* in *Nature* Vol. 555 61-66 (Nature Publishing Group, 2018).

19 Demarest, S. J. *et al.* Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549-553 (2002).

20 Elbaum-Garfinkle, S. *et al.* The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences* **112**, 7189-7194 (2015).

21 Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology* **21**, 432-440 (2011).
22 Uversky, V. N. in *Advances in Protein Chemistry and Structural Biology* Vol. 110 (ed Rossen Donev) 85-121 (Academic Press, 2018).
23 Ruan, H., Sun, Q., Zhang, W., Liu, Y. & Lai, L. Targeting intrinsically disordered proteins at the edge of chaos. *Drug Discovery Today* **24**, 217-227 (2019).
24 Sammak, S. & Zinzalla, G. Targeting protein–protein interactions (PPIs) of transcription factors: Challenges of intrinsically disordered proteins (IDPs) and regions (IDRs). *Progress in Biophysics and Molecular Biology* **119**, 41-46 (2015).
25 Maxime, W., Xavier, D. & Ignacio, I. Geometry of the nucleus: a perspective on gene expression regulation. *Current Opinion in Chemical Biology* **20**, 112-119 (2014).
26 Hetzer, M. W. The Nuclear Envelope. *Cold Spring Harbor Perspectives in Biology* **2** (2010).
27 Dion, V. & Gasser, Susan M. Chromatin Movement in the Maintenance of Genome Stability. *Cell* **152**, 1355-1364 (2013).
28 Natsume, T. & Tanaka, T. U. Spatial regulation and organization of DNA replication within the nucleus. *Chromosome Research* **18**, 7-17 (2010).
29 Frege, T. & Uversky, V. N. Intrinsically disordered proteins in the nucleus of human cells. *Biochemistry and Biophysics Reports* **1**, 33-51 (2015).
30 Weng, J. & Wang, W. Dynamic multivalent interactions of intrinsically disordered proteins. *Current Opinion in Structural Biology* **62**, 9-13 (2020).
31 Teilum, K., Olsen, J. G. & Kragelund, B. B. On the specificity of protein–protein interactions in the context of disorder. *Biochemical Journal* **478**, 2035-2050 (2021).
32 Liu, Z. & Huang, Y. Advantages of proteins being disordered. *Protein Science* **23**, 539-550 (2014).
33 McGinty, R. K. & Tan, S. Nucleosome Structure and Function. *Chemical Reviews* **115**, 2255-2273 (2015).
34 Luger, K. Dynamic nucleosomes. *Chromosome Research* **14**, 5-16 (2006).
35 Zentner, G. E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology* **20**, 259 (2013).
36 Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications — cause and consequence of genome function. *Nature Reviews Genetics* **23**, 563-580 (2022).
37 Ghoneim, M., Fuchs, H. A. & Musselman, C. A. Histone Tail Conformations: A Fuzzy Affair with DNA. *Trends in Biochemical Sciences* **46**, 564-578 (2021).
38 Tsunaka, Y., Furukawa, A. & Nishimura, Y. Histone tail network and modulation in a nucleosome. *Current Opinion in Structural Biology* **75**, 102436 (2022).
39 Hergeth, S. P. & Schneider, R. The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO reports* **16**, 1439-1453-1453 (2015).
40 Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61-66 (2018).
41 Már, M., Nitsenko, K. & Heidarsson, P. O. Multifunctional Intrinsically Disordered Regions in Transcription Factors. *Chemistry – A European Journal* **29**, e202203369 (2023).
42 Badis, G. *et al.* Diversity and Complexity in DNA Recognition by Transcription Factors. *Science* **324**, 1720-1723 (2009).

43 Brodsky, S., Jana, T. & Barkai, N. Order through disorder: The role of intrinsically disordered regions in transcription factor binding specificity. *Current Opinion in Structural Biology* **71**, 110-115 (2021).
44 Zabet, N. R. & Adryan, B. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Research* **43**, 84-94 (2015).
45 Guertin, M. J. & Lis, J. T. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Current Opinion in Genetics & Development* **23**, 116-123 (2013).
46 Ramazi, S., Allahverdi, A. & Zahiri, J. Evaluation of post-translational modifications in histone proteins: A review on histone modification defects in developmental and neurological disorders. *Journal of Biosciences* **45**, 135 (2020).
47 Koutná, E. *et al.* Multivalency of nucleosome recognition by LEDGF. *Nucleic Acids Research* **51**, 10011-10025 (2023).
48 Swinstead, E. E., Paakinaho, V., Presman, D. M. & Hager, G. L. Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. *Bioessays* **38**, 1150-1157 (2016).
49 Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry* **78**, 273-304 (2009).
50 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes & Development* **25**, 2227-2241 (2011).
51 Ji, D. *et al.* FOXA1 forms biomolecular condensates that unpack condensed chromatin to function as a pioneer factor. *Molecular Cell*.
52 Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568 (2015).
53 Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460-1465 (2020).
54 Mayran, A. & Drouin, J. Pioneer transcription factors shape the epigenetic landscape. *Journal of Biological Chemistry* **293**, 13795-13804 (2018).
55 Zaret, K. S. & Mango, S. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr Opin Genet Dev* **37**, 76-81 (2016).
56 Henikoff, S. & Greally, J. M. Epigenetics, cellular memory and gene regulation. *Current Biology* **26**, R644-R648 (2016).
57 Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663-676 (2006).
58 Shi, G. & Jin, Y. Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Research & Therapy* **1**, 39 (2010).
59 Wegner, M. All purpose Sox: The many roles of Sox proteins in gene expression. *The International Journal of Biochemistry & Cell Biology* **42**, 381-390 (2010).
60 Kiefer, J. C. Back to basics: Sox genes. *Developmental Dynamics* **236**, 2356-2366 (2007).
61 Miyagi, S., Kato, H. & Okuda, A. Role of SoxB1 transcription factors in development. *Cellular and Molecular Life Sciences* **66**, 3675-3684 (2009).
62 Xue, B., Oldfield, C. J., Van, Y.-Y., Dunker, A. K. & Uversky, V. N. Protein intrinsic disorder and induced pluripotent stem cells. *Molecular BioSystems* **8**, 134-150 (2012).

63 Williams, C. A. C., Soufi, A. & Pollard, S. M. Post-translational modification of SOX family proteins: Key biochemical targets in cancer? *Seminars in Cancer Biology* **67**, 30-38 (2020).
64 Tsuruzoe, S. *et al.* Inhibition of DNA binding of Sox2 by the SUMO conjugation. *Biochem Biophys Res Commun* **351**, 920-926 (2006).
65 Novak, D. *et al.* SOX2 in development and cancer biology. *Seminars in Cancer Biology* **67**, 74-82 (2020).
66 Keramari, M. *et al.* Sox2 Is Essential for Formation of Trophectoderm in the Preimplantation Embryo. *PLOS ONE* **5**, e13952 (2010).
67 Pevny, L. H. & Nicolis, S. K. Sox2 roles in neural stem cells. *The International Journal of Biochemistry & Cell Biology* **42**, 421-424 (2010).
68 Vilas, J. M. *et al.* Transcriptional regulation of Sox2 by the retinoblastoma family of pocket proteins. *Oncotarget* **6**, 2992-3002 (2015).
69 Williams, C. A. C., Soufi, A. & Pollard, S. M. Post-translational modification of SOX family proteins: Key biochemical targets in cancer? *Semin Cancer Biol* **67**, 30-38 (2020).
70 Kim, D. K. *et al.* O-GlcNAcylation of Sox2 at threonine 258 regulates the self-renewal and early cell fate of embryonic stem cells. *Experimental & Molecular Medicine* **53**, 1759-1768 (2021).
71 Myers, S. A. *et al.* SOX2 O-GlcNAcylation alters its protein-protein interactions and genomic occupancy to modulate gene expression in pluripotent cells. *eLife* **5**, e10647 (2016).
72 Javaeed, A. & Ghauri, S. K. Metastatic potential and prognostic significance of SOX2: A meta-analysis. *World J Clin Oncol* **10**, 234-246 (2019).
73 Meng, Y., Xu, Q., Chen, L., Wang, L. & Hu, X. The function of SOX2 in breast cancer and relevant signaling pathway. *Pathology - Research and Practice* **216**, 153023 (2020).
74 Wuebben, E. L. & Rizzino, A. The dark side of SOX2: cancer - a comprehensive overview. *Oncotarget* **8**, 44917-44943 (2017).
75 Ding, L. N., Yu, Y. Y., Ma, C. J., Lei, C. J. & Zhang, H. B. SOX2-associated signaling pathways regulate biological phenotypes of cancers. *Biomedicine & Pharmacotherapy* **160**, 114336 (2023).
76 Hüser, L., Novak, D., Umansky, V., Altevogt, P. & Utikal, J. Targeting SOX2 in anticancer therapy. *Expert Opinion on Therapeutic Targets* **22**, 983-991 (2018).
77 Zhang, S., Xiong, X. & Sun, Y. Functional characterization of SOX2 as an anticancer target. *Signal Transduction and Targeted Therapy* **5**, 135 (2020).
78 Yan, C. & Higgins, P. J. Drugging the undruggable: Transcription therapy for cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1835**, 76-85 (2013).
79 Staby, L. *et al.* Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *Biochemical Journal* **474**, 2509-2532 (2017).
80 Yamaguchi, S., Hirano, K., Nagata, S. & Tada, T. Sox2 expression effects on direct reprogramming efficiency as determined by alternative somatic cell fate. *Stem Cell Research* **6**, 177-186 (2011).
81 Ye, P. *et al.* SOX family transcription factors as therapeutic targets in wound healing: A comprehensive review. *International Journal of Biological Macromolecules* **253**, 127243 (2023).
82 Chen, Y. *et al.* Reversible reprogramming of cardiomyocytes to a fetal state drives heart regeneration in mice. *Science* **373**, 1537-1540 (2021).

83 Wang, L. L. *et al.* The p53 Pathway Controls SOX2-Mediated Reprogramming in the Adult Mouse Spinal Cord. *Cell Rep* **17**, 891-903 (2016).
84 Yamoah, E. N. *et al.* Using Sox2 to alleviate the hallmarks of age-related hearing loss. *Ageing Research Reviews* **59**, 101042 (2020).
85 Vedel, I. M., Papagiannoula, A., Naudi-Fabra, S. & Milles, S. Nuclear magnetic resonance/single molecule fluorescence combinations to study dynamic protein systems. *Current Opinion in Structural Biology* **82**, 102659 (2023).
86 Aznauryan, M. *et al.* Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proceedings of the National Academy of Sciences* **113**, E5389-E5398 (2016).
87 Chemes, L. B., Alonso, L. G., Noval, M. G. & de Prat-Gay, G. in *Intrinsically Disordered Protein Analysis: Volume 1, Methods and Experimental Tools* (eds Vladimir N. Uversky & A. Keith Dunker) 387-404 (Humana Press, 2012).
88 Heidarsson, P. O. *et al.* Release of linker histone from the nucleosome driven by polyelectrolyte competition with a disordered protein. *Nature Chemistry* **14**, 224-231 (2022).
89 Metskas, L. A. & Rhoades, E. Single-Molecule FRET of Intrinsically Disordered Proteins. *Annual Review of Physical Chemistry* **71**, 391-414 (2020).
90 Dyson, H. J. & Wright, P. E. NMR illuminates intrinsic disorder. *Current Opinion in Structural Biology* **70**, 44-52 (2021).
91 Schneider, R., Blackledge, M. & Jensen, M. R. Elucidating binding mechanisms and dynamics of intrinsically disordered protein complexes using NMR spectroscopy. *Current Opinion in Structural Biology* **54**, 10-18 (2019).
92 Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters* **589**, 2570-2577 (2015).
93 Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Current Opinion in Structural Biology* **42**, 147-154 (2017).
94 Holmstrom, E. D. *et al.* in *Methods in Enzymology* Vol. 611 (ed Elizabeth Rhoades) 287-325 (Academic Press, 2018).
95 Nath, A. *et al.* The conformational ensembles of α-synuclein and tau: combining single-molecule FRET and simulations. *Biophys J* **103**, 1940-1949 (2012).
96 Bacic, L., Sabantsev, A. & Deindl, S. Recent advances in single-molecule fluorescence microscopy render structural biology dynamic. *Current Opinion in Structural Biology* **65**, 61-68 (2020).
97 Ray, P. C., Fan, Z., Crouch, R. A., Sinha, S. S. & Pramanik, A. Nanoscopic optical rulers beyond the FRET distance limit: fundamentals and applications. *Chemical Society Reviews* **43**, 6370-6404 (2014).
98 Kremers, G.-J., Goedhart, J., van Munster, E. B. & Gadella, T. W. J. Cyan and Yellow Super Fluorescent Proteins with Improved Brightness, Protein Folding, and FRET Förster Radius. *Biochemistry* **45**, 6570-6580 (2006).
99 Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nature Methods* **5**, 507-516 (2008).
100 Schuler, B. Single-molecule FRET of protein structure and dynamics - a primer. *J Nanobiotechnology* **11 Suppl 1**, S2-S2 (2013).
101 Schuler, B. & Hofmann, H. Single-molecule spectroscopy of protein folding dynamics—expanding scope and timescales. *Current Opinion in Structural Biology* **23**, 36-47 (2013).

102 Choi, U. B. *et al.* Modulating the Intrinsic Disorder in the Cytoplasmic Domain Alters the Biological Activity of the N-Methyl-d-aspartate-sensitive Glutamate Receptor *. *Journal of Biological Chemistry* **288**, 22506-22515 (2013).
103 Bjarnason, S., Ruidiaz, S. F., McIvor, J., Mercadante, D. & Heidarsson, P. O. in *Progress in Molecular Biology and Translational Science* Vol. 183 (ed Vladimir N. Uversky) 295-354 (Academic Press, 2021).
104 Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**, 321-331 (1999).
105 Schad, E., Tompa, P. & Hegyi, H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol* **12**, R120 (2011).
106 van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589-6631 (2014).
107 Schuler, B. *et al.* Binding without folding - the biomolecular function of disordered polyelectrolyte complexes. *Curr Opin Struct Biol* **60**, 66-76 (2020).
108 Theillet, F. X. *et al.* Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chemical Reviews* **114**, 6661-6714 (2014).
109 Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder. *Chem Rev* **114**, 6561-6588 (2014).
110 Hsu, W. L. *et al.* Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci* **22**, 258-273 (2013).
111 Kulkarni, P. & Uversky, V. N. Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome. *Proteomics* **18**, e1800061 (2018).
112 McGinty, R. & Tan, S. Nucleosome structure and function. *Chemical Reviews* **115**, 2255-2273 (2014).
113 Bednar, J. *et al.* Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. *Mol Cell* **66**, 384-397 e388 (2017).
114 Peng, Z., Mizianty, M. J., Xue, B., Kurgan, L. & Uversky, V. N. More than just tails: intrinsic disorder in histone proteins. *Mol Biosyst* **8**, 1886-1901 (2012).
115 Fyodorov, D. V., Zhou, B.-r., Skoultchi, A. I. & Bai, Y. Emerging roles of linker histones in regulating chromatin structure and function. *Nature Publishing Group* **19**, 192-206 (2017).
116 Martire, S. & Banaszynski, L. A. The roles of histone variants in fine-tuning chromatin organization and function. *Nat Rev Mol Cell Biol* **21**, 522-541 (2020).
117 Lever, M. A., Th'ng, J. P., Sun, X. & Hendzel, M. J. Rapid exchange of histone H1.1 on chromatin in living human cells. *Nature* **408**, 873-876 (2000).
118 Heidarsson, P. O. *et al.* Disordered Proteins Enable Histone Chaperoning on the Nucleosome. *bioRxiv*, 2020.2004.2017.046243 (2020).
119 Kale, S., Goncearenco, A., Markov, Y., Landsman, D. & Panchenko, A. R. Molecular recognition of nucleosomes by binding partners. *Curr Opin Struct Biol* **56**, 164-170 (2019).
120 Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J., Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl* **44**, 7342-7372 (2005).
121 Bowman, G. D. & Poirier, M. G. Post-translational modifications of histones that influence nucleosome dynamics. *Chem Rev* **115**, 2274-2295 (2015).
122 Darling, A. L. & Uversky, V. N. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Front Genet* **9**, 158 (2018).

123 Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336-1340 (2018).
124 Liu, J. *et al.* Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873-6888 (2006).
125 Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends in biochemical sciences* **33**, 2-8 (2008).
126 Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**, 13392-13397 (2013).
127 Cuvier, O. & Fierz, B. Dynamic chromatin technologies: from individual molecules to epigenomic regulation in cells. *Nat Rev Genet* **18**, 457-472 (2017).
128 Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-1861 (2007).
129 Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260 (1997).
130 Adhireksan, Z., Sharma, D., Lee, P. L. & Davey, C. A. Near-atomic resolution structures of interdigitated nucleosome fibres. *Nat Commun* **11**, 4747 (2020).
131 Song, F. *et al.* Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* **344**, 376-380 (2014).
132 Yang, C., Van Der Woerd, M. J., Muthurajan, U. M., Hansen, J. C. & Luger, K. Biophysical analysis and small-angle X-ray scattering-derived structures of MeCP2-nucleosome complexes. *Nucleic Acids Research* **39**, 4122-4135 (2011).
133 Abramov, G., Velyvis, A., Rennella, E., Wong, L. E. & Kay, L. E. A methyl-TROSY approach for NMR studies of high-molecular-weight DNA with application to the nucleosome core particle. *Proc Natl Acad Sci U S A* **117**, 12836-12846 (2020).
134 Jain, S. S. & Tullius, T. D. Footprinting protein-DNA complexes using the hydroxyl radical. *Nat Protoc* **3**, 1092-1100 (2008).
135 Gontier, A. *et al.* Measurements of Protein-DNA Complexes Interactions by Isothermal Titration Calorimetry (ITC) and Microscale Thermophoresis (MST). *Methods Mol Biol* **2247**, 125-143 (2021).
136 Nakato, R. & Sakata, T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods* **187**, 44-53 (2021).
137 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
138 Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152-156 (2020).
139 Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159**, 995-1014 (2014).
140 Wu, M. & Lander, G. C. Present and Emerging Methodologies in Cryo-EM Single-Particle Analysis. *Biophys J* **119**, 1281-1289 (2020).
141 Loveland, A. B., Demo, G. & Korostelev, A. A. Cryo-EM of elongating ribosome with EF-Tu*GTP elucidates tRNA proofreading. *Nature* **584**, 640-645 (2020).
142 Garcia-Saez, I. *et al.* Structure of an H1-Bound 6-Nucleosome Array Reveals an Untwisted Two-Start Chromatin Fiber Conformation. *Mol Cell* **72**, 902-915 e907 (2018).

143 Sottini, A. *et al.* Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat Commun* **11**, 5736 (2020).
144 Milles, S. *et al.* An ultraweak interaction in the intrinsically disordered replication machinery is essential for measles virus function. *Sci Adv* **4**, eaat7778 (2018).
145 Theillet, F. X. *et al.* Structural disorder of monomeric alpha-synuclein persists in mammalian cells. *Nature* **530**, 45-50 (2016).
146 Kay, L. E. NMR studies of protein structure and dynamics - a look backwards and forwards. *J Magn Reson* **213**, 492-494 (2011).
147 van Emmerik, C. L. & van Ingen, H. Unspinning chromatin: Revealing the dynamic nucleosome landscape by NMR. *Prog Nucl Magn Reson Spectrosc* **110**, 1-19 (2019).
148 Zhou, B. R. *et al.* Structural Mechanisms of Nucleosome Recognition by Linker Histones. *Mol Cell* **59**, 628-638 (2015).
149 Bernado, P. & Svergun, D. I. Analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Methods Mol Biol* **896**, 107-122 (2012).
150 Henriques, J., Arleth, L., Lindorff-Larsen, K. & Skepo, M. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J Mol Biol* **430**, 2521-2539 (2018).
151 Misteli, T., Gunjan, A., Hock, R., Bustin, M. & Brown, D. T. Dynamic binding of histone H1 to chromatin in living cells. *Nature* **408**, 877-881 (2000).
152 Bharath, M. S., Chandra, N. R. & Rao, M. Molecular modeling of the chromatosome particle. *Nucleic acids research* **31**, 4264-4274 (2003).
153 Wong, H., Victor, J.-M. & Mozziconacci, J. An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PloS one* **2**, e877 (2007).
154 Kepper, N., Foethke, D., Stehr, R., Wedemann, G. & Rippe, K. Nucleosome geometry and internucleosomal interactions control the chromatin fiber conformation. *Biophysical journal* **95**, 3692-3705 (2008).
155 Stehr, R., Kepper, N., Rippe, K. & Wedemann, G. The effect of internucleosomal interaction on folding of the chromatin fiber. *Biophysical journal* **95**, 3677-3691 (2008).
156 Öztürk, M. A., Cojocaru, V. & Wade, R. C. Dependence of chromatosome structure on linker histone sequence and posttranslational modification. *Biophysical journal* **114**, 2363-2375 (2018).
157 Brucale, M., Schuler, B. & Samori, B. Single-molecule studies of intrinsically disordered proteins. *Chem Rev* **114**, 3281-3317 (2014).
158 Asher, W. B. *et al.* Single-molecule FRET imaging of GPCR dimers in living cells. *Nat Methods* **18**, 397-405 (2021).
159 Lerner, E. *et al.* FRET-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices. *Elife* **10** (2021).
160 Schuler, B. & Hofmann, H. Single-molecule spectroscopy of protein folding dynamics--expanding scope and timescales. *Current opinion in structural biology* **23**, 36-47 (2013).
161 Gunther, E. *et al.* The in vivo mechanics of the magnetotactic backbone as revealed by correlative FLIM-FRET and STED microscopy. *Sci Rep* **9**, 19615 (2019).
162 Lerner, E. *et al.* Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science* **359**, eaan1133 (2018).
163 Heidarsson, P. O. & Cecconi, C. From folding to function: complex macromolecular reactions unraveled one-by-one with optical tweezers. *Essays Biochem* **65**, 129-142 (2021).

164 Rudnizky, S., Khamis, H., Malik, O., Melamed, P. & Kaplan, A. The base pair-scale diffusion of nucleosomes modulates binding of transcription factors. *Proc Natl Acad Sci U S A* **116**, 12161-12166 (2019).
165 Gomes, G. W. *et al.* Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc* **142**, 15697-15710 (2020).
166 Vendruscolo, M. & Dobson, C. M. Protein dynamics: Moore's law in molecular biology. *Current Biology* **21**, R68-R70 (2011).
167 Fan, L. & Roberts, V. A. Complex of linker histone H5 with the nucleosome and its implications for chromatin packing. *Proceedings of the National Academy of Sciences* **103**, 8384-8389 (2006).
168 Collepardo-Guevara, R. & Schlick, T. Chromatin fiber polymorphism triggered by variations of DNA linker lengths. *Proceedings of the National Academy of Sciences* **111**, 8061-8066 (2014).
169 Teif, V. B., Kepper, N., Yserentant, K., Wedemann, G. & Rippe, K. Affinity, stoichiometry and cooperativity of heterochromatin protein 1 (HP1) binding to nucleosomal arrays. *Journal of Physics: Condensed Matter* **27**, 064110 (2015).
170 Watanabe, S., Mishima, Y., Shimizu, M., Suetake, I. & Takada, S. Interactions of HP1 Bound to H3K9me3 Dinucleosome by Molecular Simulations and Biochemical Assays. *Biophys J* **114**, 2336-2351 (2018).
171 Alston, J. J., Soranno, A. & Holehouse, A. S. Integrating single-molecule spectroscopy and simulations for the study of intrinsically disordered proteins. *Methods* (2021).
172 Sanders, J. C. & Holmstrom, E. D. Integrating single-molecule FRET and biomolecular simulations to study diverse interactions between nucleic acids and proteins. *Essays Biochem* **65**, 37-49 (2021).
173 Rhodes, D. & Robinson, P. J. J. Structure of the '30 nm' chromatin fibre: A key role for the linker histone. *Current Opinion in Structural Biology* **16**, 336-343 (2006).
174 Hergeth, S. P. & Schneider, R. The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO reports* **16**, 1439-1453 (2015).
175 Hansen, J. C., Lu, X., Ross, E. D. & Woody, R. W. Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains*. *Journal of Biological Chemistry* **281**, 1853-1856 (2006).
176 Goytisolo, F. A. *et al.* Identification of two DNA-binding sites on the globular domain of histone H5. *EMBO Journal* **15**, 3421-3429 (1996).
177 Gibbs, E. B. & Kriwacki, R. W. Linker histones as liquid-like glue for chromatin. *Proc Natl Acad Sci U S A* **115**, 11868-11870 (2018).
178 Izzo, A., Kamieniarz, K. & Schneider, R. The histone H1 family: specific members, specific functions? *Biological chemistry* **389**, 333-343 (2008).
179 Th'ng, J. P., Sung, R., Ye, M. & Hendzel, M. J. H1 family histones in the nucleus: control of binding and localization by the C-terminal domain. *Journal of Biological Chemistry* **280**, 27809-27814 (2005).
180 Lorch, Y., LaPointe, J. W. & Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203-210 (1987).
181 Roque, A., Ponte, I. & Suau, P. Post-translational modifications of the intrinsically disordered terminal domains of histone H1: effects on secondary structure and chromatin dynamics. *Chromosoma* **126**, 83-91 (2017).

182 Talasz, H., Helliger, W., Puschendorf, B. & Lindner, H. In vivo phosphorylation of histone H1 variants during the cell cycle. *Biochemistry* **35**, 1761-1767 (1996).
183 Hergeth, S. P. *et al.* Isoform-specific phosphorylation of human linker histone H1. 4 in mitosis by the kinase Aurora B. *Journal of cell science* **124**, 1623-1628 (2011).
184 Chu, C.-S. *et al.* Protein kinase A-mediated serine 35 phosphorylation dissociates histone H1. 4 from mitotic chromosome. *Journal of Biological Chemistry* **286**, 35843-35851 (2011).
185 Clausell, J., Happel, N., Hale, T. K., Doenecke, D. & Beato, M. Histone H1 subtypes differentially modulate chromatin condensation without preventing ATP-dependent remodeling by SWI/SNF or NURF. *PloS one* **4**, e0007243 (2009).
186 Daujat, S., Zeissler, U., Waldmann, T., Happel, N. & Schneider, R. HP1 binds specifically to Lys26-methylated histone H1.4, whereas simultaneous Ser27 phosphorylation blocks HP1 binding. *J Biol Chem* **280**, 38090-38095 (2005).
187 Alexandrow, M. G. & Hamlin, J. L. Chromatin decondensation in S-phase involves recruitment of Cdk2 by Cdc45 and histone H1 phosphorylation. *J Cell Biol* **168**, 875-886 (2005).
188 Meergans, T., Albig, W. & Doenecke, D. Varied expression patterns of human H1 histone genes in different cell lines. *DNA Cell Biol* **16**, 1041-1049 (1997).
189 Weiss, T. *et al.* Histone H1 variant-specific lysine methylation by G9a/KMT1C and Glp1/KMT1D. *Epigenetics & chromatin* **3**, 1-13 (2010).
190 Vaquero, A. *et al.* Human SirT1 interacts with histone H1 and promotes formation of facultative heterochromatin. *Molecular cell* **16**, 93-105 (2004).
191 Terme, J.-M. *et al.* Dynamics and dispensability of variant-specific histone H1 Lys-26/Ser-27 and Thr-165 post-translational modifications. *FEBS letters* **588**, 2353-2362 (2014).
192 Kamieniarz, K. *et al.* A dual role of linker histone H1. 4 Lys 34 acetylation in transcriptional activation. *Genes & development* **26**, 797-802 (2012).
193 Arents, G., Burlingame, R. W., Wang, B.-C., Love, W. E. & Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proceedings of the National Academy of Sciences* **88**, 10148-10152 (1991).
194 Kurumizaka, H., Kujirai, T. & Takizawa, Y. Contributions of Histone Variants in Nucleosome Structure and Function. *J Mol Biol* **433**, 166678 (2021).
195 Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of molecular biology* **319**, 1097-1113 (2002).
196 Zentner, G. E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol* **20**, 259-266 (2013).
197 Allan, J., Harborne, N., Rau, D. C. & Gould, H. Participation of core histone" tails" in the stabilization of the chromatin solenoid. *The Journal of cell biology* **93**, 285-297 (1982).
198 Zheng, C. & Hayes, J. J. Structures and interactions of the core histone tail domains. *Biopolymers: Original Research on Biomolecules* **68**, 539-546 (2003).
199 Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nat Struct Mol Biol* **12**, 46-53 (2005).
200 Andresen, K., Jimenez-Useche, I., Howell, S. C., Yuan, C. & Qiu, X. Solution scattering and FRET studies on nucleosomes reveal DNA unwrapping effects of H3 and H4 tail removal. *PLoS One* **8**, e78587 (2013).

201 Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *Journal of molecular biology* **327**, 85-96 (2003).
202 Shogren-Knaak, M. *et al.* Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311**, 844-847 (2006).
203 Brower-Toland, B. *et al.* Specific contributions of histone tails and their acetylation to the mechanical stability of nucleosomes. *Journal of molecular biology* **346**, 135-146 (2005).
204 Anderson, J., Lowary, P. & Widom, J. Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites. *Journal of molecular biology* **307**, 977-985 (2001).
205 Kurdistani, S. K., Tavazoie, S. & Grunstein, M. Mapping global histone acetylation patterns to gene expression. *Cell* **117**, 721-733 (2004).
206 Martin, B. J. E. *et al.* Transcription shapes genome-wide histone acetylation patterns. *Nat Commun* **12**, 210 (2021).
207 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124 (2001).
208 Sugiyama, K. *et al.* Aurora-B associated protein phosphatases as negative regulators of kinase activation. *Oncogene* **21**, 3103-3111 (2002).
209 Goto, H., Yasui, Y., Nigg, E. A. & Inagaki, M. Aurora-B phosphorylates Histone H3 at serine28 with regard to the mitotic chromosome condensation. *Genes to cells* **7**, 11-17 (2002).
210 Rossetto, D., Avvakumov, N. & Côté, J. Histone phosphorylation: a chromatin modification involved in diverse nuclear events. *Epigenetics* **7**, 1098-1108 (2012).
211 Tatsuka, M. *et al.* Multinuclearity and increased ploidy caused by overexpression of the aurora-and Ipl1-like midbody-associated protein mitotic kinase in human cancer cells. *Cancer research* **58**, 4811-4816 (1998).
212 Wang, H. *et al.* Methylation of histone H4 at arginine 3 facilitating transcriptional activation by nuclear hormone receptor. *Science* **293**, 853-857 (2001).
213 Wysocka, J. *et al.* WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **121**, 859-872 (2005).
214 Lomberk, G., Bensi, D., Fernandez-Zapico, M. E. & Urrutia, R. Evidence for the existence of an HP1-mediated subcode within the histone code. *Nature cell biology* **8**, 407-415 (2006).
215 Azzaz, A. M. *et al.* Human heterochromatin protein 1α promotes nucleosome associations that drive chromatin condensation. *Journal of Biological Chemistry* **289**, 6850-6861 (2014).
216 Kilic, S., Bachmann, A. L., Bryan, L. C. & Fierz, B. Multivalency governs HP1alpha association dynamics with the silent chromatin state. *Nat Commun* **6**, 7313 (2015).
217 Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080-2083 (2002).
218 Hiragami-Hamada, K. *et al.* N-Terminal Phosphorylation of HP1α Promotes Its Chromatin Binding. *Mol Cell Biol* **31**, 1186-1200 (2011).
219 Bryan, L. C. *et al.* Single-molecule kinetic analysis of HP1-chromatin binding reveals a dynamic network of histone modification and DNA interactions. *Nucleic acids research* **45**, 10504-10517 (2017).
220 Hayakawa, T., Haraguchi, T., Masumoto, H. & Hiraoka, Y. Cell cycle behavior of human HP1 subtypes: distinct molecular domains of HP1 are required for their

centromeric localization during interphase and metaphase. *Journal of cell science* **116**, 3327-3338 (2003).
221 Bosch-Presegué, L. *et al.* Mammalian HP1 isoforms have specific roles in heterochromatin structure and organization. *Cell reports* **21**, 2048-2057 (2017).
222 Canzio, D., Larson, A. & Narlikar, G. J. Mechanisms of functional promiscuity by HP1 proteins. *Trends in cell biology* **24**, 377-386 (2014).
223 Mishima, Y. *et al.* Hinge and chromoshadow of HP1α participate in recognition of K9 methylated histone H3 in nucleosomes. *Journal of molecular biology* **425**, 54-70 (2013).
224 Muchardt, C. *et al.* Coordinated methyl and RNA binding is required for heterochromatin localization of mammalian HP1α. *EMBO reports* **3**, 975-981 (2002).
225 Badugu, R., Yoo, Y., Singh, P. B. & Kellum, R. Mutations in the heterochromatin protein 1 (HP1) hinge domain affect HP1 protein interactions and chromosomal distribution. *Chromosoma* **113**, 370-384 (2005).
226 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-120 (2001).
227 LeRoy, G. *et al.* Heterochromatin protein 1 is extensively decorated with histone code-like post-translational modifications. *Molecular & Cellular Proteomics* **8**, 2432-2442 (2009).
228 Nishibuchi, G. *et al.* N-terminal phosphorylation of HP1α increases its nucleosome-binding specificity. *Nucleic acids research* **42**, 12498-12511 (2014).
229 Shimojo, H. *et al.* Extended string-like binding of the phosphorylated HP1alpha N-terminal tail to the lysine 9-methylated histone H3 tail. *Sci Rep* **6**, 22527 (2016).
230 Liang, J. F., Yang, V. C. & Vaynshteyn, Y. The minimal functional sequence of protamine. *Biochemical and Biophysical Research Communications* **336**, 653-659 (2005).
231 Ward, W. S. & Coffey, D. S. DNA packaging and organization in mammalian spermatozoa: Comparison with somatic cells. *Biology of Reproduction* **44**, 569-574 (1991).
232 D'Ippolito, R. A. *et al.* Protamines from liverwort are produced by post-translational cleavage and C-terminal di-aminopropanelation of several male germ-specific H1 histones. *Journal of Biological Chemistry* **294**, 16364-16373 (2019).
233 Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochim Biophys Acta* **1839**, 155-168 (2014).
234 Domenjoud, L., Kremling, H., Burfeind, P., Maier, W. M. & Engel, W. On the expression of protamine genes in the testis of man and other mammals. *Andrologia* **23**, 333-337 (1991).
235 Carrell, D. T., Emery, B. R. & Hammoud, S. Altered protamine expression and diminished spermatogenesis: what is the link? *Hum Reprod Update* **13**, 313-327 (2007).
236 Balhorn, R. The protamine family of sperm nuclear proteins. *Genome Biol* **8**, 227 (2007).
237 Bode, J., Willmitzer, L. & Opatz, K. On the competition between protamines and histones: studies directed towards the understanding of spermiogenesis. *Eur J Biochem* **72**, 393-403 (1977).
238 Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins.

*Proceedings of the National Academy of Sciences of the United States of America* **107**, 8183-8188 (2010).
239 Balhorn, R. *et al.* Protamine mediated condensation of DNA in mammalian sperm. *Male Gamete*, 55-70 (1999).
240 Bench, G. S., Friz, A. M., Corzett, M. H., Morse, D. H. & Balhorn, R. DNA and total protamine masses in individual sperm from fertile semen of selected mammals. *Cytometry* **23**, 263-271 (1996).
241 Balhorn, R., Corzett, M., Mazrimas, J. & Watkins, B. Identification of Bull Protamine Disulfides. *Biochemistry* **30**, 175-181 (1991).
242 Vilfan, I. D., Conwell, C. C. & Hud, N. V. Formation of Native-like Mammalian Sperm Cell Chromatin with Folded Bull Protamine. *Journal of Biological Chemistry* **279**, 20088-20095 (2004).
243 Ukogu, O. A. *et al.* Protamine loops DNA in multiple steps. *Nucleic Acids Res* **48**, 6108-6119 (2020).
244 Hsiang, M. W. & Cole, R. D. Structure of histone H1-DNA complex: effect of histone H1 on DNA condensation. *Proc Natl Acad Sci U S A* **74**, 4852-4856 (1977).
245 Brunner, A., Nanni, P. & Mansuy, I. Epigenetic marking of sperm by post-translational modification of histones and protamines. *Epigenetics & Chromatin* **7**, 2 (2014).
246 Bustin, M. High Mobility Group Proteins. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1799**, 1-2 (2010).
247 Nicolas, R. H. & Goodwin, G. H. in *The Chromosomal Proteins* (ed E W Johns) 41-68 (Academic Press, 1982).
248 Bustin, M. Chromatin unfolding and activation by HMGN* chromosomal protiens. *Trends in Biochemical Sciences* **26**, 431-437 (2001).
249 Phair, R. D. & Misteli, T. High mobility of proteins in the mammalian cell nucleus. *Nature* **404**, 604-609 (2000).
250 Alfonso, P. J., Crippa, M. P., Hayes, J. J. & Bustin, M. The footprint of chromosomal proteins HMG-14 and HMG-17 on chromatin subunits. *J Mol Biol* **236**, 189-198 (1994).
251 Postnikov, Y. V., Trieschmann, L., Rickers, A. & Bustin, M. Homodimers of chromosomal proteins HMG-14 and HMG-17 in nucleosome cores. *Journal of Molecular Biology* **252**, 423-432 (1995).
252 Ueda, T., Postnikov, Y. V. & Bustin, M. Distinct domains in high mobility group N variants modulate specific chromatin modifications. *J Biol Chem* **281**, 10182-10187 (2006).
253 Postnikov, Y. V. & Bustin, M. Functional interplay between histone H1 and HMG proteins in chromatin. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1859**, 462-467 (2016).
254 Ueda, T., Catez, F., Gerlitz, G. & Bustin, M. Delineation of the Protein Module That Anchors HMGN Proteins to Nucleosomes in the Chromatin of Living Cells. *Mol Cell Biol* **28**, 2872-2883 (2008).
255 Trieschmann, L., Martin, B. & Bustin, M. The chromatin unfolding domain of chromosomal protein HMG-14 targets the N-terminal tail of histone H3 in nucleosomes. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5468-5473 (1998).
256 Kato, H. *et al.* Architecture of the high mobility group nucleosomal protein 2-nucleosome complex as revealed by methyl-based NMR. *Proceedings of the*

*National Academy of Sciences of the United States of America* **108**, 12283-12288 (2011).
257 Lim, J. H. *et al.* Chromosomal protein HMGN1 modulates histone H3 phosphorylation. *Molecular Cell* **15**, 573-584 (2004).
258 Zhang, Q. & Wang, Y. High mobility group proteins and their post-translational modifications. *Biochim Biophys Acta* **1784**, 1159-1166 (2008).
259 Bergel, M. *et al.* Acetylation of novel sites in the nucleosomal binding domain of chromosomal protein HMG-14 by p300 alters its interaction with nucleosomes. *J Biol Chem* **275**, 11514-11520 (2000).
260 Murphy, K. J. *et al.* HMGN1 and 2 remodel core and linker histone tail domains within chromatin. *Nucleic Acids Research* **45**, 9917-9930 (2017).
261 Albright, S. C., Wiseman, J. M., Lange, R. A. & Garrard, W. T. Subunit structures of different electrophoretic forms of nucleosomes. *J Biol Chem* **255**, 3673-3684 (1980).
262 van Houte, L. P. *et al.* Solution structure of the sequence-specific HMG box of the lymphocyte transcriptional activator Sox-4. *J Biol Chem* **270**, 30516-30524 (1995).
263 Sheflin, L. G. & Spaulding, S. W. High Mobility Group Protein 1 Preferentially Conserves Torsion in Negatively Supercoiled DNA. *Biochemistry* **28**, 5658-5664 (1989).
264 Štros, M., Štokrová, J. & Thomas, J. O. Dna looping by the HMG-box domains of HMG1 and modulation of DNA binding by the acidic C-terminal domain. *Nucleic Acids Research* **22**, 1044-1051 (1994).
265 Travers, A. Recognition of distorted DNA structures by HMG domains. *Current Opinion in Structural Biology* **10**, 102-109 (2000).
266 Thomas, J. O. & Travers, A. A. HMG1 and 2, and related 'architectural' DNA-binding proteins. *Trends in Biochemical Sciences* **26**, 167-174 (2001).
267 Jaouen, S. *et al.* Determinants of specific binding of HMGB1 protein to hemicatenated DNA loops. *Journal of Molecular Biology* **353**, 822-837 (2005).
268 Ugrinova, I., Pasheva, E. A., Armengaud, J. & Pashev, I. G. In vivo acetylation of HMG1 protein enhances its binding affinity to distorted DNA structures. *Biochemistry* **40**, 14655-14660 (2001).
269 Knapp, S. *et al.* The long acidic tail of high mobility group box 1 (HMGB1) protein forms an extended and flexible structure that interacts with specific residues within and between the HMG boxes. *Biochemistry* **43**, 11992-11997 (2004).
270 Stott, K., Watson, M., Howe, F. S., Grossmann, J. G. & Thomas, J. O. Tail-mediated collapse of HMGB1 is dynamic and occurs via differential binding of the acidic tail to the A and B domains. *Journal of Molecular Biology* **403**, 706-722 (2010).
271 Kawase, T., Sato, K., Ueda, T. & Yoshida, M. Distinct domains in HMGB1 Are involved in specific intramolecular and nucleosomal interactions. *Biochemistry* **47**, 13991-13996 (2008).
272 An, W., Van Holde, K. & Zlatanova, J. The non-histone chromatin protein HMG1 protects linker DNA on the side opposite to that protected by linker histones. *Journal of Biological Chemistry* **273**, 26289-26291 (1998).
273 Jackson, J. B. & Rill, R. L. Circular Dichroism, Thermal Denaturation, and Deoxyribonuclease I Digestion Studies of Nucleosomes Highly Enriched in High Mobility Group Proteins HMG 1 and HMG 2. *Biochemistry* **20**, 1042-1046 (1981).
274 Catez, F. *et al.* Network of Dynamic Interactions between Histone H1 and High-Mobility-Group Proteins in Chromatin. *Mol Cell Biol* **24**, 4321-4328 (2004).

275 Cato, L., Stott, K., Watson, M. & Thomas, J. O. The Interaction of HMGB1 and Linker Histones Occurs Through their Acidic and Basic Tails. *Journal of Molecular Biology* **384**, 1262-1272 (2008).
276 Štros, M. HMGB proteins: Interactions with DNA and chromatin. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1799**, 101-113 (2010).
277 Polanska, E., Pospisilova, S. & Stros, M. Binding of histone H1 to DNA is differentially modulated by redox state of HMGB1. *PLoS One* **9**, e89070 (2014).
278 Lehn, D. A., Elton, T. S., Johnson, K. R. & Reeves, R. A conformational study of the sequence specific binding of HMG-I (Y) with the bovine interleukin-2 cDNA. *Biochemistry international* **16**, 963-971 (1988).
279 Evans, J. N. S. *et al.* 1H and 13C NMR assignments and molecular modelling of a minor groove DNA-binding peptide from the HMG-I protein. *International Journal of Peptide and Protein Research* **45**, 554-560 (1995).
280 Huth, J. R. *et al.* The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. *Nature Structural Biology* **4**, 657-665 (1997).
281 Cui, T. & Leng, F. Specific recognition of AT-rich DNA sequences by the mammalian high mobility group protein AT-hook 2: A SELEX study. *Biochemistry* **46**, 13059-13066 (2007).
282 Reeves, R. & Nissen, M. S. Interaction of high mobility group-I(Y) nonhistone proteins with nucleosome core particles. *Journal of Biological Chemistry* **268**, 21137-21146 (1993).
283 Li, O., Vasudevan, D., Davey, C. A. & Dröge, P. High-level expression of DNA architectural factor HMGA2 and its association with nucleosomes in human embryonic stem cells. *Genesis* **44**, 523-529 (2006).
284 Saitoh, Y. & Laemmli, U. K. Metaphase chromosome structure: Bands arise from a differential folding path of the highly AT-rich scaffold. *Cell* **76**, 609-622 (1994).
285 Zhao, K., Kas, E., Gonzalez, E. & Laemmli, U. K. SAR-dependent mobilization of histone H1 by HMG-I/Y in vitro: HMG-I/Y is enriched in H1-depleted chromatin. *EMBO Journal* **12**, 3237-3247 (1993).
286 Kishi, Y., Fujii, Y., Hirabayashi, Y. & Gotoh, Y. HMGA regulates the global chromatin state and neurogenic potential in neocortical precursor cells. *Nature Neuroscience* **15**, 1127-1133 (2012).
287 Wei, T. *et al.* Phosphorylation-regulated HMGA1a-P53 interaction unveils the function of HMGA1a acidic tail phosphorylations via synthetic proteins. *Cell Chem Biol* (2021).
288 Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev* **28**, 2679-2692 (2014).
289 Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568 (2015).
290 Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420 (1993).
291 Cirillo, L. A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO Journal* **17**, 244-254 (1998).
292 Chaya, D., Hayamizu, T., Bustin, M. & Zaret, K. S. Transcription Factor FoxA (HNF3) on a Nucleosome at an Enhancer Complex in Liver Chromatin. *Journal of Biological Chemistry* **276**, 44385-44389 (2001).

293 Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular Cell* **9**, 279-289 (2002).
294 Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Molecular Cell* **62**, 79-91 (2016).
295 Iwafuchi, M. *et al.* Gene network transitions in embryos depend upon interactions between a pioneer transcription factor and core histones. *Nat Genet* **52**, 418-427 (2020).
296 Taube, J. H., Allton, K., Duncan, S. A., Shen, L. & Barton, M. C. Foxa1 functions as a pioneer transcription factor at transposable elements to activate Afp during differentiation of embryonic stem cells. *Journal of Biological Chemistry* **285**, 16135-16144 (2010).
297 Gómez-Márquez, J. & Rodríguez, P. Prothymosin α is a chromatin-remodelling protein in mammalian cells. *Biochemical Journal* **333**, 1-3 (1998).
298 George, E. M. & Brown, D. T. Prothymosin α is a component of a linker histone chaperone. *FEBS Letters* **584**, 2833-2836 (2010).
299 Mosoian, A. Intracellular and extracellular cytokine-like functions of prothymosin α: Implications for the development of immunotherapies. *Future Medicinal Chemistry* **3**, 1199-1208 (2011).
300 Cast, K. *et al.* Prothymosin α: A Biologically Active Protein with Random Coil Conformation. *Biochemistry* **34**, 13211-13218 (1995).
301 Uversky, V. N. *et al.* Natively unfolded human prothymosin α adopts partially folded collapsed conformation at acidic pH. *Biochemistry* **38**, 15009-15016 (1999).
302 Bednar, J., Hamiche, A. & Dimitrov, S. H1-nucleosome interactions and their functional implications. *Biochim Biophys Acta* **1859**, 436-443 (2016).
303 Turner, A. L. *et al.* Highly disordered histone H1-DNA model complexes and their condensates. *Proc Natl Acad Sci U S A* **115**, 11964-11969 (2018).
304 Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693-705 (2007).
305 Gangaraju, V. K. & Bartholomew, B. Mechanisms of ATP dependent chromatin remodeling. *Mutat Res* **618**, 3-17 (2007).
306 Tagami, H., Ray-Gallet, D., Almouzni, G. & Nakatani, Y. Histone H3.1 and H3.3 Complexes Mediate Nucleosome Assembly Pathways Dependent or Independent of DNA Synthesis. *Cell* **116**, 51-61 (2004).
307 Mossink, B., Negwer, M., Schubert, D. & Nadif Kasri, N. The emerging role of chromatin remodelers in neurodevelopmental disorders: a developmental perspective. *Cell Mol Life Sci* **78**, 2517-2563 (2021).
308 Längst, G. & Manelyte, L. Chromatin Remodelers: From Function to Dysfunction. *Genes* **6**, 299-324 (2015).
309 Sandhu, K. S. Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. *J Mol Recognit* **22**, 1-8 (2009).
310 El Hadidy, N. & Uversky, V. N. Intrinsic Disorder of the BAF Complex: Roles in Chromatin Remodeling and Disease Development. *Int J Mol Sci* **20**, 5260 (2019).
311 Aoki, D. *et al.* Ultrasensitive Change in Nucleosome Binding by Multiple Phosphorylations to the Intrinsically Disordered Region of the Histone Chaperone FACT. *Journal of Molecular Biology* **432**, 4637-4657 (2020).
312 Santofimia-Castano, P. *et al.* Intrinsically disordered chromatin protein NUPR1 binds to the C-terminal region of Polycomb RING1B. *Proc Natl Acad Sci U S A* **114**, E6332-E6341 (2017).

313 Yan, L., Wu, H., Li, X., Gao, N. & Chen, Z. Structures of the ISWI–nucleosome complex reveal a conserved mechanism of chromatin remodeling. *Nature Structural & Molecular Biology* **26**, 258-266 (2019).
314 Warren, C. & Shechter, D. Fly Fishing for Histones: Catch and Release by Histone Chaperone Intrinsically Disordered Regions and Acidic Stretches. *Journal of Molecular Biology* **429**, 2401-2426 (2017).
315 Hodges, C., Kirkland, J. G. & Crabtree, G. R. The Many Roles of BAF (mSWI/SNF) and PBAF Complexes in Cancer. *Cold Spring Harbor perspectives in medicine* **6**, a026930 (2016).
316 Goldberg, A. D. *et al.* Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell* **140**, 678-691 (2010).
317 Drané, P., Ouararhni, K., Depaux, A., Shuaib, M. & Hamiche, A. The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. *Genes & development* **24**, 1253-1265 (2010).
318 Dyer, M. A., Qadeer, Z. A., Valle-Garcia, D. & Bernstein, E. ATRX and DAXX: Mechanisms and Mutations. *Cold Spring Harb Perspect Med* **7** (2017).
319 Gibbons, R. J. *et al.* Mutations in the chromatin-associated protein ATRX. *Hum Mutat* **29**, 796-802 (2008).
320 Mitson, M., Kelley, L. A., Sternberg, M. J. E., Higgs, D. R. & Gibbons, R. J. Functional significance of mutations in the Snf2 domain of ATRX. *Human Molecular Genetics* **20**, 2603-2610 (2011).
321 Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388**, 598-602 (1997).
322 Scacchetti, A. *et al.* CHRAC/ACF contribute to the repressive ground state of chromatin. *Life Science Alliance* **1**, e201800024 (2018).
323 Hartlepp, K. F. *et al.* The histone fold subunits of Drosophila CHRAC facilitate nucleosome sliding through dynamic DNA interactions. *Mol Cell Biol* **25**, 9886-9896 (2005).
324 Bonaldi, T., Längst, G., Strohner, R., Becker, P. B. & Bianchi, M. E. The DNA chaperone HMGB1 facilitates ACF/CHRAC-dependent nucleosome sliding. *EMBO J* **21**, 6865-6873 (2002).
325 Asfaha, Y. *et al.* Recent advances in class IIa histone deacetylases research. *Bioorganic & Medicinal Chemistry* **27**, 115087 (2019).
326 Gurova, K., Chang, H.-W., Valieva, M. E., Sandlesh, P. & Studitsky, V. M. Structure and function of the histone chaperone FACT – Resolving FACTual issues. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1861**, 892-904 (2018).
327 Di Giorgio, E. & Brancolini, C. Regulation of class IIa HDAC activities: it is not only matter of subcellular localization. *Epigenomics* **8**, 251-269 (2016).
328 Zhang, C. L., McKinsey, T. A. & Olson, E. N. The transcriptional corepressor MITR is a signal-responsive inhibitor of myogenesis. *Proceedings of the National Academy of Sciences* **98**, 7354 (2001).
329 Wang, A. H. & Yang, X. J. Histone deacetylase 4 possesses intrinsic nuclear import and export signals. *Mol Cell Biol* **21**, 5992-6005 (2001).
330 Black, Joshua C., Van Rechem, C. & Whetstine, Johnathan R. Histone Lysine Methylation Dynamics: Establishment, Regulation, and Biological Impact. *Molecular Cell* **48**, 491-507 (2012).
331 Martin, C. & Zhang, Y. The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology* **6**, 838-849 (2005).

332 Bedford, M. T. & Richard, S. Arginine Methylation: An Emerging Regulatorof Protein Function. *Molecular Cell* **18**, 263-272 (2005).
333 Sarmento, O. F. *et al.* The Role of the Histone Methyltransferase Enhancer of Zeste Homolog 2 (EZH2) in the Pathobiological Mechanisms Underlying Inflammatory Bowel Disease (IBD)*. *Journal of Biological Chemistry* **292**, 706-722 (2017).
334 Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nature Reviews Genetics* **13**, 343-357 (2012).
335 Separovich, R. J. *et al.* Post-translational modification analysis of Saccharomyces cerevisiae histone methylation enzymes reveals phosphorylation sites of regulatory potential. *Journal of Biological Chemistry* **296**, 100192 (2021).
336 Sen, P. *et al.* H3K36 methylation promotes longevity by enhancing transcriptional fidelity. *Genes & development* **29**, 1362-1376 (2015).
337 Tennen, R. I., Berber, E. & Chua, K. F. Functional dissection of SIRT6: Identification of domains that regulate histone deacetylase activity and chromatin localization. *Mechanisms of Ageing and Development* **131**, 185-192 (2010).
338 Pan, P. W. *et al.* Structure and Biochemical Functions of SIRT6*. *Journal of Biological Chemistry* **286**, 14575-14587 (2011).
339 Gil, R., Barth, S., Kanfi, Y. & Cohen, H. Y. SIRT6 exhibits nucleosome-dependent deacetylase activity. *Nucleic Acids Res* **41**, 8537-8545 (2013).
340 Liu, W. H. *et al.* Multivalent interactions drive nucleosome binding and efficient chromatin deacetylation by SIRT6. *Nature Communications* **11**, 5244 (2020).
341 Levendosky, R. F. & Bowman, G. D. Asymmetry between the two acidic patches dictates the direction of nucleosome sliding by the ISWI chromatin remodeler. *eLife* **8**, e45472 (2019).
342 Zhou, K., Liu, Y. & Luger, K. Histone chaperone FACT FAcilitates Chromatin Transcription: mechanistic and structural insights. *Current Opinion in Structural Biology* **65**, 26-32 (2020).
343 Saunders, A. *et al.* Tracking FACT and the RNA polymerase II elongation complex through chromatin in vivo. *Science* **301**, 1094-1096 (2003).
344 Gurard-Levin, Z. A., Quivy, J.-P. & Almouzni, G. Histone Chaperones: Assisting Histone Traffic and Nucleosome Dynamics. *Annual Review of Biochemistry* **83**, 487-517 (2014).
345 Winkler, D. D., Muthurajan, U. M., Hieb, A. R. & Luger, K. Histone Chaperone FACT Coordinates Nucleosome Interaction through Multiple Synergistic Binding Events*. *Journal of Biological Chemistry* **286**, 41883-41892 (2011).
346 Wang, T. *et al.* The histone chaperone FACT modulates nucleosome structure by tethering its components. *Life Science Alliance* **1**, e201800107 (2018).
347 Liu, Y. *et al.* FACT caught in the act of manipulating the nucleosome. *Nature* **577**, 426-431 (2020).
348 Mayanagi, K. *et al.* Structural visualization of key steps in nucleosome reorganization by human FACT. *Scientific Reports* **9**, 10183 (2019).
349 Tsunaka, Y., Ohtomo, H., Morikawa, K. & Nishimura, Y. Partial Replacement of Nucleosomal DNA with Human FACT Induces Dynamic Exposure and Acetylation of Histone H3 N-Terminal Tails. *iScience* **23**, 101641 (2020).
350 Szőllősi, E. *et al.* Intrinsic Structural Disorder of DF31, a Drosophila Protein of Chromatin Decondensation and Remodeling Activities. *Journal of Proteome Research* **7**, 2291-2299 (2008).

351 Crevel, G., Huikeshoven, H. & Cotterill, S. Df31 is a novel nuclear protein involved in chromatin structure in Drosophila melanogaster. *Journal of Cell Science* **114**, 37 (2001).
352 Guillebault, D. & Cotterill, S. The Drosophila Df31 Protein Interacts with Histone H3 Tails and Promotes Chromatin Bridging In vitro. *Journal of Molecular Biology* **373**, 903-912 (2007).
353 Schubert, T. *et al.* Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. *Mol Cell* **48**, 434-444 (2012).
354 Fuxreiter, M. *et al.* Malleable machines take shape in eukaryotic transcriptional regulation. *Nature chemical biology* **4**, 728-737 (2008).
355 Staby, L. *et al.* Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *Biochem J* **474**, 2509-2532 (2017).
356 Staller, M. V. *et al.* A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst* **6**, 444-455 e446 (2018).
357 Vuzman, D. & Levy, Y. Intrinsically disordered regions as affinity tuners in protein–DNA interactions. *Molecular BioSystems* **8**, 47 (2012).
358 Tafvizi, A., Huang, F., Fersht, A. R., Mirny, L. A. & van Oijen, A. M. A single-molecule characterization of p53 search on DNA. *Proc Natl Acad Sci U S A* **108**, 563-568 (2011).
359 Vuzman, D., Azia, A. & Levy, Y. Searching DNA via a "Monkey Bar" Mechanism: The Significance of Disordered Tails. *Journal of Molecular Biology* **396**, 674-684 (2010).
360 Brodsky, S. *et al.* Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol Cell* **79**, 459-471 e454 (2020).
361 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).
362 Lu, Y. *et al.* Reprogramming to recover youthful epigenetic information and restore vision. *Nature* **588**, 124-129 (2020).
363 Jang, J. *et al.* Disease-specific induced pluripotent stem cells: a platform for human disease modeling and drug discovery. *Experimental & Molecular Medicine* **44**, 202-213 (2012).
364 Xue, B., Oldfield, C. J., Van, Y. Y., Dunker, A. K. & Uversky, V. N. Protein intrinsic disorder and induced pluripotent stem cells. *Mol Biosyst* **8**, 134-150 (2012).
365 Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460-1465 (2020).
366 Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J. & Cramer, P. Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature* **580**, 669-672 (2020).
367 Weiss, M. A. Floppy SOX: mutual induced fit in hmg (high-mobility group) box-DNA recognition. *Mol Endocrinol* **15**, 353-362 (2001).
368 Holmes, Z. E. *et al.* The Sox2 transcription factor binds RNA. *Nat Commun* **11**, 1805 (2020).
369 Tan, C. & Takada, S. Nucleosome allostery in pioneer transcription factor binding. *Proc Natl Acad Sci U S A* **117**, 20586-20596 (2020).
370 Shammas, S. L. Mechanistic roles of protein disorder within transcription. *Curr Opin Struct Biol* **42**, 155-161 (2017).
371 Xhani, S. *et al.* Intrinsic disorder controls two functionally distinct dimers of the master transcription factor PU.1. *Sci Adv* **6**, eaay3178 (2020).

372 Robinson, J. L. & Carroll, J. S. FoxA1 is a key mediator of hormonal response in breast and prostate cancer. *Front Endocrinol (Lausanne)* **3**, 68 (2012).
373 Sutinen, P., Rahkama, V., Rytinki, M. & Palvimo, J. J. Nuclear mobility and activity of FOXA1 with androgen receptor are regulated by SUMOylation. *Mol Endocrinol* **28**, 1719-1728 (2014).
374 Hou, L. *et al.* Concurrent binding to DNA and RNA facilitates the pluripotency reprogramming activity of Sox2. *Nucleic Acids Res* **48**, 3869-3887 (2020).
375 Aksoy, I. *et al.* Sox transcription factors require selective interactions with Oct4 and specific transactivation functions to mediate reprogramming. *Stem Cells* **31**, 2632-2646 (2013).
376 Minderjahn, J. *et al.* Mechanisms governing the pioneering and redistribution capabilities of the non-classical pioneer PU.1. *Nat Commun* **11**, 402 (2020).
377 Chanda, S. *et al.* Generation of induced neuronal cells by the single reprogramming factor ASCL1. *Stem Cell Reports* **3**, 282-296 (2014).
378 Baronti, L. *et al.* Fragment-Based NMR Study of the Conformational Dynamics in the bHLH Transcription Factor Ascl1. *Biophys J* **112**, 1366-1373 (2017).
379 Fernandez Garcia, M. *et al.* Structural Features of Transcription Factors Associating with Nucleosome Binding. *Mol Cell* **75**, 921-932 e926 (2019).
380 Hiragami-Hamada, K. & Nakayama, J.-i. Do the charges matter?—balancing the charges of the chromodomain proteins on the nucleosome. *The Journal of Biochemistry* **165**, 455-458 (2019).
381 Pepenella, S., Murphy, K. J. & Hayes, J. J. Intra-and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma* **123**, 3-13 (2014).
382 Rowell, J. P., Simpson, K. L., Stott, K., Watson, M. & Thomas, J. O. HMGB1-facilitated p53 DNA binding occurs via HMG-Box/p53 transactivation domain interaction, regulated by the acidic tail. *Structure* **20**, 2014-2024 (2012).
383 Štros, M., Kučírek, M., Sani, S. A. & Polanská, E. HMGB1-mediated DNA bending: Distinct roles in increasing p53 binding to DNA and the transactivation of p53-responsive gene promoters. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1861**, 200-210 (2018).
384 Jayaraman, L. *et al.* High mobility group protein-1 (HMG-1) is a unique activator of p53. *Genes & development* **12**, 462-472 (1998).
385 Thomas, J. O. & Stott, K. H1 and HMGB1: Modulators of chromatin structure. *Biochemical Society Transactions* **40**, 341-346 (2012).
386 Kim, K. *et al.* Functional interplay between p53 acetylation and H1. 2 phosphorylation in p53-regulated transcription. *Oncogene* **31**, 4290-4301 (2012).
387 Tanaka, H. *et al.* HANP1/H1T2, a novel histone H1-like protein involved in nuclear formation and sperm fertility. *Mol Cell Biol* **25**, 7107-7119 (2005).
388 Holmstrom, E. D., Liu, Z., Nettels, D., Best, R. B. & Schuler, B. Disordered RNA chaperones can enhance nucleic acid folding via local charge screening. *Nat Commun* **10**, 2453 (2019).
389 Soto, L. F. *et al.* Compendium of human transcription factor effector domains. *Mol Cell* **82**, 514-526 (2022).
390 Mar, M., Nitsenko, K. & Heidarsson, P. O. Multifunctional Intrinsically Disordered Regions in Transcription Factors. *Chemistry* **29**, e202203369 (2023).
391 Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855 e1816 (2018).

392 Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol* (2023).
393 Taneja, I. & Holehouse, A. S. Folded domain charge properties influence the conformational behavior of disordered tails. *Curr Res Struct Biol* **3**, 216-228 (2021).
394 Keul, N. D. *et al.* The entropic force generated by intrinsically disordered segments tunes protein function. *Nature* **563**, 584-588 (2018).
395 Wiebe, M. S., Nowling, T. K. & Rizzino, A. Identification of novel domains within Sox-2 and Sox-11 involved in autoinhibition of DNA binding and partnership specificity. *J Biol Chem* **278**, 17901-17911 (2003).
396 Zaharias, S., Fargason, T., Greer, R., Song, Y. & Zhang, J. Electronegative clusters modulate folding status and RNA binding of unstructured RNA-binding proteins. *Protein Sci* **32**, e4643 (2023).
397 Zaharias, S. *et al.* Intrinsically disordered electronegative clusters improve stability and binding specificity of RNA-binding proteins. *J Biol Chem* **297**, 100945 (2021).
398 Newcombe, E. A., Delaforge, E., Hartmann-Petersen, R., Skriver, K. & Kragelund, B. B. How phosphorylation impacts intrinsically disordered proteins and their function. *Essays Biochem* **66**, 901-913 (2022).
399 Abdelalim, E. M., Emara, M. M. & Kolatkar, P. R. The SOX transcription factors as key players in pluripotent stem cells. *Stem Cells Dev* **23**, 2687-2699 (2014).
400 Shi, Y., Inoue, H., Wu, J. C. & Yamanaka, S. Induced pluripotent stem cell technology: a decade of progress. *Nature Reviews Drug Discovery* **16**, 115-130 (2017).
401 Werner, M. H., Huth, J. R., Gronenborn, A. M. & Clore, G. M. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell* **81**, 705-714 (1995).
402 Cox, J. L., Mallanna, S. K., Luo, X. & Rizzino, A. Sox2 uses multiple domains to associate with proteins present in Sox2-protein complexes. *PLoS One* **5**, e15486 (2010).
403 DelRosso, N. *et al.* Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* **616**, 365-372 (2023).
404 Erijman, A. *et al.* A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. *Mol Cell* **79**, 1066 (2020).
405 Piskacek, M., Otasevic, T., Repko, M. & Knight, A. The 9aaTAD Activation Domains in the Yamanaka Transcription Factors Oct4, Sox2, Myc, and Klf4. *Stem Cell Rev Rep* **17**, 1934-1936 (2021).
406 Nowling, T. K., Johnson, L. R., Wiebe, M. S. & Rizzino, A. Identification of the transactivation domain of the transcription factor Sox-2 and an associated co-activator. *J Biol Chem* **275**, 3810-3818 (2000).
407 Gagliardi, A. *et al.* A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J* **32**, 2231-2247 (2013).
408 Frederick, M. A. *et al.* A pioneer factor locally opens compacted chromatin to enable targeted ATP-dependent nucleosome remodeling. *Nat Struct Mol Biol* **30**, 31-37 (2023).
409 Nguyen, T. *et al.* Chromatin sequesters pioneer transcription factor Sox2 from exerting force on DNA. *Nat Commun* **13**, 3988 (2022).
410 Chen, J. J. *et al.* Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. *Cell* **156**, 1274-1285 (2014).

411 Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857-863 (2015).
412 Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annual Review of Biophysics* **45**, annurev-biophys-062215-010915 (2016).
413 Holmstrom, E. D. *et al.* Accurate Transfer Efficiencies, Distance Distributions, and Ensembles of Unfolded and Intrinsically Disordered Proteins From Single-Molecule FRET. *Methods Enzymol* **611**, 287-325 (2018).
414 Hofmann, H. *et al.* Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc Natl Acad Sci U S A* **109**, 16155-16160 (2012).
415 Krois, A. S., Dyson, H. J. & Wright, P. E. Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A* **115**, E11302-E11310 (2018).
416 Bugge, K. *et al.* Interactions by Disorder - A Matter of Context. *Front Mol Biosci* **7**, 110 (2020).
417 Vancraenenbroeck, R., Harel, Y. S., Zheng, W. & Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc Natl Acad Sci U S A* **116**, 19506-19512 (2019).
418 Müller-Späth, S. *et al.* Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci* **1211022107**, 4451-4457 (2010).
419 Higgs, P. G. & Joanny, J. F. Theory of Polyampholyte Solutions. *J Chem Phys* **94**, 1543-1554 (1991).
420 Heidarsson, P. O. *et al.* Release of linker histone from the nucleosome driven by polyelectrolyte competition with a disordered protein. *Nat Chem* **14**, 224-231 (2022).
421 Chafiaa, B., Rania Ghouil, A., Dingli, F., Loew, D. & Theillet, F. Structural characterization of stem cell factors Oct4, Sox2, Nanog, and Esrrb disordered domains, and a method to identify their phospho-dependent binding partners. *BioRxiv* (2023).
422 Werwein, E., Biyanee, A. & Klempnauer, K. H. Intramolecular interaction of B-MYB is regulated through Ser-577 phosphorylation. *FEBS Lett* **594**, 4266-4279 (2020).
423 Sun, X., Dyson, H. J. & Wright, P. E. A phosphorylation-dependent switch in the disordered p53 transactivation domain regulates DNA binding. *Proc Natl Acad Sci U S A* **118** (2021).
424 Schutz, S. *et al.* The Disordered MAX N-terminus Modulates DNA Binding of the Transcription Factor MYC:MAX. *J Mol Biol* **434**, 167833 (2022).
425 Justilien, V. *et al.* The PRKCI and SOX2 oncogenes are coamplified and cooperate to activate Hedgehog signaling in lung squamous cell carcinoma. *Cancer Cell* **25**, 139-151 (2014).
426 Schaefer, T. & Lengerke, C. SOX2 protein biochemistry in stemness, reprogramming, and cancer: the PI3K/AKT/SOX2 axis and beyond. *Oncogene* **39**, 278-292 (2020).
427 S, H.-P. *et al.* Sox2-phosphorylation toggles a bistable differentiation-switch in squamous cell carcinoma. *BioRxiv* (2021).
428 Müller, B. K., Zaychikov, E., Bräuchle, C. & Lamb, D. C. Pulsed interleaved excitation. *Biophysical journal* **89**, 3508-3522 (2005).
429 Pace, C. N. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol* **131**, 266-280 (1986).

430 Hillger, F., Nettels, D., Dorsch, S. & Schuler, B. Detection and analysis of protein aggregation with confocal single molecule fluorescence spectroscopy. *J Fluoresc* **17**, 759-765 (2007).
431 Zheng, W. *et al.* Inferring properties of disordered chains from FRET transfer efficiencies. *J Chem Phys* **148**, 123329 (2018).
432 Sisamakis, E., Valeri, A., Kalinin, S., Rothwell, P. J. & Seidel, C. A. M. Accurate Single-Molecule FRET Studies Using Multiparameter Fluorescence Detection. *Methods in Enzymology* **475**, 455-514 (2010).
433 Forster, T. *Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Annalen Der Physik* **2**, 55-75 (1948).
434 Nettels, D., Gopich, I. V., Hoffmann, A. & Schuler, B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2655-2660 (2007).
435 Gopich, I. V., Nettels, D., Schuler, B. & Szabo, A. Protein dynamics from single-molecule fluorescence intensity correlation functions. *J Chem Phys* **131**, 1-5 (2009).
436 Zheng, W. *et al.* Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **148**, 123329 (2018).
437 Hoffmann, A. *et al.* Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc Natl Acad Sci U S A* **104**, 105-110 (2007).
438 Schellman, J. A. Fifty years of solvent denaturation. *Biophys Chem* **96**, 91-101 (2002).
439 Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-293 (1995).
440 Skinner, S. P. *et al.* CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J Biomol NMR* **66**, 111-124 (2016).
441 Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J Biomol NMR* **70**, 141-165 (2018).
442 Mulder, F. A., Schipper, D., Bott, R. & Boelens, R. Altered flexibility in the substrate-binding site of related native and engineered high-alkaline Bacillus subtilisins. *J Mol Biol* **292**, 111-123 (1999).
443 Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc* **73**, 1-16 (2013).
444 Teilum, K., Kunze, M. B., Erlendsson, S. & Kragelund, B. B. (S)Pinning down protein interactions by NMR. *Protein Sci* **26**, 436-451 (2017).
445 Yang, Z. *et al.* UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* **179**, 269-278 (2012).
446 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).
447 Karanicolas, J. & Brooks, C. L., 3rd. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci* **11**, 2351-2361 (2002).
448 Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**, 43-56 (1995).
449 Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319-2327 (2011).

450 Lescop, E., Kern, T. & Brutscher, B. Guidelines for the use of band-selective radiofrequency pulses in hetero-nuclear NMR: example of longitudinal-relaxation-enhanced BEST-type 1H-15N correlation experiments. *J Magn Reson* **203**, 190-198 (2010).
451 Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, W276-W279 (2022).
452 Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes & Development* **28**, 2679-2692 (2014).
453 Alireza, L., George, Y., Sweta, S. & Unmesh, J. DNA methylation-dependent and -independent binding of CDX2 directs activation of distinct developmental and homeostatic genes. *bioRxiv*, 2024.2002.2011.579850 (2024).
454 Horisawa, K. & Suzuki, A. The role of pioneer transcription factors in the induction of direct cellular reprogramming. *Regenerative Therapy* **24**, 112-116 (2023).
455 Sunkel, B. D. & Stanton, B. Z. Pioneer factors in development and cancer. *iScience* **24**, 103132 (2021).
456 Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* **9**, 279-289 (2002).
457 Williams, D. C., Cai, M. & Clore, G. M. Molecular Basis for Synergistic Transcriptional Activation by Oct1 and Sox2 Revealed from the Solution Structure of the 42-kDa Oct1·Sox2·Hoxb1-DNA Ternary Transcription Factor Complex*. *Journal of Biological Chemistry* **279**, 1449-1457 (2004).
458 Weikum, E. R., Tuntland, M. L., Murphy, M. N. & Ortlund, E. A. A Structural Investigation into Oct4 Regulation by Orphan Nuclear Receptors, Germ Cell Nuclear Factor (GCNF), and Liver Receptor Homolog-1 (LRH-1). *Journal of Molecular Biology* **428**, 4981-4992 (2016).
459 Choi, Y. *et al.* FOXL2 and FOXA1 cooperatively assemble on the TP53 promoter in alternative dimer configurations. *Nucleic Acids Research* **50**, 8929-8946 (2022).
460 Ji, D. *et al.* FOXA1 forms biomolecular condensates that unpack condensed chromatin to function as a pioneer factor. *Molecular Cell* **84**, 244-260.e247 (2024).
461 Hou, L. *et al.* Concurrent binding to DNA and RNA facilitates the pluripotency reprogramming activity of Sox2. *Nucleic Acids Research* **48**, 3869-3887 (2020).
462 Frederick, M. A. *et al.* A pioneer factor locally opens compacted chromatin to enable targeted ATP-dependent nucleosome remodeling. *Nature Structural & Molecular Biology* **30**, 31-37 (2023).
463 Zhang, S. & Cui, W. Sox2, a key factor in the regulation of pluripotency and neural differentiation. *World J Stem Cells* **6**, 305-311 (2014).
464 Giorgetti, A. *et al.* Generation of induced pluripotent stem cells from human cord blood cells with only two factors: Oct4 and Sox2. *Nature Protocols* **5**, 811-820 (2010).
465 Wu, S. M. & Hochedlinger, K. Harnessing the potential of induced pluripotent stem cells for regenerative medicine. *Nature Cell Biology* **13**, 497-505 (2011).
466 Singh, V. K., Kalsan, M., Kumar, N., Saini, A. & Chandra, R. Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery. *Frontiers in Cell and Developmental Biology* **3** (2015).
467 Bjarnason, S. *et al.* DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2. *Nature Communications* **15**, 1445 (2024).
468 Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science*, eabb0074 (2020).

469 Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J. & Cramer, P. Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature* (2020).
470 Malaga Gadea, F. C. & Nikolova, E. N. Structural Plasticity of Pioneer Factor Sox2 and DNA Bendability Modulate Nucleosome Engagement and Sox2-Oct4 Synergism. *Journal of Molecular Biology* **435**, 167916 (2023).
471 Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning11Edited by T. Richmond. *Journal of Molecular Biology* **276**, 19-42 (1998).
472 Fung, H. Y. J., Birol, M. & Rhoades, E. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Current Opinion in Structural Biology* **49**, 36-43 (2018).
473 Bax, A. & Clore, G. M. Protein NMR: Boundless opportunities. *J Magn Reson* **306**, 187-191 (2019).
474 Kumar, D., Gautam, A. & Hosur, R. V. A unified NMR strategy for high-throughput determination of backbone fold of small proteins. *Journal of Structural and Functional Genomics* **13**, 201-212 (2012).
475 Uversky, V. N. Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1834**, 932-951 (2013).
476 Öztürk, M. A., Cojocaru, V. & Wade, R. C. Toward an Ensemble View of Chromatosome Structure: A Paradigm Shift from One to Many. *Structure* **26**, 1050-1057 (2018).
477 Burge, N. L. *et al.* H1.0 C Terminal Domain Is Integral for Altering Transcription Factor Binding within Nucleosomes. *Biochemistry* **61**, 625-638 (2022).
478 Thåström, A. *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences11Edited by T. Richmond. *Journal of Molecular Biology* **288**, 213-229 (1999).
479 Bai, Y. & Zhou, B.-R. Structures of Native-like Nucleosomes: One Step Closer toward Understanding the Structure and Function of Chromatin. *Journal of Molecular Biology* **433**, 166648 (2021).
480 Zhou, J., Ng, S.-B. & Chng, W.-J. LIN28/LIN28B: An emerging oncogenic driver in cancer stem cells. *The International Journal of Biochemistry & Cell Biology* **45**, 973-978 (2013).
481 Tsialikas, J. & Romer-Seibert, J. LIN28: roles and regulation in development and beyond. *Development* **142**, 2397-2404 (2015).
482 Huertas, J., Schöler, H. R. & Cojocaru, V. Histone tails cooperate to control the breathing of genomic nucleosomes. *PLOS Computational Biology* **17**, e1009013 (2021).
483 Espinoza-Fonseca, L. M. Aromatic residues link binding and function of intrinsically disordered proteins. *Molecular BioSystems* **8**, 237-246 (2012).
484 Shukla, S., Agarwal, P. & Kumar, A. Disordered regions tune order in chromatin organization and function. *Biophysical Chemistry* **281**, 106716 (2022).
485 Ferreira, H., Somers, J., Webster, R., Flaus, A. & Owen-Hughes, T. Histone Tails and the H3 αN Helix Regulate Nucleosome Mobility and Stability. *Molecular and Cellular Biology* **27**, 4037-4048 (2007).
486 Pepenella, S., Murphy, K. J. & Hayes, J. J. Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma* **123**, 3-13 (2014).
487 Luzete-Monteiro, E. & Zaret, K. S. Structures and consequences of pioneer factor binding to nucleosomes. *Current Opinion in Structural Biology* **75**, 102425 (2022).

488 Iwafuchi, M. *et al.* in *Nature Genetics* Vol. 52 418-427 (Springer US, 2020).
489 Veerapandian, V. *et al.* Directed Evolution of Reprogramming Factors by Cell Selection and Sequencing. *Stem Cell Reports* **11**, 593-606 (2018).
490 Millar, A. H. *et al.* The Scope, Functions, and Dynamics of Posttranslational Protein Modifications. *Annual Review of Plant Biology* **70**, 119-151 (2019).
491 Lanouette, S., Mongeon, V., Figeys, D. & Couture, J. F. The functional diversity of protein lysine methylation. *Molecular Systems Biology* **10**, 724 (2014).
492 Drazic, A., Myklebust, L. M., Ree, R. & Arnesen, T. The world of protein acetylation. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1864**, 1372-1401 (2016).
493 Peng, L., Li, E.-M. & Xu, L.-Y. From start to end: Phase separation and transcriptional regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**, 194641 (2020).
494 Galvanetto, N. *et al.* Extreme dynamics in a biomolecular condensate. *Nature* **619**, 876-883 (2023).