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

Sveinn Bjarnason[1#], Jordan A.P. McIvor[2#] Andreas Prestel[3], Kinga S. Demény[1], Jakob T. Bullerjahn[4], Birthe B. Kragelund[3], Davide Mercadante[2*], and Pétur O. Heidarsson[1*]

[1]*Department of Biochemistry, Science Institute, University of Iceland, Sturlugata 7, 102 Reykjavík, Iceland.*

[2]*School of Chemical Science, University of Auckland, Auckland, New Zealand.*

[3]*REPIN and Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark.*

[4]*Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Straße 3, 60438, Frankfurt am Main, Germany.*

[#]*These authors contributed equally: Sveinn Bjarnason, Jordan McIvor.*

Correspondence to davide.mercadante@auckland.ac.nz, pheidarsson@hi.is

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.

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[1,2]. 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[3], contributing competence for phase separation[4], or regulating DNA binding specificity[5]. In recent years, the importance of electrostatic interactions for the conformational dynamics of IDRs has become increasingly evident[6,7]. Computational modelling has suggested that charged patches on folded domains modulate the dimensions of adjacent IDRs, which might have direct functional consequences[8,9], and charge modulation by posttranslational modifications (PTMs) such as phosphorylation can have a large impact on the ensemble[2,10]. 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[11]. Sox2 is classified as a pioneer transcription factor due to its ability to target its cognate binding sequence in condensed, nucleosome-rich DNA[12]. 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[13,14]. Sox2 has 317 residues and consists of a small HMG-box DBD[15] flanked by a short 40-residue low-complexity N-terminal region, and a long ~200-residue C-terminal IDR (C-IDR), both of which are predicted to be disordered (**Fig. 1a**). The DBD is rich in positively charged residues (net charge = +13)– as commonly observed in DNA-binding proteins[2,16]– 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

2

IDRs[17], and AD2 (residues ~250-300)[18-21] (**Fig. 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[22]. There is evidence that the IDRs of Sox2 are neccessary for pioneering function[23] 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[24,25]. 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[26], RNA binding[23,27], and DNA scanning and target site selection[28]. 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[2], 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.

RESULTS

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[29-31], 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 (**Fig. 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 (**Fig. 1b, Supplementary Table 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.
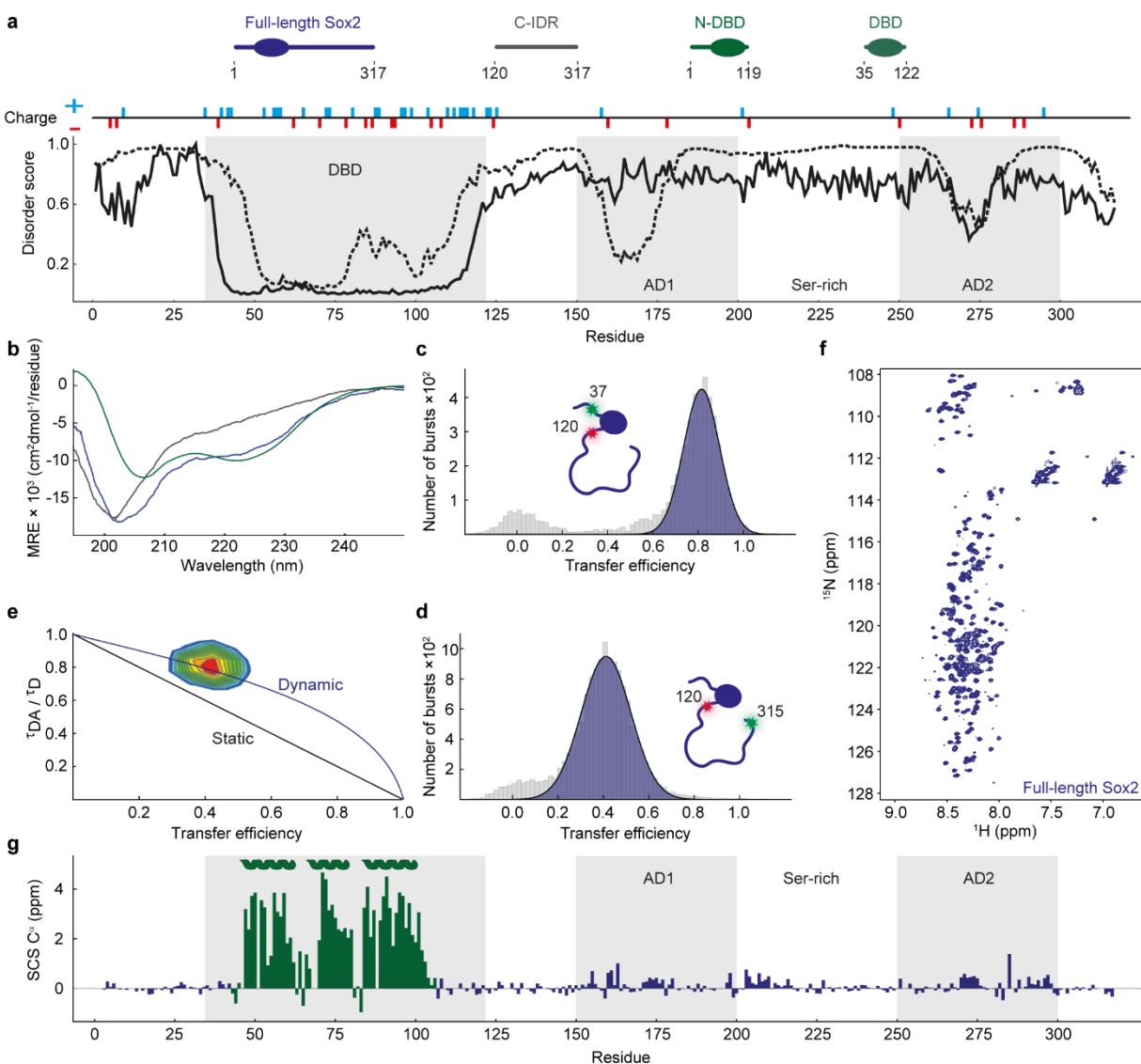

Fig. 1. Sox2 C-IDR is disordered and dynamic. **a)** Schematics of Sox2 illustrating the main constructs used in this study. The plot shows disorder predictions as a function of residue number, based on two different predictors (Disopred3[32] (dashed line), AlphaFold[19] normalized pLDDT (solid line)). 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). **e**) Fluorescence lifetime analysis of the Sox2 C-IDR. The 2D-correlation plot shows fluorescence lifetimes of the Cy3b donor (τ_{DA}) relative to the intrinsic donor fluorescence (τ_D). The dynamic line is based on a SAW-ν 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-

4

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[33,34]. 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, **Fig. 1c**) we measured an $\langle E \rangle \sim 0.8$, which corresponds to an average distance between the dyes close to that expected from the NMR solution structure (PDB 2LE4), indicating that the DBD remains folded in our experiments (**Methods** and **Supplementary Table 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 $\langle E \rangle = 0.43$ (**Fig. 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 v (SAW-v) to determine the root mean square distance (R_{RMS}) between the two dyes (**Methods**). The SAW-v model has recently been shown to describe well the dimensions of intrinsically disordered proteins (IDPs)[35]. The $\langle E \rangle$ of the C-IDR leads to an R_{RMS} of 7.5 nm and a scaling exponent v of 0.57, which is within the range expected for an IDP[36].

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 (τ_D) and presence (τ_{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 (**Fig. 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 $^1H^{15}N$-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 (**Fig. 1f**), with little dispersion of resonances in the proton dimension, characteristic of an IDR[6]. 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 (**Fig. 2, Fig. S1, and Methods**). A secondary chemical shift (SCS) analysis of C^{α} and C^{β} 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 (**Fig. 1a,g**).

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[37-39]. Charged residues can partake in long-range interactions and play a primary role in the conformational dynamics of IDRs[6]. 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[40]. 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)(**Fig. 2a**), whereas the end-to-end distance of the DBD (fluorescently labelled in residues 37 and 120) was largely independent of context (**Fig. 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-v distance distribution using the measured $\langle E \rangle$ at each denaturant concentration), was plotted as a function of titrant

concentration (**Fig. 2c,d, Fig. S2**). 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% (**Fig. 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 (**Fig. 2f, Fig. S2**). 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[41] and can be explained by polyampholyte theory[42,43]; 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[41]. Overall, even though the C-IDR contains relatively few charges causing it to adopt a collapsed state [40], charged interactions with the DBD sensitively control its dimensions further.
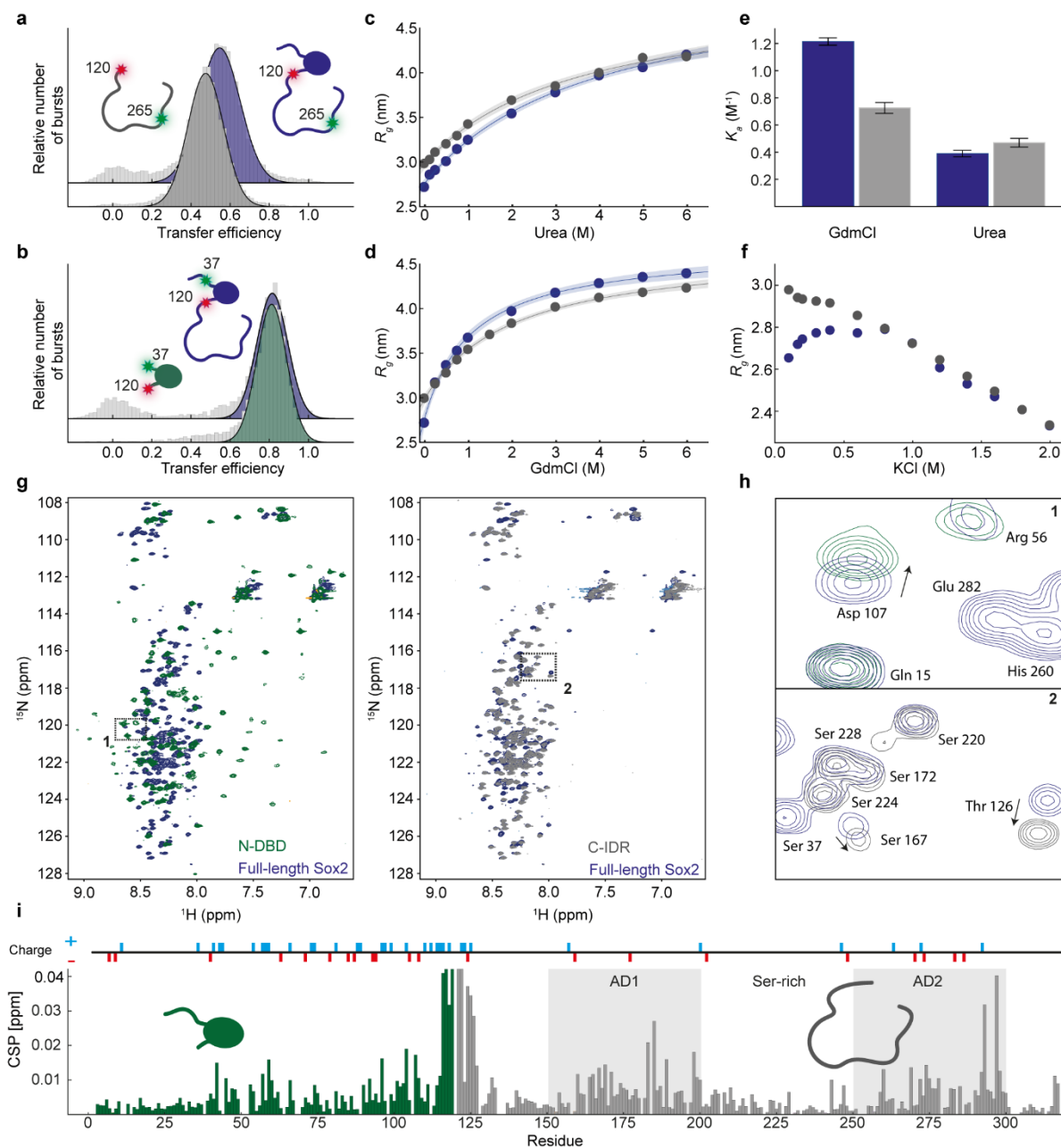
Fig. 2. Interdomain interactions between Sox2 DBD and C-IDR. a-b) Single-molecule transfer efficiency histograms of full-length Sox2 and either an isolated C-IDR, both labelled at positions 120 and 265 (a), or an isolated DBD, both labelled at positions 37 and 120 (b). **c-d)** Apparent R_g (see text 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)** ^{1}H^{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). 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 (**Fig. 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 (**Fig. S3**). 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 (**Fig. 2a**). The regions with the largest chemical shift perturbations (CSPs) overlapped with regions of the highest charge density (**Fig. 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 mixing 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 (**Fig. S4**).

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[29] by using fluorescently labelled oligonucleotides carrying a Sox2 binding site (ACGTGA) (**Supplementary Table 3**). At physiological salt concentrations (165 mM KCl), the free 15 bp dsDNA had a FRET efficiency $\langle E \rangle$ ~0.4 (**Fig. 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[23,27] (**Fig. 3b, Fig. S5, and Supplementary Table 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 (**Fig. S5**). 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 (**Fig. S5**) 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 (**Fig. 3c**). Given the lack of effects on binding affinity, we were surprised to observe a substantial change in FRET efficiency; the C-IDR expanded considerably upon binding DNA, with FRET decreasing from 0.43 to 0.28 (**Fig. 3c**). This is in contrast to the DBD end-to-end distance which even compacted slightly (**Fig. S5**). 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 (**Fig. 3e**).

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[29] (**Fig. 3d, Fig. S6, 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 (**Fig. 3c,d**), and fluorescence lifetime analysis showed slightly dampened dynamics (**Fig. 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 (**Fig. S6**)[44]. Overall, these data thus indicate that the conformational ensemble of the Sox2 C-IDR is similar in complex with DNA and nucleosomes.
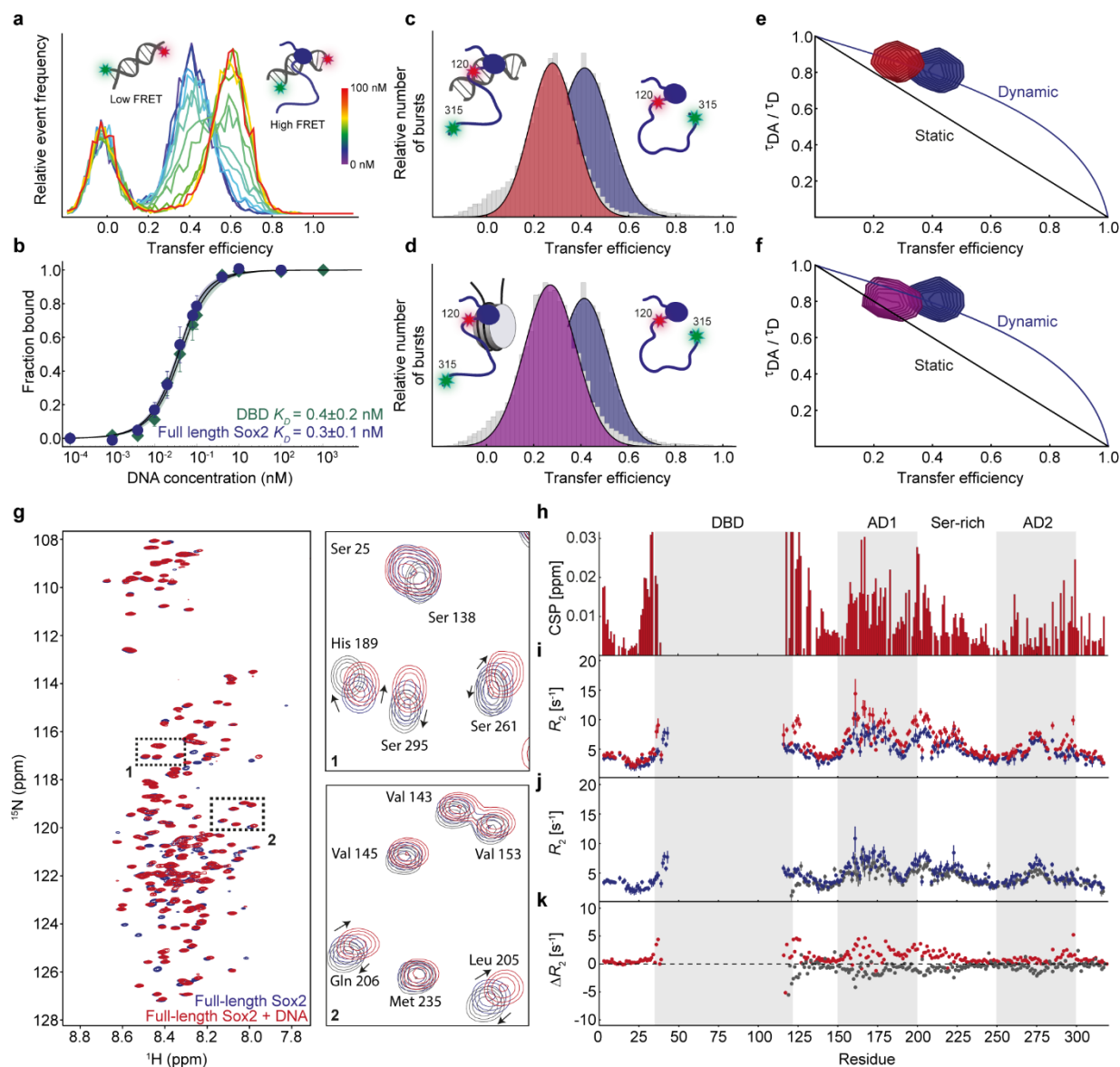
Fig. 3. Conformational rearrangements of the Sox2 C-IDR upon binding DNA and nucleosomes. a) Single-molecule transfer efficiency histograms of fluorescently labelled 15 bp DNA containing Sox2 binding site, with different concentrations of unlabelled full-length Sox2. The peak at E~0 corresponds to a 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)** ^{1}H^{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).

11

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 (**Fig. 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 (**Fig. 3h**, **Fig. 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 (**Fig. 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 (**Fig. 3g, zooms**). We then measured the fast time scale dynamics of the different states using NMR. Residue-specific relaxation rates (**Fig. 3i,j,k, Fig. S7**), 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.

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 (**Fig. 4a,b, Fig. S8**). 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, ε_{pp}, as it defines the interaction strength between the beads modelling the disordered regions of the protein (**Fig. 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 (**Fig. 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[6,41,44-46].
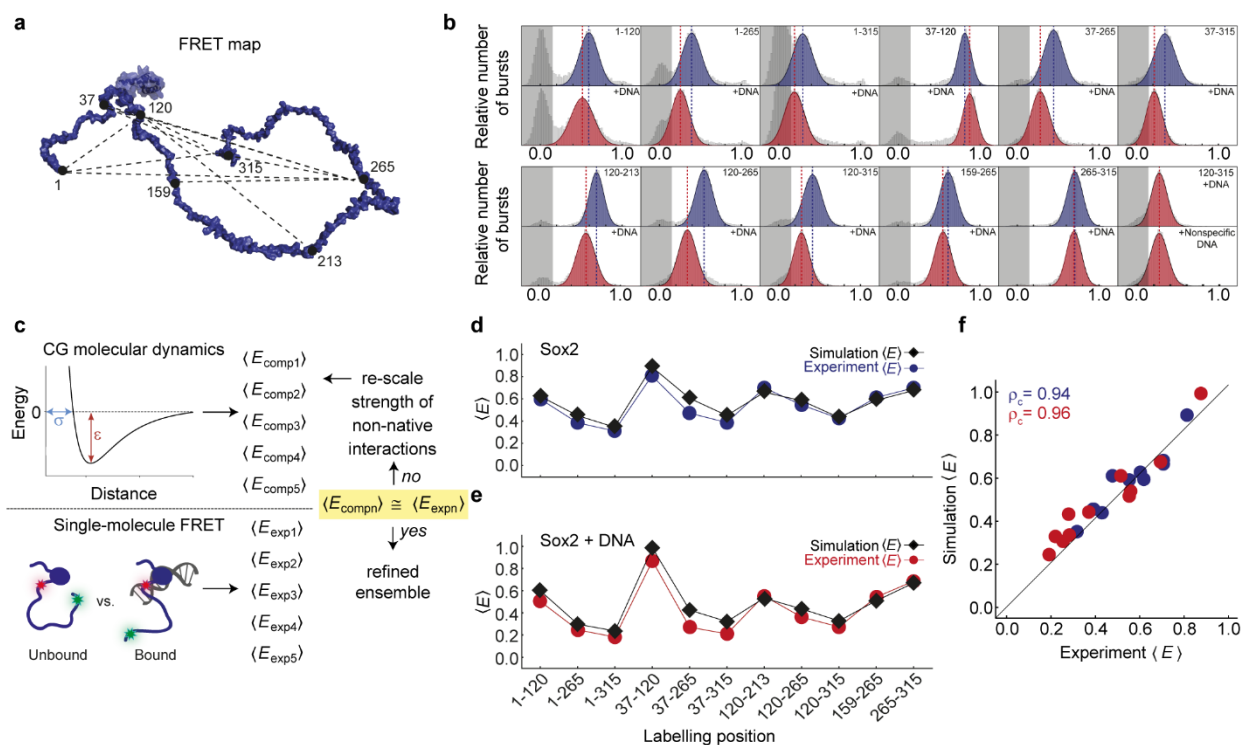

Fig. 4. Langevin dynamics simulations reproduce FRET efficiencies from smFRET experiments. a) Model of Sox2 showing FRET labelling positions that probe 11 unique intramolecular distances. **b)** Single-molecule transfer efficiency histograms of free and DNA-bound Sox2 variants, fluorescently labelled 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 (ε_{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 (concordance correlation coefficient ρ_c=0.94) and Sox2 bound to DNA (ρ_c=0.96). 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 ρ_c of 0.94 and 0.96, respectively (**Fig. 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,

13

in agreement with the FRET and NMR data (**Fig. 2, Fig. 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. S9**). The ensemble of free Sox2 expands as a function of salt concentration (**Fig. S9**) due to charge screening, but its dimensions reach a plateau at salt concentrations in proximity of the physiological range, in line with the experiments (**Fig. 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 (**Fig. 5a,b, Supplementary Movies 1 and 2**). In agreement with the experiments, the C-IDR dimensions are modulated by dynamic interactions with the DBD. When Sox2 binds to DNA, the C-IDR ensemble expands with more frequent excursions to extended states and thus a larger apparent R_g (**Fig. 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-and C-IDR, and decreased contacts of both IDRs with the DBD (**Fig. 5c**). When Sox2 binds the DNA, the region experiencing the largest variation in contact space is the C-IDR AD1 (**Fig. 5c,d,e**) directly in line with significant DNA-induced CSPs in the AD1 region (**Fig. 3h**). Projecting the average number of contacts for each residue onto the Sox2 structure (**Fig. 5d**) reveals an increase in proximity of regions overlapping with AD1 and DBD, when bound to DNA, but a decrease for AD2 (**Fig. 5e**). This differential engagement of the two ADs is re-iterated by the lifetimes of the contacts made by AD1 and AD2 with the DBD (**Fig. 5f,g**). The correlation function of contact formation over time, fits better to a double exponential with distinct slow and fast components. The contact lifetimes for the two ADs are similar in the absence of DNA. However, in complex with DNA, the lifetime 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 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 lifetimes before and after DNA binding (**Fig. S10**).
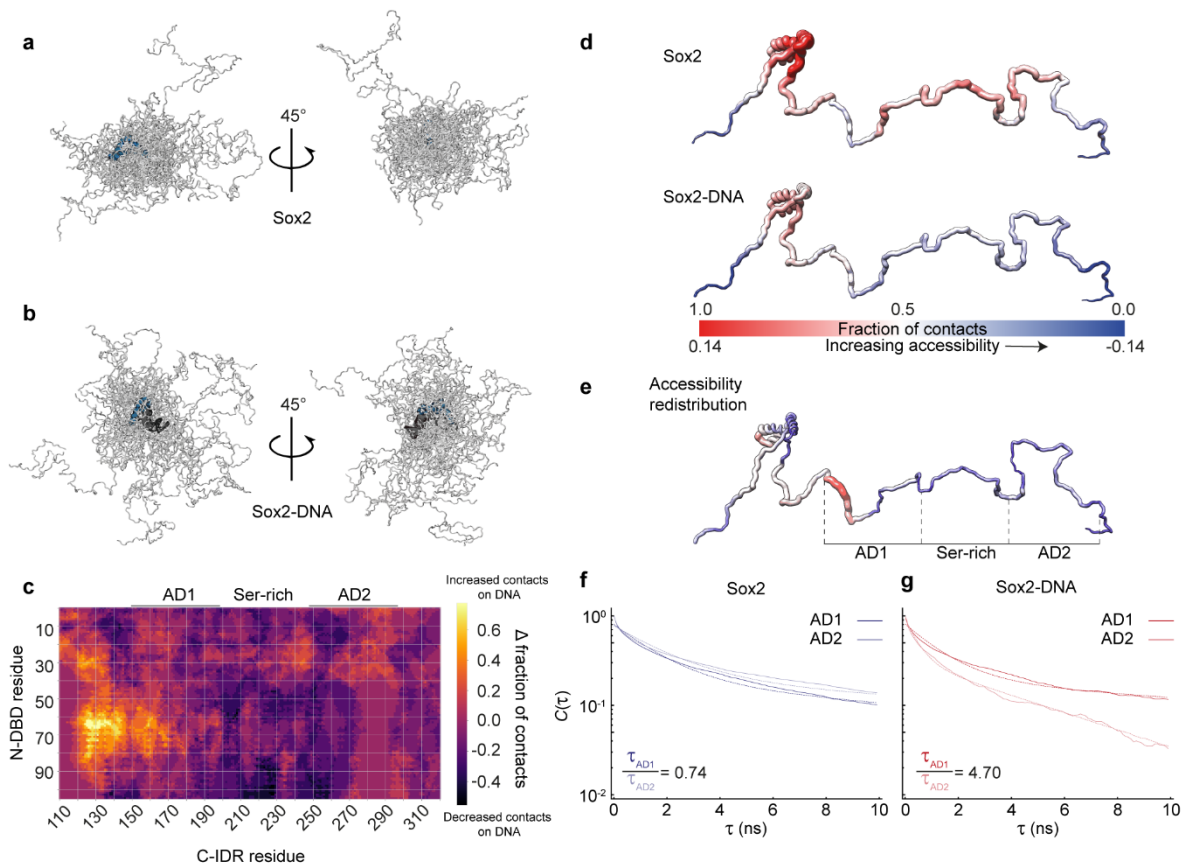
Fig. 5. Dynamic structural model of Sox2 ensembles, free and in complex with DNA. a,b) 20 representative snapshots from the simulation for a) free Sox2 and b) DNA-bound Sox2. The DBD is shown in blue, C-IDR in light grey, DNA in dark grey. **c)** Difference in the fraction of intramolecular contacts 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)** Fraction of contacts between the DBD and IDRs plotted on a schematic structure of Sox2 for free Sox2 (upper) and DNA-bound Sox2 (lower). **e)** Difference in fraction of contacts, plotted on a schematic structure of Sox2, between the unbound and bound states. Red regions correspond to increased contacts between DBD and C-IDR upon DNA binding, whereas blue regions indicate less contacts and thus more accessibility. The scale is shown in panel d, bottom numbers. **f-g)** Contact time autocorrelation functions (solid lines) for the regions corresponding to AD1 (solid colour) and AD2 (faded colour), for both free (f) and DNA-bound (g) Sox2. Double-exponential fits (dotted lines) reveal that the contact lifetimes of both ADs are approximately equal in free Sox2 but diverge in complex with DNA, where contact lifetimes are more than four times as long for AD1 compared to AD2.

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

15

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[40], 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. 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[47] soluble, until the right genomic binding site or coregulator is located. We found, in agreement with others[23,27], 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. Other TFs have recently been reported to have similar interdomain interactions, including cases such as B-MYB where no effects were observed on DNA binding affinity[48]. The N-terminal AD of the oncoprotein p53 has also been demonstrated to partake in intramolecular electrostatic interactions with its DBD, to a degree dependent on its phosphorylation state[38,49]. 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[50]. 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[51]. 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 and longer contact lifetimes upon DNA binding (**Fig. 5c,d,e**) whereas much of the remainder of the C-IDR, including AD2, has more than fourfold increased accessibility through much shorter contact lifetimes. 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[40] will dictate the exact

16

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. Increasing 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[52]. 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[12,29,30]. 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 (**Fig. S11**). 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)[2]. 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

METHODS

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$_4$Cl or ^{15}N-NH$_4$Cl 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$_2$PO$_4$, 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$_2$PO$_4$, 6 M Urea, pH 8.0) overnight and loaded onto

18

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_2O, 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.

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.

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 (ddH₂O) to a final concentration of 2.5 µM and stored at -20°C. Oligonucleotides intended for smFRET measurements were resuspended in DNA

19

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.

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.

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[53](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[54].

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[55]. The resulting bursts were then filtered to remove bursts where the acceptor bleaches during the transit of the molecule through the confocal volume[56], 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 $S = 0.3 - 0.7$ 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-v model[57], given by

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

which is characterized by the critical exponents v and $\gamma \approx 1.1615$. The constants A and α are determined by requiring $P(r)$ to be normalized and to satisfy $\langle r^2 \rangle = R^2$, respectively. The dependency on v in $P(r)$ is removed by assuming that a scaling law $R = b\, N^v$ must hold and substituting $v = \ln\left(\frac{R}{b}\right)/\ln(N)$ 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

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

In denaturation experiments, the Förster radius was corrected for changes in refractive index according to[41]:

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

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[58], 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[59].

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.

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 τ_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

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

23

Here, the variance σ^2 is given by

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

Determination of denaturant association coefficients

Association constants (K_a) of GdmCl and urea were determined using a weak denaturant binding model[60,61] 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, ΔE, and E_0 being fit parameters.

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 (θ) 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.

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

24

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

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

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[62] and analysed using CcpNmr software[63]. 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 $^1H^{15}N$ 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 (**Fig. 2 and Fig. S1**). Secondary structure content in Sox2 was determined from secondary C^{α} chemical shifts using a random coil reference for intrinsically disordered proteins[64].

T_1 and T_2 ^{15}N relaxation times were determined from 2×2 series of $^1H^{15}N$ 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[63] 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[65]

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

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[30]). The disordered regions, not available in the starting structure, were modelled using the modeller plugin[66] embedded in UCSF Chimera[67]. Each residue

25

of Sox2 was represented as a single bead mapped to the C^{α} atom of the starting full-atom structure. The simulation parameters used in the current work are identical to those outlined in Heidarsson et al[44]. We used the following potential energy function describing protein-protein interactions[6]:

$$V_{protein-protein} = \frac{1}{2}\sum_{i=1}^{N} k_b\left(d_i - d_i^0\right)^2 + \frac{1}{2}\sum_{i=1}^{N} k_\theta\left(\theta_i - \theta_i^0\right)^2$$

$$+ \sum_{i=1}^{N-2}\sum_{m=1}^{4} k_{i,m}\left(1 + cos\left(n\phi_i - \delta_{i,m}\right)\right) + \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_d\epsilon_0 d_{ij}} e^{-\frac{d_{ij}}{\lambda_D}}$$

$$+ \sum_{(i,j)\,\in Native} \varepsilon_{ij}\left[13\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right]$$

$$+ \sum_{(i,j)\,\notin Native} 4\varepsilon_{pp}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right]$$

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_b, k_θ for bond lengths and equilibrium values d_i^0 and θ_i^0 for angles. Both assignments are based on the distances and angles between the C^{α} 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[68]. 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 pK_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 (ϵ_d) set to a value of 80, the permittivity of the medium (ϵ_0), and the Debye screening length (λ_D), which is given by:

$$\lambda_D = \frac{\epsilon_0\epsilon_d k_B T}{2 N_A e^2 I},$$

where the Boltzmann constant and temperature are described by k_B and T, respectively, in addition to the Avogadro's number N_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[68]. In this potential, ε_{ij} describes the strength of the interaction calculated in accordance with a native-centric model[68], with $\sigma_{ij} = \frac{(\sigma_i + \sigma_j)}{2}$, determined based on C^α-C^α 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 ε_{pp} set to a value of 0.16 $k_B T$ (~0.4 kJ mol^{-1}) and σ_{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[6,44].

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[30]. 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:

$$V_{DNA} = \frac{1}{2}\sum_{i<N} k_b(d_i - d_i^0)^2 + \frac{1}{2}\sum_{i<N-1} k_\theta(\theta_i - \theta_i^0)^2 + \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_d \epsilon_0 d_{ij}} e^{-\frac{d_{ij}}{\lambda_D}}$$

$$+ \sum_{(i,j)\in\text{stack}} 4\varepsilon_{stack}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^6\right] + \sum_{(i,j)\in\text{pair}} 4\varepsilon_{pair}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^6\right]$$

$$+ \sum_{(i,j)\notin\text{pair,stack}} 4\varepsilon_{ns}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^6\right].$$

27

The bond distances and bond angles between the beads are described in the first and second terms, respectively. These two terms are identical to the ones used to treat the same interactions for protein beads. Specific values for these parameters were chosen to reproduce a typical persistence length (~50 nm) of double helical DNA at the equilibrium[69]. As described above, a modified Coulomb potential where ϵ_0 is the permittivity of free space and ϵ_d a dielectric constant set to 80 to mimic an aqueous environment, was used to account for electrostatic interactions, with a screening between charged beads defined by the Debye length λ_D. Non-bonded intearactions between DNA beads were accounted for by stacking and pairing terms and described by Lennard-Jones potentials, as shown by the third and fourth terms, respectively. While ε_{stack} was set to 3.0 $k_{B}T$ (~7.5 kJ mol^{-1}), ε_{pair} was set to 3.5 $k_{B}T$ (~8.8 kJ mol^{-1}), consistent with previously estimated free energy values for base stacking and shown to effectively reproduce the dynamics of B-DNA[44]. For DNA beads not involved in stacking or pairing interactions, a weakly attractive potential of $\varepsilon_{ns} = 0.04\ k_{B}T$ (~0.10 kJ mol^{-1}) was employed. All values used to parameterize the DNA in the current work have been previously used to succesfully describe DNA in a protein-DNA complex[44].

Protein-DNA interaction potential. The potential energy describing the interactions between protein and DNA beads has the following functional form:

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

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[30]). 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 ε_{ij} would be set to 2 $k_{B}T$ (~5 kJ mol^{-1}). Otherwise the contact was considered non-native and ε_{pd} was set to 0.06 $k_{B}T$ (~0.15 kJ mol^{-1}). For all contacts, σ_{ij} was set to a value of 0.5 nm. The values of ε_{ij} and ε_{pd} have previously been optimized to yield the best agreement between experimental and simulated FRET efficiencies[44].

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[70]. 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 λ_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[71].

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

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

AUTHOR CONTRIBUTIONS

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.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1 Soto, L. F. *et al.* Compendium of human transcription factor effector domains. *Mol Cell* **82**, 514-526 (2022).

2 Mar, M., Nitsenko, K. & Heidarsson, P. O. Multifunctional Intrinsically Disordered Regions in Transcription Factors. *Chemistry* **29**, e202203369 (2023).

3 Vuzman, D. & Levy, Y. Intrinsically disordered regions as affinity tuners in protein–DNA interactions. *Molecular BioSystems* **8**, 47 (2012).

4 Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855 e1816 (2018).

5 Brodsky, S. *et al.* Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol Cell* **79**, 459-471 e454 (2020).

6 Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61-66 (2018).

7 Schuler, B. *et al.* Binding without folding - the biomolecular function of disordered polyelectrolyte complexes. *Curr Opin Struct Biol* **60**, 66-76 (2020).

8 Taneja, I. & Holehouse, A. S. Folded domain charge properties influence the conformational behavior of disordered tails. *Curr Res Struct Biol* **3**, 216-228 (2021).

9 Keul, N. D. *et al.* The entropic force generated by intrinsically disordered segments tunes protein function. *Nature* **563**, 584-588 (2018).

10 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).

11 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).

12 Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568 (2015).

13 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).

14 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).

15 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).

16 Vuzman, D. & Levy, Y. DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21004-21009 (2010).

17 DelRosso, N. *et al.* Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* **616**, 365-372 (2023).

18 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).

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

20 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).

21 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).

22   Gagliardi, A. *et al.* A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.* **32**, 2231-2247 (2013).

23   Hou, L. *et al.* Concurrent binding to DNA and RNA facilitates the pluripotency reprogramming activity of Sox2. *Nucleic Acids Res* **48**, 3869-3887 (2020).

24   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).

25   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).

26   Nguyen, T. *et al.* Chromatin sequesters pioneer transcription factor Sox2 from exerting force on DNA. *Nat Commun* **13**, 3988 (2022).

27   Holmes, Z. E. *et al.* The Sox2 transcription factor binds RNA. *Nat Commun* **11**, 1805 (2020).

28   Chen, J. J. *et al.* Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. *Cell* **156**, 1274-1285 (2014).

29   Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460-1465 (2020).

30   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).

31   Remenyi, A. *et al.* Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* **17**, 2048-2059 (2003).

32   Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857-863 (2015).

33   Lerner, E. *et al.* Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science* **359**, eaan1133 (2018).

34   Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu Rev Biophys* **45**, 207-231 (2016).

35   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).

36   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).

37   Lee, G. M. *et al.* The structural and dynamic basis of Ets-1 DNA binding autoinhibition. *J. Biol. Chem.* **280**, 7088-7099 (2005).

38   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).

39   Bugge, K. *et al.* Interactions by Disorder - A Matter of Context. *Front Mol Biosci* **7**, 110 (2020).

40   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).

41   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).

42   Müller-Späth, S. *et al.* Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci* **1211022107**, 4451-4457 (2010).

43   Higgs, P. G. & Joanny, J. F. Theory of Polyampholyte Solutions. *J. Chem. Phys.* **94**, 1543-1554 (1991).

44   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).

45   Sottini, A. *et al.* Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat Commun* **11**, 5736 (2020).

46   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).

47   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).

48   Werwein, E., Biyanee, A. & Klempnauer, K. H. Intramolecular interaction of B-MYB is regulated through Ser-577 phosphorylation. *FEBS Lett.* **594**, 4266-4279 (2020).

49   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).

50   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).

51   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).

52   S, H.-P. *et al.* Sox2-phosphorylation toggles a bistable differentiation-switch in squamous cell carcinoma. *BioRxiv* (2021).

53   Müller, B. K., Zaychikov, E., Bräuchle, C. & Lamb, D. C. Pulsed interleaved excitation. *Biophys. J.* **89**, 3508-3522 (2005).

54   Pace, C. N. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131**, 266-280 (1986).

55   Schuler, B., Sonja, M., Soranno, A. & Nettels, D. in *Intrinsically disordered protein analysis. Methods in Molecular Biology (Methods and Protocols)* Vol. 896 (eds Vladimir N. Uversky & A. Keith Dunker) 21-45 (Springer New York, 2012).

56   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).

57   Zheng, W. *et al.* Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **148**, 123329 (2018).

58   Sisamakis, E., Valeri, A., Kalinin, S., Rothwell, P. J. & Seidel, C. A. M. Accurate Single-Molecule FRET Studies Using Multiparameter Fluorescence Detection. *Methods Enzymol.* **475**, 455-514 (2010).

59   Forster, T. *Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Annalen Der Physik* **2**, 55-75 (1948).

60   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).

61   Schellman, J. A. Fifty years of solvent denaturation. *Biophys. Chem.* **96**, 91-101 (2002).

62   Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277-293 (1995).

63   Skinner, S. P. *et al.* CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J. Biomol. NMR* **66**, 111-124 (2016).

64   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).

65   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).

66   Yang, Z. *et al.* UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* **179**, 269-278 (2012).

67   Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605-1612 (2004).

68   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).

69   Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* **34**, 564-574 (2006).

70   Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43-56 (1995).

71   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).