Phosphorylation regulated conformational diversity and topological dynamics of an intrinsically disordered nuclear receptor

Vasily Akulov[1,2,3], Alba Jiménez Panizo[4], Eva Estébanez-Perpiñá[4], John van Noort[2,3], Alireza Mashaghi[1,3]*

[1] Medical Systems Biophysics and Bioengineering, Division of Systems Pharmacology and Pharmacy, Leiden Academic Centre for Drug Research, Leiden University, the Netherlands
[2] Leiden Institute of Physics, Leiden University, the Netherlands
[3] Centre for Interdisciplinary Genome Research, Faculty of Science, Leiden University, the Netherlands
[4] Department of Biochemistry and Molecular Biomedicine, Institute of Biomedicine (IBUB) of the University of Barcelona (UB), Barcelona, Spain

*Correspondence: a.mashaghi.tabari@lacdr.leidenuniv.nl

Abstract

Site-specific phosphorylation of disordered proteins is often considered as a marker of protein activity, yet it is unclear how phosphorylation alters conformational dynamics of disordered protein chains, such as those in the nuclear receptor superfamily. In the case of disordered human glucocorticoid receptor N-terminal domain (GR NTD), a negatively charged region known as core activation function 1 (AF1c) features three phosphorylation sites, regulating its function and intracellular localization. Deletion of this sequence reduces GR transcriptional activation ability dramatically in cell experiments. By developing a circuit topology-based fold analysis approach, combined with atomistic simulations, we reveal that site-specific phosphorylation facilitates formation of non-local contacts, leading to the emergence of disordered compact topologies with significant entanglement, which are distinct from solvent exposed topologies. While we observe that the topological buildup of solvent-exposed states is similar in different phosphovariants, it depends on the exact phosphorylation site for the disordered compact states. This study thus reveals the complex regulatory role of the GR phosphorylation and introduces a unique analysis framework that can be broadly applied to studying topological dynamics of disordered proteins.

Key words: glucocorticoid receptor; nuclear receptor; phosphorylation; post-translational modification; circuit topology; molecular dynamics simulations; intrinsically disordered protein

Introduction

Protein molecules can adopt a broad spectrum of conformational states, ranging from disordered conformations to remarkably stable amyloid fibers [1-3]. In contrast to stably folded proteins, which tend to form unique, dense globules stabilized by interchain contacts, disordered proteins exhibit dynamic, solute-exposed conformations. The absence of a dense globule leads to significant structural fluctuations in disordered proteins. However, these are extreme cases; many proteins lie somewhere in between. Folded proteins can contain dynamic, disordered loops and turns, and intrinsically disordered proteins (IDPs) can expectedly be found in compact, disordered states, where certain residues form transient contacts and thus the chain forms a highly connected yet dynamic plexus. Evidence suggests that “disordered compact” states can be observed in isolated IDPs depending on their sequence [4],are linked to function [3, 5] and may be modulated by post-translational modifications. Reversible phosphorylation, for example, is thought to be a mechanism nature uses to modulate IDP function, presumably by regulating the conformational dynamics and compaction of these proteins. Consequently, IDPs presumably adopt a wide range of shapes and interact with different binding partners or partition to different subcellular spaces in a context-dependent manner [6], through largely unknown structural mechanisms. Despite active research on IDPs, these types of protein states have not been thoroughly studied, due to technical challenges.

Our understanding of IDPs is hindered by the lack of a proper description of the dynamics that capture topological motifs hidden within conformational noise. We do not yet know what patterns the transient contacts and dynamic loopy topologies form in disordered compact states, and whether these patterns can be used to identify these states. Recently, the topology of protein chains has been defined based on the arrangement of loops or the associated intrachain contacts. This approach, known as circuit topology (CT) [7], has been applied to stable folded proteins for various applications [8, 9] and has proven effective for modeling polymer folding reactions [10]. The approach has also been applied to disordered proteins, enabling the capture of conserved features in their topological fluctuations [11]. Additionally, it has been used to detect topological similarities between IDPs with similar functions, providing a new metric for quantifying structural similarity suitable for both IDPs and proteins with stable 3D structures [11]. As such, one may envision this method being applied to study phosphorylation-induced order in disordered chains, aiding in the identification and study of disordered compact conformations.

In this research, we develop an analysis framework by combining extensive molecular dynamics (MD) simulations and CT to study phosphorylation-regulated structural and dynamic properties of a disordered protein. We achieve this by studying the disordered N-terminal domain of the human glucocorticoid receptor (GR NTD) (Figure 1). This protein is a key transcription factor and a member of the nuclear receptor superfamily, which is implicated in many physiological and pathological processes such as metabolism, homeostasis and inflammation [12]. The segment of the NTD, known as activation function 1 (AF1; comprising the residue region 77-262) and, more specifically, the AF1 core (AF1c, including residues 187-244) is of particular importance, as is the case in many other related nuclear receptors. The hormone-independent AF1 is required for maximal transcriptional activation of GR [13] . Deletion of this sequence reduces GR transcription activation ability for GR-dependent test genes by at least 60–70% [1]. AF1 is subject to post-translational modifications, that contribute to the complexity and context-dependence of GR signaling [14, 15]. Site-specific phosphorylation of GR NTD is often considered a marker of protein activation [2-5] and a

determinant of its subcellular localization. However, its implications for secondary or higher-order structural changes are largely unknown despite advances [2, 6-8]. Unfortunately, there is no detailed experimental structure neither for the full-length, nor for NTD fragments such as AF1c. Modeling data on its dynamics is also lacking, as the intrinsically disordered nature of AF1c precludes conformational analysis. Our IDP analysis approach reveals that phosphorylation tunes the structural diversity of this disordered protein and leads to the emergence of disordered compact states, which are detectable using the CT-based contact analysis approach.

Results

All-atom molecular dynamics simulations reveal the disordered nature of GR-AF1c under various phosphorylation states

We selected five phosphovariants of AF1c, which are reportedly visited by the GR protein as it transitions from the cytoplasm to the nucleus, to study the effect of individual phosphorylations on the ensemble of conformations (Figure 1A). The chain may get phosphorylated at three serine residues, namely residues 203, 211, and 226 (Figure 1B), which are followed by prolines in the primary sequence. The phosphorylation of these sites correlates with GR activation and localization as depicted in Figure 1A [16-21]. The overlaid conformational samples of the simulated AF1c variants are shown in Figure 1C (see Methods). These phosphorylation variants are: non-phosphorylated variant (NP); AF1c with phosphorylated serine 203 (pSer 203); AF1c with phosphorylated serine 211 (pSer 211); AF1c with phosphorylated serine residues 203 and 211 (pSer 203-211); AF1c with phosphoserines at positions 203, 211 and 226 (pSer 203-211-226). All these phosphorylation variants, except pSer 203-211-226 were directly observed *in vivo* [22]. The triple phosphorylated variant was observed *in vitro* [18]. Here, we focus our research on AF1c with all prolines in the trans isomer, which is commonly seen in cells and leave the analysis of trans-to-cis isomerization to future studies.

During the course of simulations, we observed that all variants were showing significant conformational disorder (Figure 1C). As can be seen in Figure 1C, overlaid structures of different phosphovariants representing the distribution of monomers are characteristic of a disordered chain. Additionally, the high fluctuations in solvent accessible surface area (SASA) (Figure S1) per residue show the dynamic nature of the proteins studied. The disordered nature of AF1c can be attributed to the unfavorable folding balance between charges (-13 in non-phosphorylated case at neutral pH) and hydrophobicity.

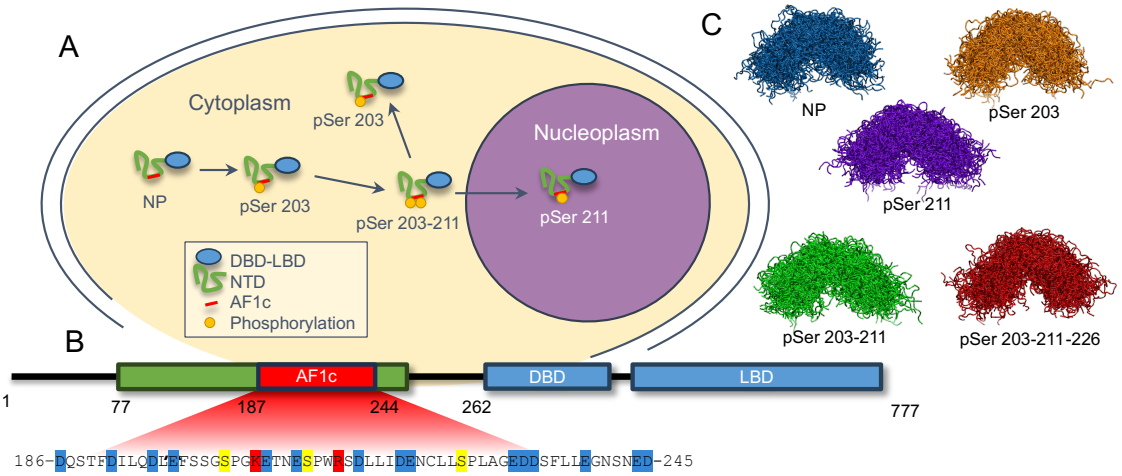


*Figure 1 Role of phosphorylation in GR structure-function relationship. (**A**) The GR phosphorylation pathway. In non-activated form GR is not phosphorylated, but upon activation first get phosphorylated Ser 203, then Ser 211. After that, protein is transported to the nucleus, and loses Ser203 phosphorylation. Taken with changes from [19] (**B**) The schematic representation of GR with the sequence of AF1c region. GR has a modular architecture consisting of an intrinsically disordered amino-terminal domain (NTD; colored in green), including core activation function region 1 (AF1c in red), a zinc-finger DNA-binding domain (DBD), and a carboxy-terminal ligand-binding domain (LBD), each contributing to the receptor's function. In the AF1c sequence, residues are color-coded: blue – negatively charged residues at neutral pH; red – positively charged residues at neutral pH; yellow – serine residues, which could be phosphorylated. (**C**) Overlaid structures of simulated AF1c phosphovariants.*

NMR shifts of simulated protein correlate well to experimentally measured values

From MD data of NP variants, we calculated NMR chemical shifts using SPARTA+ [23] and compared them with available experimental NMR chemical shifts [24]. Eight independent trajectories of each NP variant (8 microseconds in total) were combined for this analysis and the predicted NMR shift was correlated to the experimental data using the Pearson correlation test (*p-value*<0.05 for all comparisons) (Figure S2). The close agreement between simulated spectra and experimental data confirmed the validity of our simulation protocol. For all backbone atoms, shifts correlation coefficient *R* was more than 0.76. The deviation of HN shifts calculated from simulation structures is higher than in other backbone atoms. This could be explained by the disordered nature of the AF1c and the different geometries of protons interacting with solvent molecules or backbone oxygen atoms. As such, HN shifts are the most sensitive to the protein structure. The low *p-value* between simulated and experimental signals validates our simulation protocols, which we then use to study the phosphovariants, for which NMR data is lacking to date.

Sequential phosphorylation opens up protein structure

To quantify the effect of phosphorylation on the protein geometry, we visualized distributions of the radius of gyration (*Rg*) and the *SASA* (Figure 2 A,B). The NP molecule has the most compact shape, and upon phosphorylation, AF1c elongates and becomes more solvent-exposed. This trend is expected as more phosphorylation leads to more negative charge in the chain. Detailed analysis reveals that the *Rg* distributions of double and triple phosphorylated variants are relatively close. A single phosphorylation of Ser 211 alters the *SASA* and *Rg* more than a single phosphorylation of Ser 203. The *SASA* metric appears to be more sensitive to the protein structure, and in the case of phosphovariants two-peak distributions are observed, suggesting the co-existence of compact and solvent exposed

protein conformations. This indicates that local site-specific changes contribute to regulation of overall compaction.

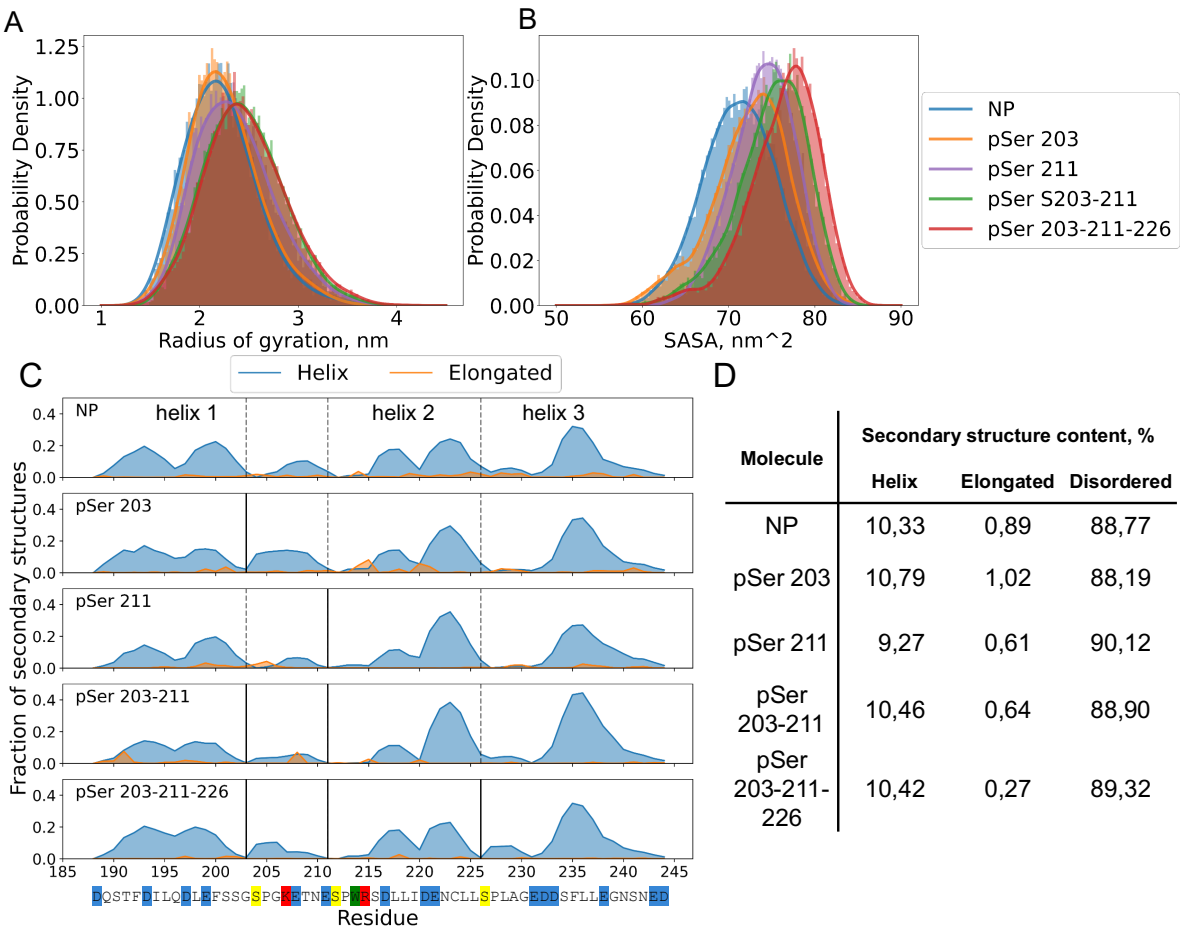
The observed radii of gyration are close to one another, and are consistent with theoretical expectations from the theory of disordered proteins. Our simulations yield an *Rg* for *NP* of 2.2±0.4 nm, which is slightly smaller than what is predicted for excluded volume of the chain, i.e. 2.5 nm. Interestingly, sequential phosphorylation shifts *Rg* to this limit. The pSer 203-211 and pSer 203-211-226 variants yield radii of 2.5±0.4 nm. Such an increase in *Rg* agrees with earlier literature [25, 26].

While we focused our research on using a conventional atomistic unbiased MD simulation protocol, we also tested the replica exchange with solute tempering (REST2) enhanced sampling method [27] (See Supplementary Information). Similar trends for the *Rg*, *SASA* and other parameters (Figure S6) were observed for both methods of simulation. However, we decided not to pursue REST2 for the remainder of this study because most of the generated conformations had a low likelihood, caused by the deviated Hamiltonian of the system (See Supplementary Information). This resulted in the lower speed of the modeling convergence relative to multiple replicas of conventional simulations.

Phosphorylation leads to the helical structure redistribution

Next, we studied the secondary structure content of different phosphovariants, using the DSSP tool [28]. As shown in Figure 2C the distribution of secondary structures showed a dependency on the phosphorylation state while the time-averaged amount of secondary structures was almost identical for all variants, see Figure 2D.

The secondary structure distribution of NP molecule corresponds well to what was expected from previous NMR experiments [24, 29]. More specifically, three helical regions could be observed: 189-203 (helix 1), 215-226 (helix 2), 231-245 (helix 3). Ser-Pro residues separate these helical regions. Single phosphorylation of the 203 site promotes an increase of the helical conformation in the helix 1 fragment. This could be attributed to charge interaction between phosphoserine and lysine in the sequence: pSer-Pro-Gly-Lys. A proline residue is known to be not only a helix breaker, but also a helix initiator residue. In combination with the small Gly residue, it tends to form a charge interaction between pSer and Lys, which are suitable for hydrogen backbone bonds in a helical structure. In contrast, pSer211 does not stabilize helix 2. In this case, phosphoserine and arginine are separated by tryptophan (pSer-Pro-Trp-Lys), which competes with Lys in interaction with pSer. This competition makes it harder to create stable hydrogen bonds and to form a helix. The same could be observed in the double phosphorylated variants. The pSer 203-211 and pSer 203-211-226 variants have more stable helix 1 and helix 3, respectively, than other phosphorylation variants. This could be attributed to structure elongation and the increased role of the short-range contacts within the protein backbone.



Molecule	Secondary structure content, %		
	Helix	Elongated	Disordered
NP	10,33	0,89	88,77
pSer 203	10,79	1,02	88,19
pSer 211	9,27	0,61	90,12
pSer 203-211	10,46	0,64	88,90
pSer 203-211-226	10,42	0,27	89,32

*Figure 2 An increase in the protein net charge opens up its structure and promotes the redistribution of secondary structures. The radius of gyration (**A**) and the Solvent Accessible Surface Area (**B**) distributions for different phosphovariants are displayed. (**C**) The fraction of secondary structure per residue for different phosphovariants. Blue – helical structure; orange – elongated. Vertical lines representing phosphorylation sites, dashed – non-phosphorylated, solid – phosphorylated. (**D**) The average secondary structure for different phosphovariants. Secondary structure was calculated from combined trajectories for each phosphorylation type with help of DSSP tool [28].*

Contact analysis reveals distinct subpopulations

Next, we focus on intra-chain contact analysis. We identified important intra-chain contacts by performing a principal component analysis (PCA) to determine and visualize phosphorylation-induced patterns in contact space. First, each frame from the simulation was represented with a vector of protein intra-chain contacts, with a cutoff of 0.5 nm. Then PCA was performed on the trajectories of all phosphovariants projecting them onto the same space. Note that the PCA is an unsupervised method, where the identified components are defined solely by data variability.

The first two eigenvectors of the PCA show the importance of short-range contacts (Figure 3A), identifying mainly the ones that form helical secondary structure. Yet, the first principal component could be attributed to the formation of helix 3 and some destabilization of helix 1, while the second principal component represents stabilization of helix 1 and helix 2, see Figure 3A.

Our PCA analysis of the studied phosphovariants revealed distinct conformational distributions, which we plotted as energy landscapes in Figure 3B. The energy landscapes of NP and pSer 211 are close to each other: a shallow energy basin is elongated along PCA axis 2, showing minima along the two helix component (eigen vector 2) together with unstructured conformations. The pSer 203 variant contains a more pronounced axis 1 component, and a less pronounced axis 2 character. Importantly, pSer 203-211 has a unique pattern of energy surface, featuring two wells: one representing conformations with stable helices 1 and 2 and the second one representing conformations with stable helix 3. This double phosphovariant has one percent higher amount of secondary structure than others. The most stable conformations of triple phosphovariant are disordered with a well at the bottom-left side of the energy landscape; this makes conformation of the triple variant close to that of NP and pSer 211.

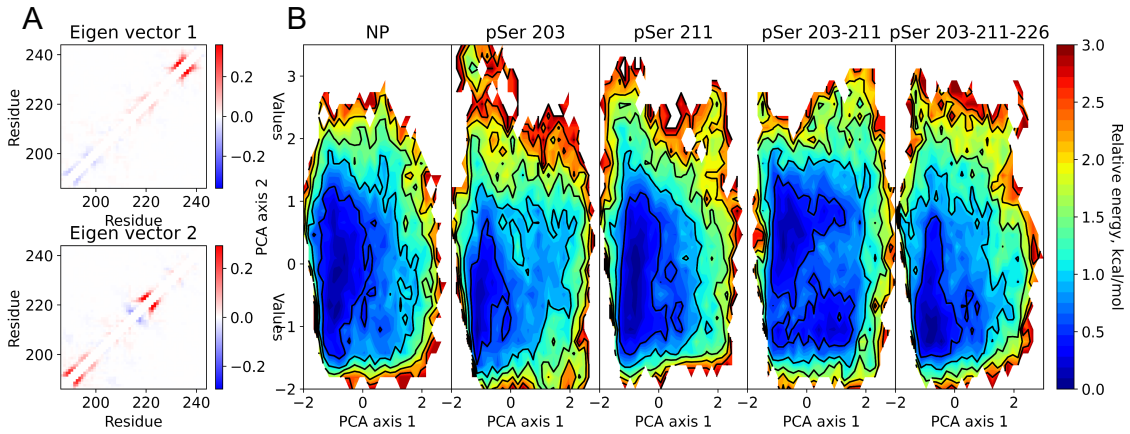


*Figure 3 Principal component analysis identifies folding of helices 1 and 2 in AF1c independently from folding of helix 3. (**A**) Contributions to the first two eigenvectors of the PCA transformation per residue. (**B**) Energy landscape of different phosphorylation types.*

Circuit topology as a tool to identify compact structures

To better resolve transient conformations in the predominantly disordered AF1c structures, we developed and applied a novel circuit topology-based approach (Figure 4A) and classified disordered conformations based on the complexity of their topological structure. The contact analysis presented in the previous section revealed the importance of secondary structures formed by short-range contacts. Topological arrangements of contacts, particularly non-local ones, may also be informative in understanding the conformational dynamics of disordered chains on a more global level. Importantly, AFc1 does not fold in isolation within the timescale of our simulations, consistent with previous experimental reports [24, 30]. However, they may form Disordered Compact States (DCS), which can be resolved using CT.

CT formalizes the pairwise arrangement of contacts in a chain, which can be in Series (S), in Parallel (P), or cross each other (X), as shown in Figure 6A. To resolve the emergence of DCSs, we introduce a simple metric called *CT Folding Score*. The main idea of the *CT Folding Score* is that, upon folding of a protein chain, the amount of topological relations should increase, because of the increase in number of contacts. Compared to simply counting the overall number of contacts, this approach is more sensitive to the formation of non-local contacts and contains information about the arrangement of contacts. Consequently, the protein conformations classified as a “compact”, using *CT Folding Score,* have a smaller radius of gyration *Rg*, despite a varying amount of helical structure (Figure 4B).

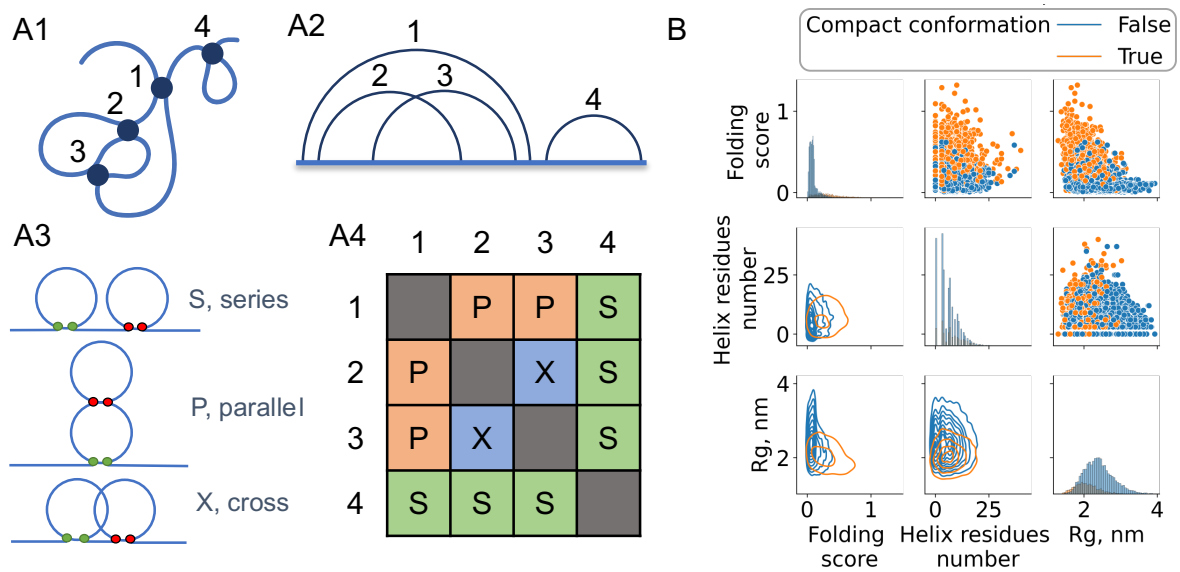


*Figure 4 Circuit topology (CT) as metric to track conformational changes in IDPs. (**A1**) CT classifies the relation between intra-chain contacts in a protein. (**A2**) Circuit diagram (arc diagram), where arcs represent contacts between two residues, clearly reveals how contacts and associated loops are entangled. In this example, a circuit is shown which includes four contacts. (**A3**) A given contact pair in circuits could have three basic arrangements: series (S), parallel (P), and cross (X). (**A4**) A protein structure represented with a CT matrix. In the CT matrix, rows/columns represent contacts, and each cell represents the relationship between these contacts. (**B**) The pairwise relationships between the radius of gyration (Rg), helix residues number, and folding score (see Methods for definition). Colors represent the distribution of compact conformations (orange) and solvent exposed (blue) conformations.*

Phosphorylation of AF1c affects the propensity and kinetics to form disordered compact conformations in a site-specific manner. In Figure 5A, the *CT Folding Score* dynamics for each phosphovariant is represented, from combining all eight replicate trajectories. The temporal evolution of the CT Folding Score during the simulations is represented by the *Frame Number* on the x-axis. Considering *CT Folding Score* as a metric of protein compaction, a burst can be interpreted as a compaction event. In the case of pSer 203, pSer 203-211, and pSer 203-211-226 molecules, several long living compact topologies were observed. Interestingly both NP and pSer 211 have kinetics patterns that are distinct from those of other phosphovariants. This difference could be explained by fast transitions between disordered and helix containing states, consistent with the absence of a free energy barrier between them (see Figure 3). In the case of double and triple phosphovariant, stretches of a low score followed by long-living bursts were observed, reflecting periods of elongated protein structures, followed by sudden compaction.

Contact maps show that solvent-exposed disordered structures are very common in all phosphovariants of AF1c. With the *CT Folding Score,* we can now analyze the effect of phosphorylation on contact formation propensity separately for disordered compact and disordered solvent-exposed conformations (Figure 5B). The contact maps show that the compact conformations contain not only non-local contacts, but also have a higher propensity to form helical structures. In particular, compact conformations of pSer 203 and pSer 203-211-226 molecules show interactions between helix 1 and helix 2. Interestingly, in the case of pSer 203, residues 203-211 form a helix (Figure 4B), but in pSer 203-211, a disordered region could be observed between helix 1 and residues 203-211. Some helices contain kinks within them (Figure 5B). When the helix is interrupted by a disordered region, we observed that the two helix parts tend to interact with each other. Interestingly, despite NP having the smallest

average *Rg* and *SASA*, its compact shape is formed by intra-helical region contacts and increased non-ordered contacts in the 203-211 region. Although the solvent-exposed conformations still contain a significant amount of helical structures, they lack non-local contacts.

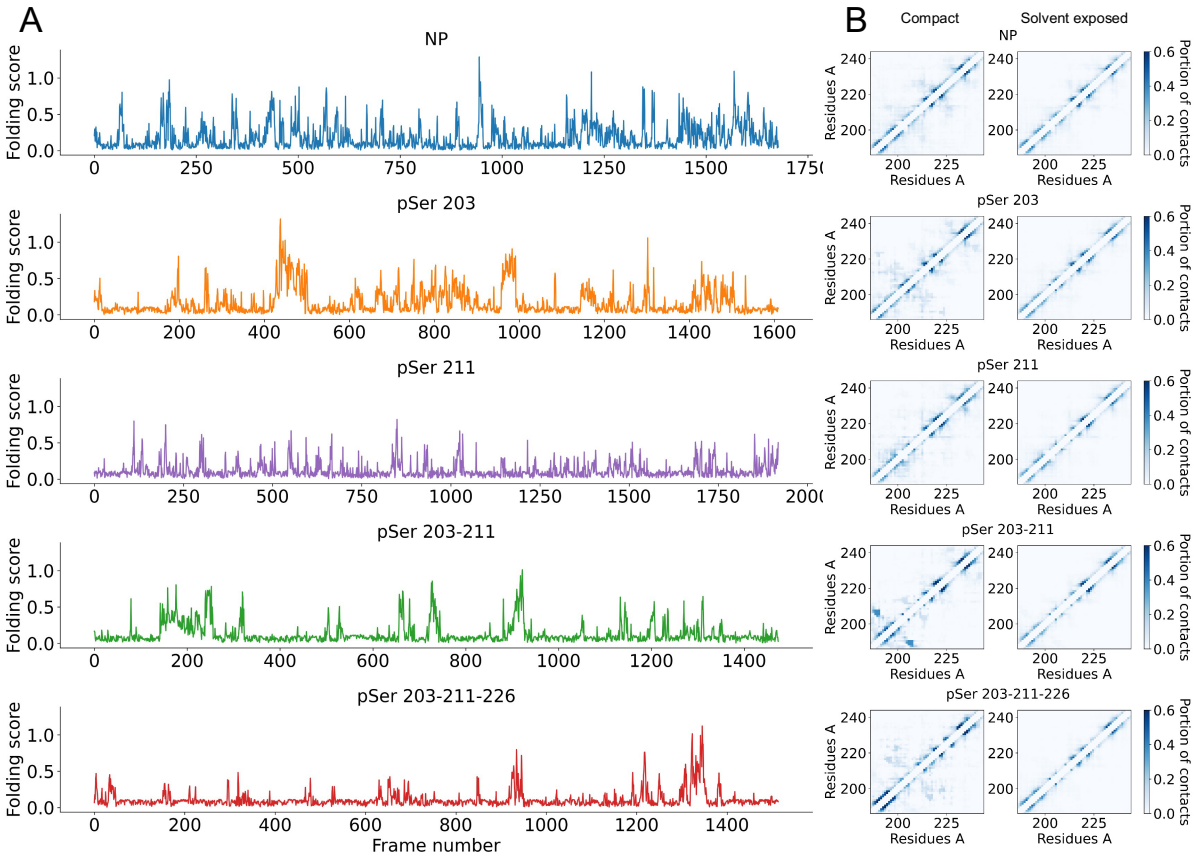


*Figure 5 Circuit topology identifies long-range contacts in disordered protein conformations. (**A**) CT Folding score dynamics during the simulation for different phosphorylation variants. (**B**) The averaged contact matrix for compact and solvent exposed protein conformations.*

The conformational kinetics is modulated by phosphorylation

Finally, we investigate the relationship between contact analysis and the dynamics of the CT Folding Score. From our contacts analysis, we have constructed energy landscapes for five phosphovariants of AF1c which were mapped to the first two principal components (Figure 3). This data suggests that NP and pSer 211 phosphovariants have one broad energy minimum, that includes solvent exposed and double helix conformations. Because these states are not separated by the energy barrier, protein structure could freely fluctuate between them. This observation is in line with the *CT Folding Score* kinetics of these variants, where we see the absence of long-living disordered compact states, while still frequently featuring relatively high peaks. In contrast, we see a clear separation for solution exposed and disordered compact states on the energy landscape of pSer 203 and pSer 203-211 variants. The pSer 203 has a distinct double helix state (Figure 3 top of the panel pSer 203). The pSer 203-211 has a double helix state (helix 1 and helix 2) and single helix 3 state, separated by an energy barrier. The kinetics of disordered compact state formation could be revealed with autocorrelation analysis (Figure 6A). The *CT Folding Score* autocorrelation of NP and pSer 211 is dominated by fast compaction events within the 10–30 ns time range, while the curve for pSer 203, pSer 203-211 and pSer 203-211-226 is shifted and shows slower (50 – 200 ns) compaction events. An interesting exception from these observations is the triple phosphorylated variant. As seen in

Figure 5A, this variant clearly visits a long-living disordered compact state, but on the energy surface it has only a single stable state that is disordered. This could originate from the fact that we used only the first two PCA principal components and these vectors appear not to contain information on the non-local contacts.

The dwell time distribution of the *CT Folding Score* showed phosphorylation-induced stabilization of disordered compact states. Folding score traces were transformed to dwell time histograms by applying a threshold that represents the compact state. Analysis of dwell time revealed a long-living component for decompaction of pSer 203, pSer 203-211 and pSer 203-211-226 molecules (Figure 6B), suggesting the stabilization of compact conformation in these cases. Interestingly, decompaction slows down more in these cases than compaction. This could relate to the charge of binding partners for positively charged residues, promoting the formation of strong interaction with phosphoserine residues (Figure 6C, D). The partial screening of the positively charged Lys 206 (in case of AF1c pSer 203,) and Arg 214 (AF1c pSer 203-211) caused by phosphorylation promotes change in the average contact map of the protein compact state (Figure 5B). The energy barrier in non-phosphorylated and pSer 211 cases could be easily crossed, but lowering of the ion concentration slows down the compaction kinetics and promotes the formation of the long living disordered compact states, in the same manner as in the case of multiple phosphorylations at native salt conditions (Figure S5). All these observations show the role of the electrostatic interaction balance on the disordered protein conformational ensemble and kinetics.

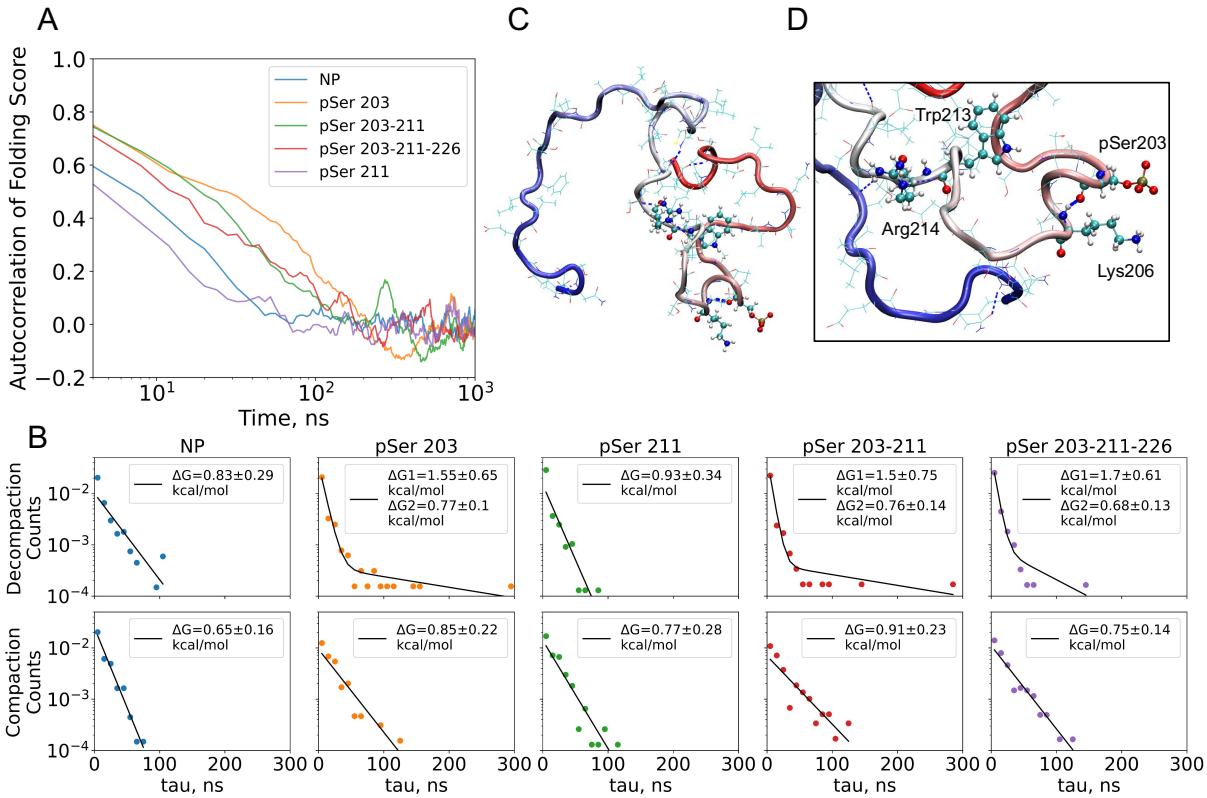


*Figure 6 Phosphorylation promotes formation of long-living compact states and alters distribution of protein structures. (**A**) Autocorrelation of CT Folding Score for different phosphovariants. NP and S211P show compaction events with 10-30 ns life times, while S203P, S203-211P and S203-211-226P show relatively slow (50 – 200 ns) compaction events. (**B**) Dwell-time histograms fitted with a single and double exponential decay function. The histograms were constructed from the times protein was in the disordered compact state (Decompaction) and solvent exposed state (Compaction), then counts were divided on the simulation time for the activation energy ΔG calculation. Here, we presented the histogram fits with single and double exponent function; double exponent fits*

*were done on Decompaction times for pSer203, pSer 203-211, and pSer 203-211-226. Slow exponent represents the dynamics of compact disordered states. These events are rare and have long durations, that characterize the form of the curve. (**C**) Disordered compact shape of S203P at frame 450 (CT Folding Score peak). Colors show the amino acid number: form red – N-terminal region to blue – C-terminal region. In this image, one can readily see that most of the protein is in direct contact with solvent, while there is a contact between N-terminal region and the middle of the protein (compare with contact map for S203P/compact in Figure 7B). (**D**) Close look at the helix formed by interaction of pSer203-Lys206. This helix is stabilized by charge interaction and backbone hydrogen bond formation (blue dashed line)*

Discussion

By performing more than 40 microseconds of atomistic unbiased MD simulations, this study demonstrates that the AF1c subunit of the GR and its phosphovariants are disordered and that site-specific phosphorylation modulates the kinetics and the probability of forming disordered compact states. NP and pSer 211 variants show many short-living compaction events. This could be attributed to the formation of weak contacts between distinct protein parts forming elongated (in terms of DSSP classification) secondary structure (Figure S3 A, C). In the case of pSer 203 and pSer 203-211 variants, we could observe stable elongated structure contacts more often, and these contacts are responsible for the majority of the compact, yet disordered states. There are interesting exceptions, however, such as a long living disordered folded state in pSer 203 between frames 400 and 500 (Figure S3 B, Figure 6 B C) and in pSer 203-211 between frames 900-950 (Figure S3 D). These states cannot be attributed to the elongated structure formation, but have a higher content of helical structure, including helix in 203-211 region in case of pSer 203. Solvent-exposed conformations are rarely followed by disordered compact conformations, stabilized by elongated structure or helix 1 and helix 2 interactions (Figure S3). The triple phosphorylated variant pSer 203-211-226 has a high propensity for helical structure formation (Figure 2C), and pSer 203-211-226 has the most solvent-exposed structure (Figure 2AB). Such a combination could explain the dynamics of this variant's *CT Folding Score*.

Experimental research has suggested that phosphorylation of Ser211 residue promotes folding of AF1c [18, 20, 31], which likely happens within timescales much longer (dozens of microseconds) than what our study probes. However, we have made observations that are in line with phosphorylation-induced folding (as observed in previous CD analysis [18, 20]). For example, we have observed helix formation of the 203-211 fragment after phosphorylation of Ser203 residue (Figure S3B, Figure 6D). At low salt conditions (25 mM NaCl, pSer 211) (Figure S5 D, H), pSer211-Pro212-Trp213-Arg214 formed stable helices in our simulations, in the same manner as observed earlier for pSer203-Pro204-Gly205-Lys206. These observations support the previous suggestion that pSer-Pro tends to be a helix initiator motif [32]. Thus, phosphorylation may be associated with stabilization of helical conformations by charge interaction and hydrogen bonds. These helix formation seeds could grow into well-defined helical structures and that eventually stabilize folds on long time scales.

The reversible phosphorylation of GR plays a key role in modulating the GR response by influencing its cellular localization, and it is regulated by the balance between the phosphorylating targeted kinases and dephosphorylating phosphatases that will interact with the receptor [2, 33]. Our study reveals the detailed effect of phosphorylation on the conformational diversity of AF1c, suggesting that GR adopts different topological dynamic states as it translocates from the cytoplasmic to the nuclear environment. This sheds new light

on how phosphorylation of Ser203 and Ser211 results in enhanced nuclear localization and subsequent increase of GR transcriptional activity as previously reported [2, 34]. The ability of AF1c phosphovariants to adopt multiple structural changes likely creates distinct binding surfaces within the NTD, enabling GR to selectively interact with different regulatory complexes along its signal transduction pathway. Interestingly, we have observed that pSer 203-211 has the highest propensity for secondary structure formation. This phosphovariant is localized in the perinuclear region and is believed to be the precursor of the active form of GR (pSer211), found in the nucleus [3-5]. In this context, it is noteworthy that pSer 211 and pSer 203-211 exhibit distinct structural propensities, indicating that different conformations may be needed to transition between the two phosphorylated states. Whereas Ser203 can be phosphorylated by both cyclin-dependent kinases (CDKs) and ERK kinases, p28 MAPK is the one that fully activates GR by phosphorylating Ser211 after ligand-binding, suggesting that these different structural states can influence the interaction of GR with different kinases [34]. On the other hand, the shift from one state to another is significantly influenced by the presence of phosphatases interacting with GR and reverting its modifications, which are also involved in the nuclear import of the ligand-activated GR [5]. Our results are in line with the experiments performed by Garza et al. 2010, who demonstrated that phosphorylation at Ser211 leads to significant structural changes in the GR AF1 domain, supporting the hypothesis that this modification induces a more ordered and functionally relevant conformation capable of interacting with several transcriptional coregulators [7]. We hypothesize that each phosphorylation-driven conformational state could favor the interaction of GR with specific partners by altering the accessibility of AF1c binding sites, thereby modulating its activity and transcriptional outcomes in a context-specific manner.

GR may be phosphorylated in more ways than this research covers. Zhen Wang *et al.* [22] reported observation of pSer211-pSer226, pSer203-pS226, and pSer226 phosphorylation variants. The AF1c pSer203-pS226 is characteristic of the perinuclear region. Of particular interest is the comparison of nucleus-localized AF1c pSer211, AF1c pSer226, and AF1c pS211-pSer226, which may have different transactivation activity. The listed combinations of phosphorylation could exhibit unique propensities of disordered compact state formation, which may be linked to their biological role. Combining all-atom MD simulations with CT analysis can readily be used to analyze these variants, which we plan for the follow up research.

The central result of this research is a new methodology for IDP analysis, which we demonstrated using the GR AF1 analysis as a proof-of-concept. The *CT Folding Score*, a circuit topology-based metric, calculates the normalized amount of series, parallel and cross types of contact relations. Using *CT Folding Score* we quantified the probability of disordered compact state formation, where the protein chain is solvent-exposed, yet certain residues form transient (non-local) contacts. In contrast to CT Folding Score analysis, characterization of such states by secondary structure content, *Rg* or *SASA* is not productive due to large structural variations and the incompleteness of these representations (Figure S4A). However, compact disordered states could be efficiently defined using the language of circuits. Figure S4B shows that protein conformations with higher *CT Folding Score* usually have a smaller Rg and SASA, independent from helical structure. Since it contains information about protein chain entanglement, the *CT Folding Score is* sensitive for detecting protein folding events. One can consider utilizing the *CT Folding Score* as a coordinate of folding. In future applications this metric could be used in meta-dynamics folding simulations [35]. Depending

on the application, users may refine the *CT Folding Score*, for example, by introducing different weights or normalizing it by maximum topological relations.

Fully or partly intrinsically disordered proteins account for a significant fraction of human proteome, yet our understanding of their conformational regulation is limited. It is known that IDPs often get stabilized by binding to their interacting partners, but this is not the only way that disordered to ordered transitions happen. Disordered compact states can be seen in isolated IDPs depending on their sequence, post-translational modification, and solvent condition [4, 36]. In contrast to the molten globule state, commonly seen as a transition state in the folding of stably folded proteins, disordered compact conformations (which we may also refer to as “disordered connected” conformations) do not contain a lot of secondary structure. In this way, they could be considered as precursors of a molten globule state. Despite structural instability, formation of compact disordered states is linked to function [5, 37]. Importantly, reversible protein phosphorylation provides a major regulatory mechanism in conformational dynamics and compaction of IDPs [15, 38], which can be effectively characterized using the topology-based approach developed in this study. Chains that form transient contacts at various time scales and with different interconnected chain regions can form “circuits” that feature both disorder and compactness simultaneously.

Methods

Molecular dynamics simulations

For conventional MD simulations of AF1c, initial structures were predicted with LocalColabFold program [39-41]. Because LocalColabFold does not recognize phosphoserine amino acids, it was replaced with glutamic acid in the input sequence of phosphorylated proteins. Then, after obtaining 3D-structures, these glutamic residues were replaced by the phosphoserines with the help of SwissSidechain v.2 PyMOL Plugin [42-46]. The structures with the highest quality scores for each phosphorylation variant were used as initial structures. Three different initial structures were selected for simulating each of the phosphorylation variants.

The MD simulations were performed using the free molecular simulation software GROMACS [47] version 2021.7 patched with the open-source, community-developed PLUMED library [48] version 2.8.3 [49]. The protein force field was amber99SBdisp [50, 51], with explicit solvent - amber99SBdisp water model, phosphoserine parameters were taken from [52]. The AF1c was solvated in dodecahedron box with the box vector lengths 9.5 nm 9.5 nm 9.5 nm and periodic boundariy conditions. The system was neutralized with 150 mM sodium and chloride ions. Before the production run, the system was minimized and equilibrated. The equilibration was done in NVT and then NPT ensembles for 100 ps, the heavy protein atoms were restrained by harmonic force with the default value for the constants being 1000 kJ mol^{-1} nm^{-2}. The LINCS algorithm [53] was used to constrain H-bonds vibrations. Integration was done with leap-frog algorithm, time step 2 fs. Buffered neighbor searching was done with Verlet list cutoff scheme, cutoff 1.0 nm in equilibration steps and 0.9 nm at the production step. Particle Mesh Ewald [54] with cubic interpolation was used for long-range electrostatics. In simulations, the pressure of 1 bar was controlled with a Parrinello-Rahman barostat, relaxation time 2 ps for equilibration steps and 5 ps for the production step. The isothermal compressibility of water was 4.5×10-5 bar^{-1}. The solute and solvent were coupled independently, the temperature was set for both on 310 K. During equilibration steps, a modified Berendsen thermostat was used with a time constant tau of 0.1 ps. In production runs, the temperature was controlled by a

Nose-Hoover thermostat, with time constant tau equal to 1 ps. The length of the simulation in production runs was 1000 ns for all systems. Eight replicates for each phosphorylation type were independently prepared and simulated. For analysis, the last 900 ns of simulations, with time steps of 500 ps were used. This part of the trajectory was considered as equilibrated. All frames for which the distance between the protein molecule and its periodic image was less than 2 nm were removed from the analysis. All trajectories for the same phosphorylation type were combined for the analysis. In this research, we focused on conventional MD; however, we also performed REST2 to ensure completeness. Our REST2 based simulation methodology is described in the Supplementary Information.

Data analysis

All analysis and plots preparation were done with homemade Python 3 scripts in the Anaconda based environments with help of different libraries: scipy [55], matplotlib [56], numpy [57], mdtraj [58], mdanalysis [59-61], pyemma [62], seaborn [63], pandas [64].

Calculation of NMR and CD spectra

The NMR shifts from simulated trajectories were calculated with SPARTA+ program [23]. The experimental NMR spectrum was taken from [24]. The NP sequence, used in this research, covers residues from 187 to 244 of GR, while in NMR research of *Kim et al.* the sequence ranges from 181 to 244 a.a.

The secondary structure was computed with the help of the DSSP program [28]. Protein structure images were generated with VMD package [65], (http://www.ks.uiuc.edu/Research/vmd/).

Circuit Topology

For circuit topology analysis, trajectories were subsampled with a 2 ns time step to speed up the analysis. Next, these frames were saved as pdb files and transformed to CT representation with the code ProteinCT [66]. This representation was then analyzed using newly developed code that can be found in GitHub repository: https://github.com/circuittopology/CT_Folding_Score

The *CT Folding Score* is defined as: $\sqrt{\left(\frac{S}{max(S)}\right)^2 + \left(\frac{P}{max(P)}\right)^2 + \left(\frac{X}{max(X)}\right)^2}$,

where S is – the number of series relations, P - is the number of parallel relations, X - is the number of cross relations between the contacts; and max is the maximum number of contacts relations observed in simulations of all phosphorylation variants of AF1c. A conformation is identified as “compact” if it has a *folding score* of more than 0.1. In this way, the folding score is a relative metric. Before classification, a Savitzky-Golay filter was applied to the *CT Folding Score* time series, to reduce signal noise.

Data and Code Availability Statement

The circuit topology-based fold analysis tool developed in this study is available at:
https://github.com/circuittopology/CT_Folding_Score

The simulation trajectories have been deposited in Zenodo repositories and can be found at:
Doi: 10.5281/zenodo.13820169
Doi: 10.5281/zenodo.13822438

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization & Project Administration: A.M. Investigation: V.A. Formal Analysis: V.A. Software: V.A. Methodology: V.A., A.M. Visualization: V.A., J.v.N., A.M. Writing & Draft Preparation: V.A., A.M. Writing-Review & Editing: V.A., A.J.P., E.E., J.v.N., A.M.

Acknowledgements

The authors would like to thank the technical support of Vahid Sheikhhassani from Medical Systems Biophysics and Bioengineering Laboratory, Leiden University. This work was performed using the Academic Leiden Interdisciplinary Cluster Environment (ALICE), as well as Snellius, the Dutch National supercomputer hosted at SURF (grant number: EINF-7466).

References

1. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding.* Biochim Biophys Acta, 2010. **1804**(6): p. 1231-64.
2. van der Lee, R., et al., *Classification of intrinsically disordered regions and proteins.* Chem Rev, 2014. **114**(13): p. 6589-631.
3. Chowdhury, A., D. Nettels, and B. Schuler, *Interaction Dynamics of Intrinsically Disordered Proteins from Single-Molecule Spectroscopy.* Annu Rev Biophys, 2023. **52**: p. 433-462.
4. Das, R.K. and R.V. Pappu, *Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues.* Proc Natl Acad Sci U S A, 2013. **110**(33): p. 13392-7.
5. Wang, D., et al., *The importance of the compact disordered state in the fuzzy interactions between intrinsically disordered proteins.* Chemical Science, 2022. **13**(8): p. 2363-2377.
6. Tompa, P., et al., *Intrinsically disordered proteins: emerging interaction specialists.* Curr Opin Struct Biol, 2015. **35**: p. 49-59.
7. Golovnev, A. and A. Mashaghi, *Generalized Circuit Topology of Folded Linear Chains.* iScience, 2020. **23**(9): p. 101492.
8. Scalvini, B., V. Sheikhhassani, and A. Mashaghi, *Topological principles of protein folding.* Phys Chem Chem Phys, 2021. **23**(37): p. 21316-21328.
9. Schullian, O., et al., *A Circuit Topology Approach to Categorizing Changes in Biomolecular Structure.* Frontiers in Physics, 2020. **8**.
10. Heidari, M., H. Schiessel, and A. Mashaghi, *Circuit Topology Analysis of Polymer Folding Reactions.* ACS Cent Sci, 2020. **6**(6): p. 839-847.
11. Scalvini, B., et al., *Circuit Topology Approach for the Comparative Analysis of Intrinsically Disordered Proteins.* J Chem Inf Model, 2023. **63**(8): p. 2586-2602.
12. Kadmiel, M. and J.A. Cidlowski, *Glucocorticoid receptor signaling in health and disease.* Trends Pharmacol Sci, 2013. **34**(9): p. 518-30.
13. Dieken, E.S. and R.L. Miesfeld, *Transcriptional transactivation functions localized to the glucocorticoid receptor N terminus are necessary for steroid induction of lymphocyte apoptosis.* Mol Cell Biol, 1992. **12**(2): p. 589-97.
14. Anbalagan, M., et al., *Post-translational modifications of nuclear receptors and human disease.* Nucl Recept Signal, 2012. **10**: p. e001.

15. Weikum, E.R., et al., *Glucocorticoid receptor control of transcription: precision and plasticity via allostery.* Nat Rev Mol Cell Biol, 2017. **18**(3): p. 159-174.
16. Almlof, T., A.P. Wright, and J.A. Gustafsson, *Role of acidic and phosphorylated residues in gene activation by the glucocorticoid receptor.* J Biol Chem, 1995. **270**(29): p. 17535-40.
17. Chen, W., et al., *Glucocorticoid receptor phosphorylation differentially affects target gene expression.* Mol Endocrinol, 2008. **22**(8): p. 1754-66.
18. Garza, A.M., S.H. Khan, and R. Kumar, *Site-specific phosphorylation induces functionally active conformation in the intrinsically disordered N-terminal activation function (AF1) domain of the glucocorticoid receptor.* Mol Cell Biol, 2010. **30**(1): p. 220-30.
19. Ismaili, N. and M.J. Garabedian, *Modulation of glucocorticoid receptor function via phosphorylation.* Ann N Y Acad Sci, 2004. **1024**: p. 86-101.
20. Khan, S.H., W.A. McLaughlin, and R. Kumar, *Site-specific phosphorylation regulates the structure and function of an intrinsically disordered domain of the glucocorticoid receptor.* Sci Rep, 2017. **7**(1): p. 15440.
21. Wang, Z., J. Frederick, and M.J. Garabedian, *Deciphering the phosphorylation "code" of the glucocorticoid receptor in vivo.* J Biol Chem, 2002. **277**(29): p. 26573-80.
22. Wang, Z., et al., *Modulation of glucocorticoid receptor phosphorylation and transcriptional activity by a C-terminal-associated protein phosphatase.* Mol Endocrinol, 2007. **21**(3): p. 625-34.
23. Shen, Y. and A. Bax, *SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network.* J Biomol NMR, 2010. **48**(1): p. 13-22.
24. Kim, D.H., A. Wright, and K.H. Han, *An NMR study on the intrinsically disordered core transactivation domain of human glucocorticoid receptor.* BMB Rep, 2017. **50**(10): p. 522-527.
25. 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, 2012. **109**(40): p. 16155-60.
26. Rieloff, E. and M. Skepo, *The Effect of Multisite Phosphorylation on the Conformational Properties of Intrinsically Disordered Proteins.* Int J Mol Sci, 2021. **22**(20).
27. Liu, P., et al., *Replica exchange with solute tempering: A method for sampling biological systems in explicit water.* Proceedings of the National Academy of Sciences, 2005. **102**(39): p. 13749-13754.
28. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
29. Dahlman-Wright, K., et al., *Structural characterization of a minimal functional transactivation domain from the human glucocorticoid receptor.* Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1699-703.
30. Buholzer, K.J., et al., *Multilayered allosteric modulation of coupled folding and binding by phosphorylation, peptidyl-prolyl cis/trans isomerization, and diversity of interaction partners.* J Chem Phys, 2022. **157**(23): p. 235102.
31. Kumar, R. and E.B. Thompson, *Role of Phosphorylation in the Modulation of the Glucocorticoid Receptor's Intrinsically Disordered Domain.* Biomolecules, 2019. **9**(3).
32. Hamelberg, D., T. Shen, and J.A. McCammon, *Phosphorylation effects on cis/trans isomerization and the backbone conformation of serine-proline motifs: accelerated molecular dynamics analysis.* J Am Chem Soc, 2005. **127**(6): p. 1969-74.
33. Vandevyver, S., L. Dejager, and C. Libert, *Comprehensive overview of the structure and regulation of the glucocorticoid receptor.* Endocr Rev, 2014. **35**(4): p. 671-93.
34. Miller, A.L., et al., *p38 Mitogen-activated protein kinase (MAPK) is a key mediator in glucocorticoid-induced apoptosis of lymphoid cells: correlation between p38 MAPK*

activation and site-specific phosphorylation of the human glucocorticoid receptor at serine 211. Mol Endocrinol, 2005. **19**(6): p. 1569-83.
35. Valsson, O., P. Tiwary, and M. Parrinello, *Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint.* Annu Rev Phys Chem, 2016. **67**: p. 159-84.
36. Marsh, J.A. and J.D. Forman-Kay, *Sequence determinants of compaction in intrinsically disordered proteins.* Biophys J, 2010. **98**(10): p. 2383-90.
37. Tesei, G., et al., *Conformational ensembles of the human intrinsically disordered proteome.* Nature, 2024. **626**(8000): p. 897-904.
38. Iakoucheva, L.M., et al., *The importance of intrinsic disorder for protein phosphorylation.* Nucleic Acids Res, 2004. **32**(3): p. 1037-49.
39. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat Methods, 2022. **19**(6): p. 679-682.
40. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
41. Evans, R., et al., *Protein complex prediction with AlphaFold-Multimer.* 2022.
42. David Gfeller, O.M.a.V.Z., *Expanding molecular modeling and design tools to non-natural sidechains.* Journal of Computational Chemistry, 2012. **33**: p. 1525-1535.
43. Gfeller, D., O. Michielin, and V. Zoete, *SwissSidechain: a molecular and structural database of non-natural sidechains.* Nucleic Acids Res, 2013. **41**(Database issue): p. D327-32.
44. Schrodinger, LLC, *The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8*. 2015.
45. Schrodinger, LLC, *The JyMOL Molecular Graphics Development Component, Version 1.8*. 2015.
46. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
47. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1-2**: p. 19-25.
48. Bonomi, M., et al., *Promoting transparency and reproducibility in enhanced molecular simulations.* Nature Methods, 2019. **16**(8): p. 670-673.
49. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird.* Computer Physics Communications, 2014. **185**(2): p. 604-613.
50. Robustelli, P., S. Piana, and D.E. Shaw, *Developing a molecular dynamics force field for both folded and disordered protein states.* Proc Natl Acad Sci U S A, 2018. **115**(21): p. E4758-E4766.
51. Piana, S., et al., *Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes.* J Chem Theory Comput, 2020. **16**(4): p. 2494-2507.
52. Steinbrecher, T., J. Latzer, and D.A. Case, *Revised AMBER parameters for bioorganic phosphates.* J Chem Theory Comput, 2012. **8**(11): p. 4405-4412.
53. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
54. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
55. Virtanen, P., et al., *SciPy 1.0: fundamental algorithms for scientific computing in Python.* Nat Methods, 2020. **17**(3): p. 261-272.
56. Hunter, J.D., *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 2007. **9**(3): p. 90-95.
57. Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.
58. McGibbon, R.T., et al., *MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories.* Biophys J, 2015. **109**(8): p. 1528-32.

59. Alibay, I., et al., *Building a community-driven ecosystem for fast, reproducible, and reusable molecular simulation analysis using mdanalysis.* Biophysical Journal, 2023. **122**(3): p. 420a-420a.
60. Naughton, F.B., et al., *MDAnalysis 2.0 and beyond: fast and interoperable, community driven simulation analysis.* Biophysical Journal, 2022. **121**(3): p. 272a-273a.
61. Michaud-Agrawal, N., et al., *Software News and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations.* Journal of Computational Chemistry, 2011. **32**(10): p. 2319-2327.
62. Scherer, M.K., et al., *PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models.* J Chem Theory Comput, 2015. **11**(11): p. 5525-42.
63. Waskom, M.L., *seaborn: statistical data visualization.* Journal of Open Source Software, 2021. **6**(60): p. 3021.
64. McKinney, W.a.o., *Data structures for statistical computing in python*. Proceedings of the 9th Python in Science Conference. Vol. 445. 2010, Austin, TX. 51--56.
65. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
66. Moes, D., et al., *ProteinCT: An implementation of the protein circuit topology framework.* MethodsX, 2022. **9**: p. 101861.