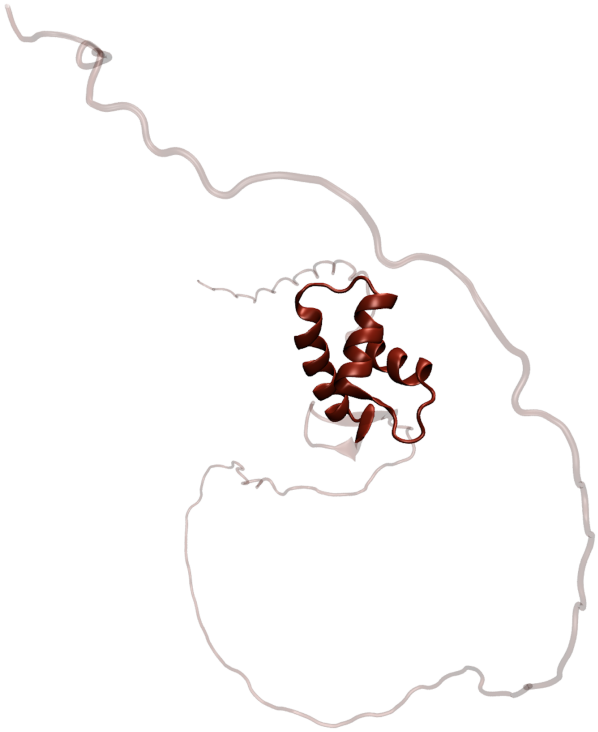
UNIVERSITY OF COPENHAGEN
DEPARTMENT OF BIOLOGY



BSc in Biochemistry

Frederik Espersen Knudsen

Functional binding with variants of the H1.0 intrinsically disordered C-terminal domain

Supervised by Kresten Lindorff-Larsen and Francesco Pesce

The University of Copenhagen, 16th of June 2023

Abstract

A key assumption in evolution is the conservation of function. In the context of proteins, this often interprets as the conservation of structure through primary sequence. Intrinsically disordered regions (IDR) complicate this notion as they are best described by a dynamic structural ensemble and exhibit low sequence identity. While amino acid content has been proposed as an alternative conserved feature in IDRs, it is insufficient for explaining structural properties and function on its own, especially in high-charge sequences. This thesis employs coarse-grained molecular dynamics simulations to study the effect of charge patterning in the intrinsically disordered C-terminal domain (CTD) of H1.0 on its conformational properties and high-affinity binding to its intrinsically disordered chaperone prothymosin α (ProTα). A general sequence and structural profile of the H1.0 CTD is derived from a set of 189 orthologs, which characterises it as a highly charged and expanded chain with a conserved composition. An evolution algorithm is employed to generate sequence variants with identical amino acid composition but varying dimensions and charge patterning. The effect of charge patterning on functional binding of the H1.0 CTD to ProTα is assessed by comparing the wild type complex to that of complexes with generated CTD variants, demonstrating that a homogenous charge distribution leads to lower binding affinity. Examining the effect of post-translational modifications or assessing other functional binding partners of the H1.0 CTD may reveal further insights into the evolutionary constraints on H1.0 CTD charge patterning.

Frederik Espersen Knudsen
Functional binding with variants of the H1.0 intrinsically disordered C-terminal domain
BSc in Biochemistry, June 2023
Supervised by Kresten Lindorff-Larsen (main) and Francesco Pesce

University of Copenhagen
Faculty of Science
Section for Biomolecular Sciences
Ole Maaløes Vej 5
DK-2200 Copenhagen N

Acknowledgements

I would like to thank Kresten Lindorff-Larsen for providing me with the supervision, ressources, and extraordinary settings to conduct this project.

I would especially like to thank Francesco Pesce for the day-to-day support, discussion, guidance, and counseling, as well as for his ideas for the contents and goals of this project.

I would also like to thank Giulio Tesei, my deskmates Arriën Symon Rauh, Emil Thomasen, and Johannes Betz, as well as the rest of the Lindorff-Larsen group for their support and inputs throughout the project. While several of its contributers have already been mentioned, I would like to give a special thanks to those who have contributed to develop *CALVADOS*, which has been at the bedrock of this project.

Lastly, I would like to thank The University of Copenhagen (UCPH), the Danish e-Infrastructure Cooperation (DeiC), and the Centre for Scientific Computing Aarhus (CSCAA) for providing me access to their computational ressources, without which many of this project's endeavours would not have been realised.

Abbreviations

Terms		**Measures**	
IDP	Intrinsically disordered protein	κ	Charge patterning
IDR	Intrinsically disordered region	R_g	Radius of gyration
H1.0	Histone H1.0	Δ	Asphericity
ProTα	Prothymosin α	S	Prolateness
CTD	C-terminal domain	ν	Flory scaling exponent
MD	Molecular dynamics	K_d	Dimer dissociation constant

Table of Contents

1 Introduction

A fundamental assumption in evolution is the conservation of function. As the function of a protein is governed by its structure, the structure is often thought to be the principal attribute conserved during protein evolution.

Intrinsically disordered proteins (IDPs) complicate this dogma. They challenge the notion that a single protein sequence dictates a single protein structure as they, or rather their intrinsically disordered regions (IDR), are best described by a highly dynamic ensemble of conformations. IDRs play crucial and varied roles in proteins, serving as flexible linkers, undergoing disorder-to-order transitions upon binding, or participating in dynamic complexes [1].

It has been proposed that IDRs evolve at a faster rate than structured regions, based on their lower sequence identity [1]. IDRs do, however, tend to have a relatively well-conserved overall amino acid composition, i.e., frequency of amino acids [1]. This has lead to the idea that amino acid composition, rather than the primary sequence, plays the primary role in maintaining the structural properties and functions of IDRs [1, 2].

The main premise underlying the idea that composition, rather than sequence, is the evolutionarily constrained characteristic in IDRs, is that IDRs with different primary sequences, but similar compositions, are able to achieve comparable structural and functional properties. This study examines that premise by exploring the extend to which the a conserved composition can accommodate variable structure and function via charge patterning, by using the intrinsically disordered C-terminal domain of histone H1.0 as a model IDR.

1.1 H1 linker histones

H1 linker histones are a family of small (∼200 residues) positively charged proteins. They bind to nucleosomes and linker DNA and mediate the formation of chromatin secondary structure, called chromatosomes [Fig. 1A], and the condensation of chromatin fibers through nucleosome oligomerisation [3–5]. Through this, they play important roles in fundamental cell processes like transcriptional regulation, DNA replication and repair, and epigenetics [3–5].

All H1 proteins share a tripartite structural division between a central helix-turn-helix folded domain (H15) and two intrinsically disordered regions situated in either terminus [Fig. 1B] [3–5]. The ∼110 residues disordered C-terminal domain (CTD) is reported to play a vital role in H1 and to have a

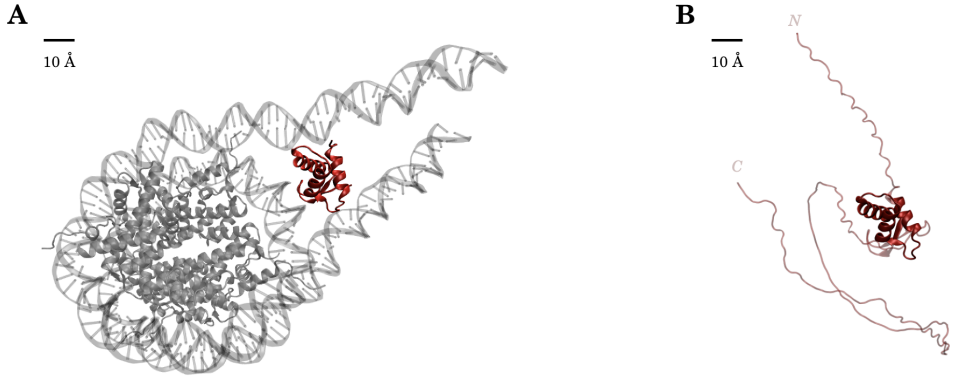


Figure 1 **A** The position of H1.0 (red) in a chromatosome (Orthographic cryo-EM structure excluding disordered regions). **B** The central folded H15 domain and the disordered tails (opaque) of H1.0 (Orthographic AlphaFold2 structure).

multitude of functions [3–5]. It has been reported that deletion of the CTD nearly eliminates the ability of H1 to bind to chromatin [3] and that the CTD itself is the binding site for an array of nuclear regulator proteins [4].

Chromatin modelling is critical for all eukaryotes, and with their intrinsic role herein, H1 histones should presumably be under strict evolutionary constraints. But while the structured domains of the histones are well-conserved [2, 6], the primary sequence of their intrinsically disordered tails has been reported to be less conserved across both para- and orthologs [2–4]. Their amino acid composition, on the other hand, is well-conserved, and many of the paralogous H1 CTDs can perform identical functional roles [2]. It has therefore been speculated that the ability of the CTD to specifically recognize and interact with many different types of macromolecules and structures may be in part credited to its intrinsic disorder and conserved amino acid composition [4, 7, 8].

1.2 Amino acid composition in the disordered C-terminal domain of H1.0

The H1 CTD composition is dominated by Lys (∼40%), Ala, and Pro, while being near devoid of the negatively charged residues Asp and Glu and highly hydrophobic residues like Phe, Tyr, and Trp [3, 4]. The abundance of Lys accounts for the high positive charge of the CTD, which is believed to help both DNA-binding and shielding of the negative charges of the DNA backbone for chromatosome compaction [4].

In a series of studies by Lu & Hansen [2, 7, 8] the importance of composition over primary sequence in the CTD of H1.0 has been elucidated. In their 2004 study [7], they divided the CTD into four equal subdomains, each with different primary sequences but similar composition. By testing different deletion constructs, they established that only two of the subdomains contributed to the ability to alter linker DNA conformation and chromatin condensation.

While the CTD is intrinsically disordered in isolation, previous studies have reported regions that undergo a disorder-to-order transition upon binding to the DNA minor groove [9–12]. Not coincidentally, these regions coincide with the functional subdomains identified by Lu et al. [7]. One group reported that the CTD forms secondary structure corresponding to 24% α-helix, 25% β-structure, 17% open loops, and 33% turns upon binding to DNA under physiological conditions [11], yielding some perspective on the intrinsic disorder of the CTD.

In a 2009 follow-up study [8], Lu et al. found that they could not only interchange the different subdomains but also randomly scramble the sequences of the subdomains without any significant functional effect. They could not, however, change the overall composition even sligthly without significant loss of function. Instead they propose that the functional roles of specific subdomains was a question of chain positioning relative to DNA, enforced by the folded domain, rather than primary sequence [7, 8].

While this demonstrates the primacy of composition in the context of chromatin condensation, one should note that other functions of the H1.0 CTD may require certain sequence motifs. It is well-documented [3, 4, 13] that chromatin condensation is modulated via phosphorylation of the H1 CTD by cyclin-dependent kinase (CDK), which phosphorylates key Ser and Thr residues in conserved consensus sequences. These phosphorylations are thought to disrupt charge interactions with the DNA backbone and to sterically enforce secondary structure that may spatially impede Lys residues from forming stabilising interactions with the linker DNA minor groove [3, 6]. It has been proposed that the observed α-helical structures during DNA-binding are dictated by conserved DNA-binding binding motifs that utilise the many Pro residues for helical initiation [6, 14]. Additionally, specific recognition of and interaction with other protein partners may require specific sequence motifs [4].

1.3 The interaction between H1.0 and its chaperone ProTα

While modulating chromatin structure is certainly one of the prime roles of H1.0, it has several other functional partners beyond the linker DNA of chromatosomes. One of these is its intrinsically disordered chaperone prothymosin α (ProTα) [15]. ProTα is of the same sequence length as the H1.0 CTD, but is conversely very negatively charged.

In a 2018 study, Borgia et al. [15] established that H1.0 and ProTα forms a very high affinity complex. They found that most interactions occur between ProTα and the CTD of H1.0, as was demonstrated by the fact that an isolated CTD construct binds just as well as the full length protein, but that a CTD-deletion leads to significantly lower affinity. Intriguingly, they demonstrated that, while both ProTα and H1.0 are IDPs, no structure forms during complex formation and the interaction between them is intrinsically disordered itself.

The bulk of the affinity of the complex has been attributed to the high amount of opposite charges between H1.0 and ProTα. Both the negative charge of ProTα and the high binding affinity reflects the that fact that ProTα competes for H1.0 binding with the abundant amount of negatively charged nuclear DNA.

The study proposed that this type of highly dynamic complex does not require structurally defined binding sites or specific persistent interactions between individual residues. Instead, intermolecular distance maps illustrated that interactions are broadly distributed along the sequences at charge-dense regions of the two peptides, while single-molecule FRET experiments demonstrated that no site-specific interactions or chain alignments seemed to persist, despite higher electrostatic attraction in charge-rich regions.

While this initially presents yet another example of an H1.0 function that appears to be dictated by composition rather than by the primary sequence, other factors beyond composition may as, if not more, important. The fact that interactions are distributed throughout the sequences indicates that, despite a conserved composition, the linear distribution of charge in the sequences has a key role to play in IDP binding.

1.4 The influence of charge patterning in IDRs

While the meaning of charge patterning can vary by context, in the topic of IDRs it often refers to whether or not charged residues are distributed in a homogeneous or heterogeneous manner across the primary sequence. To evaluate this, this study employs the κ-parameter [16], which describes charge asymmetry segment-wise across a linear protein sequence (see Methods: *Charge patterning*).

IDP Sequences with a dominant net charge and a high frequency of charged residues (FCR) are referred to as polyelectrolytes [16], and it has been found that charge patterning, as described by κ, is well-correlated with their dimensions. The lower the κ, the more evenly distributed are the charged residues along the sequence, and the larger the dimensions of the IDR tend to be. In regards to structural measures (see Methods: *Structural measures*), high FCR and κ commonly results in a relatively larger radius of gyration R_g, which reflects physical chain size, and larger Flory scaling exponents ν, which reflects how compacted or expanded a chain is [16, 17].

1.5 Investigating intrinsically disordered proteins with molecular dynamics

Both the H1.0 CTD itself and its interaction with ProTα provide opportune phenomena for assessing the extent to which charge patterning can accommodate structural and functional roles within the same amino acid composition.

To assess the H1.0 CTD structurally in a large scale manner, this study utilises coarse-grained molecular dynamics simulations of peptides. In such simulations, each residue is represented by a single bead, and peptides as strings of beads. Molecular interactions are modelled by a set of potentials that act as

approximations for the complex quantum physics that dictates atomic behaviour. In each step of the simulation, beads are moved according to classical mechanics along the gradients of the model potentials, simulating the dynamics of the peptides. In the end, a simulation yields what is known as a trajectory describing the position of all the particles of the simulated system at each sampled time frame, which can be used for later analysis.

Molecular dynamics is especially useful in a case such as this, where the subjects are intrinsically disordered peptides, since molecular simulations offers a tangible ensemble of chain structures to analyse. The approximate approach of coarse-grained simulations significantly downscales the computational effort required and enables one to sample peptides, timescales, and molecular events in bulk. A prerequisite for using simulations is that they should reflect reality; that is, the results they provide should agree with real-life observations. Most often, the challenge involved herein is to choose a proper set of potentials to model interactions.

This study employs the *CALVADOS* 2 model [18, 19] to simulate the IDPs, as it has been demonstrated to accurately predict conformational properties of IDPs. This model utilises two potentials: A salt-screened Debye-Hückel potential to represent charge-charge interactions and an Ashbaugh-Hatch potential to represent short-range interactions. The Ashbaugh-Hatch potential includes two sets of residue-specific parameters, one of which has been parameterised with experimental data on IDPs and hydrophobicity scales from the literature.

Through coarse-grained molecular dynamics simulations and artificial sequence evolution, this project demonstrates the importance of charge patterning for IDR structure as well as for the functional binding of the H1.0 CTD to ProTα.

2 Methods

Curating H1.0 CTD orthologs H1.0 orthologs were queried from OrthoDB [20] from the Eukaryota phylogenetic partition group (OrthoDB group ID: 5363206at2759).

485 orthologs were identified in OrthoDB, of which 195 returned UniProt records. 3 records were filtered off on the criteria that the UniProt description had to contain "H1" or "H5" - the names of the linker histone families. Of the remaining 192 records, 189 returned C-terminal domains that were classified as intrinsically disordered by *MobiDB* [21]. The UniProt *MobiDB* disorder annotations were used to extract the H1.0 ortholog CTD sequences.

Reference IDRs Throughout the paper, a previously curated dataset of IDR orthologs by Tesei et al. [17] is used as reference, such as for IDR Flory scaling exponents and amino acid frequencies. This dataset has been aggregated for this study such that each row represents an ortholog family. It is from this granularity that means and quantiles are evaluated.

Charge patterning The κ measure [16] calculated using *localCIDER* [22], was used to assess charge patterning in peptide sequences. It is defined as such:

If f_+ and f_- are the fractions of positive and negative charge respectively, then charge asymmetry σ_i in a sequence segment i is defined as:

$$\sigma_i = \frac{(f_+ - f_-)^2}{f_+ + f_-} \tag{1}$$

The charge asymmetry for each segment σ_i of size $g \in \{5, 6\}$ is calculated, and the segment-wise squared deviation δ from the global sequence charge asymmetry σ is defined as:

$$\delta = \frac{1}{N} \sum_{i=1}^{N} (\sigma_i - \sigma)^2 \tag{2}$$

where N is the number of segments in the sequence given the segment length g. δ is a sequence-specific value, and the maximum δ_{max} is the one for a sequence with terminally clustered charges (i.e. positive in one end, negative in the other). κ is calculated as the sequence-specific δ normalised by δ_{max}:

$$\kappa = \frac{\delta}{\delta_{max}} \tag{3}$$

and is thus bounded by $0 \leq \kappa \leq 1$, where 0 corresponds to uniform charge distribution and 1 corresponds to asymmetric charge distribution. The final κ is calculated as an average of κ-values for segment lengths $g = 5$ and $g = 6$.

Sequence identity Sequence identity i was computed from multiple sequence alignments with a measure inspired by Bellay et al [23]. Alignments were performed using the MUSCLE algorithm in *Jalview* with default settings.

To calculate sequence identity from an alignment, the frequency $f_j(aa)$ of every amino acid aa was computed for each position j in the alignment (Note that a gap position '-' is also a valid amino acid). The highest amino acid frequency $F_j = \max(f_j(aa))$ was chosen in each position, essentially generating a consensus sequence. For each position, the frequency of conforming to the consensus sequence is thus F_j. The sequence identity i was defined as the mean of F_j across the sequence:

$$i = \frac{1}{N} \sum_{j=1}^{N} F_j \tag{4}$$

Positions where the consensus amino acid was a gap position '-' were, however, excluded from the mean. This was done to correct for a few long outlier ortholog sequences that created long gaps in alignments. Filtering off gap positions yielded a consensus sequence whose length was in agreement with the mean sequence length of the alignment sequences. See Appendix F for examples of position-wise identity scores.

Single-chain simulations Single-chain molecular dynamics simulations were run with a sequence and chemiophysical conditions as input. The simulations were run with OpenMM using the*CALVADOS* 2 model (see below).

An arbitrary topology was generated from the input sequence. Simulation box side length were either 200 nm, 100 nm, or 25 nm. Simulation times ranged between 100 ns and 100 µs with 10 fs timesteps. Simulations were run with periodic boundary conditions in a NVT ensemble using a Langevin integrator with a friction coefficient of 0.01 ps^{-1}. Frames were sampled with a stride of 3000 timesteps, and the first 1000 frames were removed from each trajectory to account for equilibration. A summary of simulation conditions are found in Appendix A.

Two-chain simulations Two-chain simulations were run like single-chain simulations, but with a two-chain topology as input. Topologies were generated by combining each of the most compact structures (by R_g) from sample trajectories of the H1.0 CTD and of ProTα. The structures were transposed 10 nm apart and collectively centered.

CALVADOS model Coarse grained molecular dynamics simulations of peptides were run using the-*CALVADOS* 2 model by Tesei et al. [18, 19]. The model is developed to accurately predict conformational and LLPS properties for IDP sequences of diverse length and charge patterning, and under different salt and temperature conditions. It is parameterised against experimental data and residue-specific hydrophobicity scales from the literature.

The model represents peptides as C_α-centered residue beads connected through harmonic restraints:

$$u_{bond}(r) = \frac{1}{2}k(r - r_0)^2 \tag{5}$$

using a force constant $k = 8033\ kJ\ mol^{-1}\ nm^{-2}$ and equilibrium distance $r_0 = 0.38nm$. Non-neighbouring beads interact through an Ashbaugh-Hatch potential (with a 2 nm cutoff):

$$u_{AH}(r) = \begin{cases} u_{LJ}(r) + \epsilon(1-\lambda), & r \leq 2^{1/6}\sigma \\ \lambda\ u_{LJ}(r), & r > 2^{1/6}\sigma \end{cases} \tag{6}$$

where $\epsilon = 0.8368\ kJ\ mol^{-1}$ is the potential well depth and u_{LJ} is the Lennard-Jones potential:

$$u_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{7}$$

where λ and σ are averages $\lambda = \frac{\lambda_i+\lambda_j}{2}$, $\sigma = \frac{\sigma_i+\sigma_j}{2}$ of amino acid-specific parameters that quantify hydrophobicity and size, respectively. Salt-screened electrostatic interactions between non-neighbouring beads are modeled via a Debye-Hückel potential (with a 4 nm cutoff):

$$u_{DH} = \frac{q_i q_j e^2}{4\pi\epsilon_0\epsilon_d r}\exp(-r/D) \tag{8}$$

where q is the average amino acid charge number, e is the elemental charge, ϵ_0 is the vacuum permittivity, ϵ_d is the dielectric constant, r is the interresidue distance, and D is the Debye length. The dielectric constant is calculated as:

$$\epsilon_d = 5321\ T'^{-1} + 233.76 - 0.9297\ T' + 0.1417 \cdot 10^{-2}\ T'^2 - 0.8292 \cdot 10^{-6}\ T'^3 \tag{9}$$

where T' is the unitless absolute temperature $T' = T/^\circ K$. The Debye length is calculated as:

$$D = \sqrt{1/(8\pi B c_s)} \tag{10}$$

where B is the Bjerrum length and c_s is the ionic strength. The Bjerrum length is calculated as:

$$B = \frac{e^2}{4\pi\epsilon_0\epsilon_r}\frac{N_A}{RT} \tag{11}$$

where N_A is Avogadro's constant, R is the ideal gas constant, and T is the absolute temperature.

The Ashbaugh-Hatch and Debye-Hückel potentials were offset such that $u(r) = 0$ for their cutoff distances.

Structural measures Measures were calculated across structural ensembles. $\langle\ \rangle$-annotation indicates a trajectory average.

The *center of mass* $\mathbf{r}_{CoM}$ is used as a measure of general peptide chain position. It is calculated frame-wise as:

$$\mathbf{r}_{CoM} = \frac{1}{M}\sum_{k=1}^{N} m_k \mathbf{r}_k \tag{12}$$

where M is the total mass, N is the total number of particles in the system, and m_k and $\mathbf{r}_k$ are the mass and position of each particle k respectively.

The *gyration tensor* $\mathbf{Q}$ is a matrix used for calculating other structural measures. It is calculated frame-wise according to Aronovitz & Nelson [24]:

$$Q_{ij} = \frac{1}{N}\sum_{k=1}^{N}(r_i - \overline{r}_i)(r_j - \overline{r}_j) \tag{13}$$

where r_i and r_j are each one of the three Cartesian components x, y, and z of each particle in the simulation. The coordinates r are mass-weighted, such that $\mathbf{r}_{CoM} = (\overline{r}_x, \overline{r}_y, \overline{r}_z)$. For computations, it also exists as a trace-less form:

$$\hat{\mathbf{Q}} = \mathbf{Q} - \mathrm{tr}(\mathbf{Q}) \tag{14}$$

The *radius of gyration* R_g is used as a measure of general peptide size. It is calculated frame-wise from the gyration tensor:

$$R_g = \sqrt{\mathrm{tr}(\mathbf{Q})} \tag{15}$$

The *asphericity* Δ is used as a measure of spherical asymmetry of peptides. It is calculated from the gyration tensor:

$$\Delta = \frac{3}{2}\frac{\left\langle \mathrm{tr}(\hat{\mathbf{Q}}^2)\right\rangle}{\langle(\mathrm{tr}(\mathbf{Q}))^2\rangle} \tag{16}$$

bounded by $0 \leq \Delta \leq 1$, where 0 corresponds complete spherical symmetry and 1 to complete spherical asymmetry.

The *prolateness* S is used as a measure of prolateness/oblateness - i.e. whether the structure is generally pill-shaped/disk-shaped. It is calculated from the gyration tensor:

$$S = \frac{27 \left\langle \det(\hat{\mathbf{Q}}) \right\rangle}{\langle (\mathrm{tr}(\mathbf{Q}))^3 \rangle} \tag{17}$$

bounded by $-\frac{1}{4} \leq S \leq 2$, where $-\frac{1}{4}$ to 0 corresponds to a more oblate structure and 0 to 2 corresponds to a more prolate structure.

The *Flory scaling exponent* ν is used as a measure of peptide chain compactness/expandedness. It is found by fitting the polymer scaling law:

$$\sqrt{\langle {r_{i,j}}^2 \rangle} = r_0 \ |i-j|^{\nu} \tag{18}$$

to all trajectory-averaged interresidue distances r between residue positions i, j in a peptide chain. Only residue pairs where $|i-j| > 10$ are used for fitting. A fixed value $r_0 = 0.518$ nm, determined as the best fit for across all H1.0 CTD orthologs, is used for calculating all scaling exponents.

Sequence evolution A Monte Carlo directed evolution algorithm, provided by Pesce et al. [25], was used to generate new H1.0 CTD sequences with the same overall amino acid content as the wild type.

The algorithm takes an input sequence and generates variants with the same amino acid content. In each generation, it generates a new sequence by swapping the position of two random residues in the previous sequence. An observable o is calculated, either R_g (by simulation or reweighthing) or κ, that is used to either accept or reject the new sequence. The probability $p_{i,j}$ to accept the new sequence j over the previous sequence i is calculated according to the Metropolis-Hastings criterion against a set target observable value o_T:

$$p_{i,j} = \min \begin{cases} \exp\left(\frac{|o_j - o_T| - |o_i - o_T|}{c}\right) \\ 1 \end{cases} \tag{19}$$

where c is a control parameter that tunes the strictness of acceptance. Thus, sequences with observables closer to the target will *always* be accepted, whereas sequences further away from the target only *may* be accepted. The control parameter c is lowered by 1% every $2N$ generations, where N is the length of the sequence.

Selecting representative sequences from clusters Sets of variants generated by the evolution algorithm were clustered based on either their $\langle R_g \rangle$, κ, or both using the `sklearn.cluster.KMeans` class from *SKlearn* [26].

For each cluster, a representative was chosen based on the shortest distance from the fitted centroid value. This is trivial when only one feature is used for fitting. However, in the case of fitting on both $\langle R_g \rangle$ and κ, the procedure was as follows: first, all $\langle R_g \rangle$- and κ- distances was calculated; next, each set of distances was scaled by the standard deviation of the set; and lastly, all distances were squared and summed for each variant. The variant with the lowest summed squared distance was chosen as the representative of the cluster.

Dimer dissociation constant Inspired by Borgia et al. [15], the dissociation constant K_d between H1.0 variants and ProTα was calculated as:

$$K_d^{-1} = 4\pi N_A \int_0^{r_{max}} r^2 \mathrm{e}^{-pmf_{eff}(r)/RT} \ dr \tag{20}$$

where N_A is Avogadro's number, R the ideal gas constant, T the absolute temperature, r the difference in center of mass between the H1.0 and the ProTα peptides, and $pmf_{eff}(r)$ the effective potential mean force between the two peptides.

The effective potential mean force $pmf_{eff}(r)$ is calculated as:

$$pmf_{eff}(r) = pmf(r) + 2RT\ln(r) \tag{21}$$

where $pmf(r)$ is the potential mean force, corresponding to the interaction energy, between the two peptides and $2RT\ln(r)$ is an adjustment for entropy.

The interaction energy between the two peptides was determined according to the *CALVADOS* model, and the energy as a function of center of mass distance was determined by binning the center of mass differences ($N_{bins} = 100$), calculating the bin average energy for $pmf(r)$, and using the bin centres for r.

The energy in the bins had to be corrected for the energy offsets in the *CALVADOS* model, such that long distances (where no contacts formed) were set to have $pmf_{eff}(r_{\max}) = 0$. The definite integral was calculated using the `scipy.integrate.simpson` function from *SciPy* [27] (Fig. 7 provides an example of the energy landscape that is integrated). For a better fit against experimental data, the effective mean force potential $pmf_{eff}(r)$ was scaled down by a unitless factor of 2.48 before evaluating the K_d expression.

This method relies on a fully sampled energy landscape of the center of mass distance, which requires at least one unbinding event to occur for the calculated K_d to be reliable. An unbinding event is defined to occurr when the shortest distance between the two chains becomes greater than 4 nm, at which point no interactions can occur due to the distance cutoffs of the *CALVADOS* model.

3 Results

This study seeks to assess the importance of charge patterning and structural properties for the binding of the H1.0 CTD to its chaperone ProTα. This is achieved by simulating ProTα-binding with a set of artificial variants of the H1.0 CTD, which share its composition, but have been sampled across ranges of sequence and structural features.

To confidently be able to assess binding between the artificial variants and ProTα, complex formation of the two wild type human proteins is simulated. This is compared to experimental *in vitro* data to evaluate the ability of the simulations to accurately capture the interaction.

Artificial variants of the H1.0 CTD are generated using an evolution algorithm, which is constrained to achieve certain target measures, such as a specific size or charge patterning, without changing the sequence composition. The algorithm can generate thousands of variants, from which representative sets can be sampled. It requires, however, an initial input sequence.

For the input sequence to reflect the general profile of the H1.0 CTD, conservation in 189 H1.0 CTD orthologs is assessed, and the findings used to construct a consensus input sequence.

Initially, however, the general characteristics of the H1.0 CTD are elucidated by assessing sequence and structural measures of the wild type human ortholog. To get a general sense of the importance of charge patterning, two artificial variants of the human ortholog, one randomly scrambled and one with clustered charges, are also investigated. From there, preparations for simulating variant binding with ProTα begins.

3.1 Characterising the human H1.0 C-terminal domain

The CTD of human H1.0 is a 111 residue long IDR that mainly consists of Lys (38.7%), Ala (17.1%) and Pro (10.8%) [Appendix B (a)]. The CTD is highly positively charged. 43% of its residues are charged with 43 Lys and 3 Arg as opposed to only 2 Glu and 1 Asp giving it a net charge per residue of +0.38 (excluding the intrinsic negative charge of the C-terminal residue).

The positively charged residues are distributed all throughout the CTD, often in pairs, while the three negatively charged residues are found in the N-terminal half [Appendix B (b)]. As is common in strongly charged IDRs [16], the linear sequence distribution of charges in the CTD is relatively uniform, resulting in a low κ-value of 0.15. (see Methods: *Charge distribution*).

To assess its structural properties, the CTD was simulated as a single-chain peptide (see Methods: *Single-chain simulations*). As the CTD is intrinsically disordered, the simulation provided an ensemble of different conformations of the peptide, all of which were considered for structural measures, rather than fixating on one particular structure (see Methods: *Structural measures*). Besides the wild-type,

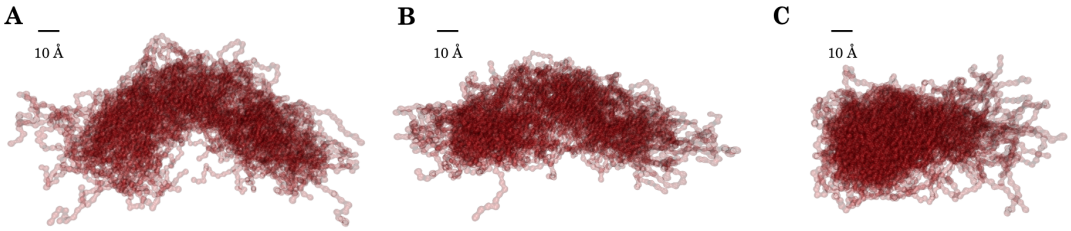


Figure 2 Representative structural ensembles from trajectories of human H1.0 CTD variants. **A** Wild type variant. **B** Randomly shuffled variant. **C** Terminally clustered charges variant.

Variant	κ *Charge patterning*	$\langle R_g \rangle$ *Size*	ν *Expandedness*	Δ *Symmetry*	S *Shape*
Wild type	0.15	4.2 nm	0.67	0.29	1.41
Randomly shuffled	0.22	4.2 nm	0.66	0.29	1.38
Terminally clustered charges	0.99	3.0 nm	0.57*	0.22	1.20

Table 1 Sequence and structural measures for human H1.0 CTD variants (distributions of structural measures can be found in Appendix C). *The variant with terminally clustered charges fitted poorly to a regular polymer scaling law [Appendix C (d)] due to its long-range intramolecular interactions [Appendix D (c)].

a randomly scrambled variant and a variant where positive charges were moved to the N-terminal and negative charges to the C-terminal were also simulated.

All variants demonstrated wide distributions of R_g [Appendix C (a)] and a lack of distinct contacts in intrachain residue contact maps [Appendix D (b)], which supports that all three variants are indeed intrinsically disordered in simulations. Additionally, their intrinsic disorder was also observable when visualising their trajectories directly [Fig. 2].

The wild type variant exhibits a very expanded structure for its sequence length, as reflected in its high Flory scaling exponent ν [Table 1], which for reference is in the 99th quantile of ν-values of IDRs (see Methods: *Reference IDRs*). As is common in highly extended IDRs [17], its structural ensemble was dominantly prolate with a high value of S and asymmetric with a relatively high value of Δ. It should be noted however, that asymmetry intrinsically follows from higher degrees of prolateness.

The structural measures [Table 1] indicate that there is little to no structural difference between the wild type and the randomly shuffled variant, while they have some κ difference. Indeed their distribution of structural measures are near-indistinguishable [Appendix C (d)]. While this nicely exemplifies how a common amino acid composition may achieve the same structural properties [1], the generation of artificial H1.0 CTD variants by simulated evolution in a later section will demonstrate that even shuffled variants do not necessarily yield similar conformations ensembles despite similar composition and κ.

The variant with terminally clustered charges appears to be smaller, more compact, and to be less prolate than the two other variants. Intramolecular contact and energy maps indicate slight, disordered contacts between the N- and C-terminal end of the CTD [Appendix D]. This can be explained by the fact that high-κ sequences, where charges are segregated towards each end, tend form more hairpin-like structures rather than random swollen coils-like [16]. This demonstrates the importance of charge patterning in IDR structure, especially in that of strong polyelectrolytes like the H1.0 CTD.

3.2 Assessing conservation in H1.0 CTD orthologs

To assess the general features of the domain, a set of 189 orthologs of the CTD of H1.0 across all of Eukaryota was assembled using phylogeny and disorder annotations from OrthoDB and UniProt (see Methods: *Curating H1.0 CTD orthologs*) to assess the conservation of both sequence and structural features.

3.2.1 The conservation of sequence features

As found in former studies [1, 2], the general amino acid composition profile was generally well-conserved across orthologs, with the human variant well-representing the set. The orthologs were characterised by a low composition complexity, as Lys, but also Ala, Pro, and to a lesser degree Ser, Val, and Thr

constituted the major part of the ortholog sequences [Fig. 3A]. In most orthologs there was little polar residues like Cys, Asn, His and Gln and next to no hydrophobic aromatic residues like Phe, Tyr, or Trp, which is common for IDRs [16, 28]. Generally, there was a low number of the negatively charged residues Asp and Glu. When compared against a reference set of IDRs (see Methods: *Reference IDRs*), Lys and Ala are especially enriched whereas there appears to be a bias against Glu, Leu, Asp, Asn, and His. It should be noted that there seems to be a relatively larger degree of variation in the content of Ala, Ser, Thr, and Val.

The primary sequence of the H1.0 CTD is only partly conserved, as is seen in a calculated average sequence identity score of 0.73 for the H1.0 CTD compared to an identity score of 0.83 for the H1.0 folded H15 domain.

It may be the case that the low amino acid composition complexity of the orthologs has lead to an artificially higher conservation score; if there are only a few different amino acid types to shuffle around, some sequences are bound to be similar by chance. To assess the sequence identity in this context, a randomly shuffled variant was generated from each of the orthologs. These shuffled variants generally exhibit lower sequence identity scores when aligned than the wild types [Fig. 3B], indicating that the primary sequence of the H1.0 CTD seems to be conserved to a degree beyond just amino acid composition. There is, however, a wide distribution of both low and high position-wise sequence identity scores across the CTD [Appendix F].

At least some of the high sequence identity motifs can be attributed to conserved CDK phosphorylation sites [13] and suspected DNA-binding motifs [9, 14] [Appendix F].

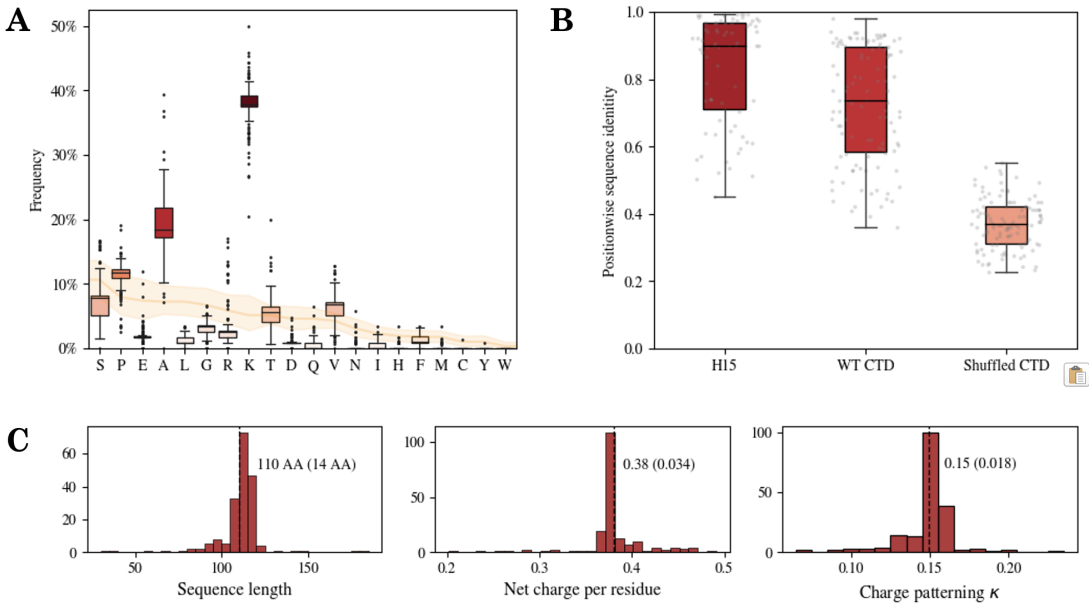


Figure 3 Conservation of sequence features across orthologs. **A** The amino acid content of orthologs (boxplots) with respect to reference frequencies of amino acids in IDRs (yellow) (see Methods: *Reference IDRs*). **B** Distributions of the position-wise sequence identity in mulitple sequence alignments of orthologs. *H15*: The wild type central H15 folded domain of H1.0. *CTD WT*: The wild type CTD of H1.0. *Shuffled CTD*: Randomly shuffled variants of the CTD. Each point represents a position in the gap-filtered alignment (see Methods: *Sequence identity*). See Appendix F for examples of sequence identity profiles for WT and shuffled CTD alignments. **C** The distribution of sequence length, average charge per residue, and κ-values. Means are displayed next to dashed lines followed by standard deviation in parenthesis.

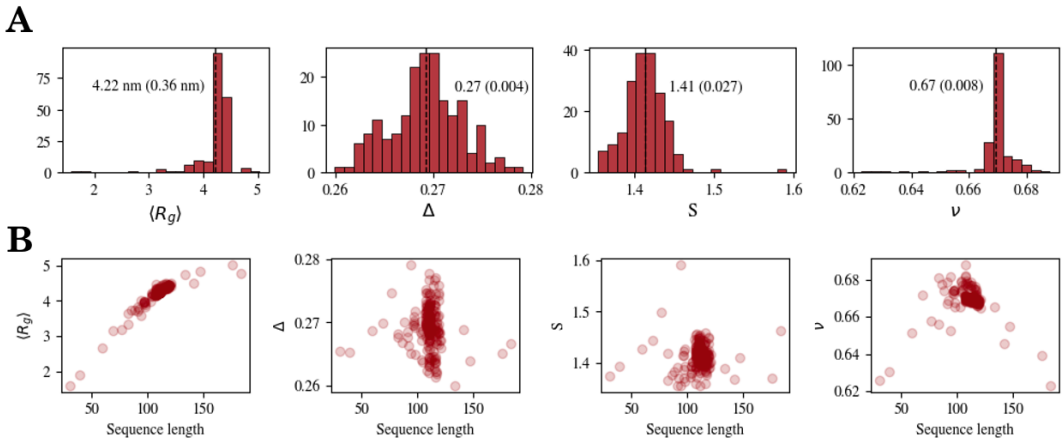


Figure 4 Conservation of structure across orthologs. **A** Distribution of structural measures. Means are displayed next to dashed lines followed by standard deviation in parenthesis. **B** Variation of structural measures with sequence length. Note that $\langle R_g \rangle$ is intrinsically length dependent.

The sequence length seems to be somewhat conserved with 95% of orthologs within 83 and 121 AA in length [Fig. 3C]. Both net charge per residue and charge patterning by κ however, appears to be very well-conserved across all orthologs.

It is worth noting that the overall density and patterning of charge manages to stay constant in spite of the larger variation in sequence length and Lys / Arg content. One possible mechanism of this can be seen in the negative Pearson correlation coefficient of -0.71 between the Lys and Arg content ($p = 1.4 \cdot 10^{-31}$, see Appendix E (b)) that implies a balancing compensation of positively charged residues to conserve the net charge per residue in the CTD (there is only inconsequential amounts of Asp, Glu, and His, so just as well a conservation of FCR).

3.2.2 The conservation of structural features

The ortholog sequences were simulated individually as single-chain peptides (see Methods: *Single-chain simulations*) to assess their structural properties.

The structural properties generally show a high degree of conservation across the orthologs. It should be noted that structural measures such as $\langle R_g \rangle$, Δ, S, and ν are all somewhat correlated [17].

The highly extended chain structure is well-conserved and exhibits little variation, as reflected in the Flory scaling exponent ν. This is in agreement with previous findings for ν conservation amongst IDR orthologs [17]. Most deviation from the general ortholog profile manifests in low scaling exponents for the shortest and longest ortholog sequences [Fig. 4B]. This slight increase in compactness with increasing sequence length is evident in the decreased scaling of $\langle R_g \rangle$ above some sequence length [Fig. 4B]. The lower compactness of shorter sequences seems to be inconsequential for the effective $\langle R_g \rangle$ however.

The overall shape of the peptide chains is well-conserved with small relative standard deviations on ortholog asphericity Δ and prolateness S [Fig. 4A], which seem to vary in a sequence length-independent manner [Fig. 4B]. This, however, is most likely a consequence of the conserved scaling exponent ν.

Overall, the general structure of the CTD of H1.0 seems to be well-conserved across orthologs, though there might be a slight propensity towards compacting longer chains to stay within an upper physical size constraint related to the $\langle R_g \rangle$.

3.3 Generating H1.0 CTD variants by sequence evolution

Knowing the general profile of the CTD of H1.0 orthologs, the next step was to generate H1.0 CTD variants for later use.

An evolution algorithm proposed by Pesce et al. [25] was implemented, which takes an input sequence and iteratively randomly switches residue positions to obtain a certain target observable - in this study a specific value of either $\langle R_g \rangle$ or κ (see Methods: *Sequence evolution*).

3.3.1 Preparing an ortholog consensus sequence as input for evolution

An input sequence for the evolution algorithm was generated to represent the set of H1.0 ortholog CTDs.

The orthologs were filtered to ignore certain outlier sequences, which were deemed too long, too short, or too compacted (with either length < 75, length > 125, or $\nu < 0.65$) to represent the general ortholog profile.

The average length and composition of this representative set, referred to as the consensus length and composition, was used to generate *one* representative consensus sequence.

While this input sequence well-represented both the length and composition of the orthologs, its primary sequence was randomly shuffled. This, however, was deemed inconsequential, as the many shuffling steps of the algorithm, eventually changes the sequence far from whence it started.

3.3.2 $\langle R_g \rangle$-constrained evolution

Initially, two runs of the evolution algorithms were set up: one to maximise $\langle R_g \rangle$ by targeting a value of 10 nm, and one to minimise $\langle R_g \rangle$ by targeting a value of 0 nm.

The maximisation run seemed to be unable to optimise the input sequence much, as can be seen in [Fig. 5A], where the algorithm gets stuck on specific sequences for many generations at a time (the 'jumps' of the thick line). Additionally, it did not seem to be able to reach a substantially higher $\langle R_g \rangle$. Ultimately, the run achieved a 0.19 nm increase in $\langle R_g \rangle$.

The minimisation run, however, was able to achieve a lower $\langle R_g \rangle$, seemingly by clustering charges together [Fig. 5B]. While in the first half of the run, negatively charged clusters are seen to form throughout the sequence, the second half of the run implements a more radical charge segregation by clustering the positive charges in either terminus. Ultimately, the run achieved a 0.78 nm decrease in $\langle R_g \rangle$.

One of the main strategies with which the algorithm achieves changes in $\langle R_g \rangle$ is the altering of charge pattering [25]. This is well explained by the fact that segregation of charges into oppositely charged blocks has been reported as a key driver of chain compaction, and vice versa [16, 17].

One of the reasons that the algorithm did not converge towards higher $\langle R_g \rangle$-value in this case may be two-fold caused by the intrinsic charge patterning of the input sequence: 1) the input sequence was randomly shuffled, but since it had a high content of positively charged residues, these charges may likely already be well-distributed across the sequence by chance; and 2) The input sequence only had three negatively charged residues, which does not give the algorithm a lot of opportunities for optimising homogenous charge patterning. The input sequence thus perhaps lacks opportunities to optimise charge patterning and thus $\langle R_g \rangle$. Alternatively, it is known from former work with the algorithm that wild type IDP sequences are often near maximally expanded relative to their sequence composition [25], and it was seen that a random shuffling of the human wild type CTD lead to a structurally near-indistinguishable variant.

While the minimisation run initially seems successful in comparison, there is still room for improvement. From the initial runs of human H1.0 CTD variants, it was observed that a variant with positive residues strictly in the N-terminus and negative residues strictly in the C-terminus was able to achieve $\langle R_g \rangle = 3.0$ nm [Table 1].

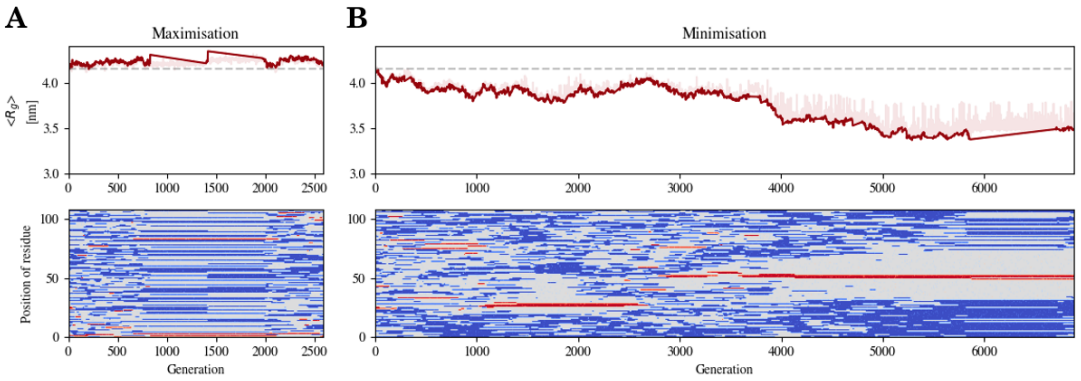


Figure 5 R_g-constrained evolution of a randomly generated input sequence with the consensus composition of H1.0 CTD orthologs (see Methods: *Sequence evolution*). The top panels show the observed $\langle R_g \rangle$ for accepted generations (thick line) as well as for non-accepted generations (shaded). Long jumps of the thick line indicates stretches where the algorithm gets stuck with specific sequences. The lower panels are kymographs showing the movement of charged residues as evolution progresses, where blue indicates positively charged residues (Lys, Arg (No His in sequence)) and red indicates negatively charged residues (Asp, Glu). The dashed line represents the $\langle R_g \rangle$ of the input sequence. **A** Sequence evolution targeting a high $\langle R_g \rangle = 10\ nm$. This evolution was terminated after 2588 generations as no noticeable progress was made in increasing $\langle R_g \rangle$ compared to the input sequence. **B** Sequence evolution targeting a low $\langle R_g \rangle = 0\ nm$

.

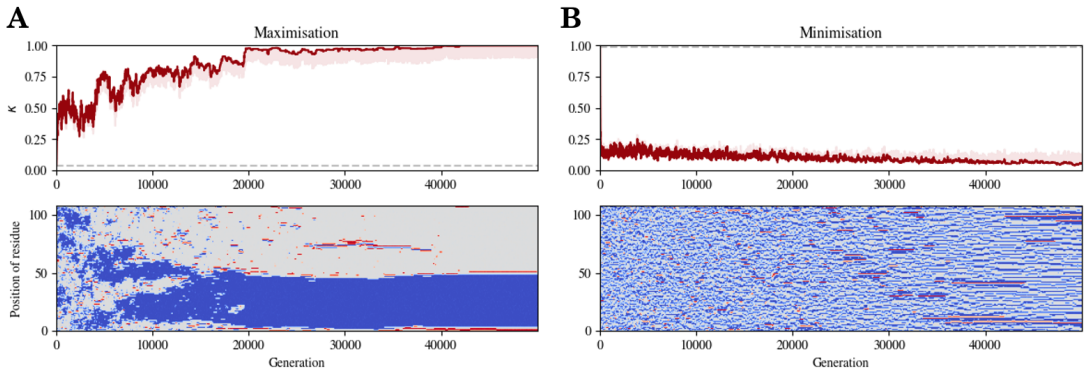


Figure 6 κ-constrained evolution of a randomly generated input sequence with the consensus composition of H1.0 CTD orthologs (see Methods: *Sequence evolution*). The top panels show the observed κ for accepted generations (thick line) as well as for non-accepted generations (shaded). Long jumps of the thick line indicates stretches where the algorithm gets stuck with specific sequences. The lower panels are kymographs showing the movement of charged residues as evolution progresses, where blue indicates positively charged residues (Lys, Arg (No His in sequence)) and red indicates negatively charged residues (Asp, Glu). The dashed line represents the κ of the input sequence. **A** Sequence evolution targeting a high $\kappa = 1$. **B** Sequence evolution targeting a low $\kappa = 0$.

.

Since the composition of the input sequence was near identical to that of the human ortholog, the same $\langle R_g \rangle$ should be achievable for the evolution. That is, there is potentially still about 0.5 nm of $\langle R_g \rangle$-reduction left for the algorithm to achieve. What seems to be the issue is that the negatively charged residues are centered, rather than clustered in a terminus. This perhaps represents a local optimum that may be difficult to navigate out of given that only single residue position swaps are performed in the algorithm. Thus, while the minimisation run 'converged' on an overall lower $\langle R_g \rangle$, it has only achieved about half of the potential $\langle R_g \rangle$-minimisation.

The minimisation run appears to make several attempts at clustering the negatively charged residues together in the first half of the generations. None of these transient negative clusters appears to give any significant payoff in $\langle R_g \rangle$-reductions when they form, in comparison to the final negative cluster in the second half of the generations. This might indicate some kind of sequence threshold that has to be surpassed for the algorithm to start converging towards a significant optimum. One thing that characterises the final sequence charge patterning is the *No man's land* in between the positively and negatively charged residues consisting of neutral residues. Such a charge profile also appears to start to form just before generation 2000, but ultimately fails to stabilise. Another characteristic of the final sequence charge patterning is that the negative charges are centralised. While segregating charges in opposite terminals presumably achieves the global minimum of $\langle R_g \rangle$, there might as mentioned exist a local minimum by clustering one type of charge centrally, when the charge content is as asymmetrical (many +, few -) as in the H1.0 CTD.

To generate a larger sequence space to sample variants from, 3 other replicates of the minimisation run was submitted, which yielded similar results.

3.3.3 κ-constrained evolution

The computation of $\langle R_g \rangle$ for each generated sequence in the evolution required simulations and reweighting, both of which were computationally costly. This challenged the efficacy of the algorithm to effectively search a large sequence space.

Charge patterning, as measured by κ is significantly correlated with IDP $\langle R_g \rangle$ [16]. This correlation is clear in the sequences generated from the $\langle R_g \rangle$-targeting algorithm runs [Appendix G (a)]. As κ is much quicker to calculate than the trajectory-based $\langle R_g \rangle$, the algorithm was repurposed into targeting κ as an observable instead.

Again, both a maximisation and a minimisation run was submitted, targeting $\kappa = 1$ and $\kappa = 0$ respectively. Both of these runs were efficiently able to sample the entire κ-range of 0.0-1.0 with clearly traceable evolution of charge patterning [Fig. 6].

By definition, terminally segregation of opposite charges is the maximum κ-value, and the maximisation of κ can be thought of as the equivalent to minimisation of $\langle R_g \rangle$ [16]. Curiously, in the κ maximisation run, the algorithm did not manage to segregate oppositely charged residues in either terminus either, instead placing all charges in one terminal domain. Running the algorithm for further generations or loosening the Monte Carlo control parameter may lead to convergence, but such a brute-force solution is hardly efficient for addressing the shortcomings that the algorithm exhibits.

3.3.4 Selecting variants by clustering

To probe the effect of $\langle R_g \rangle$ and κ (and other structural measures) on the H1.0 CTD-ProTα interaction, a set of variants that were representative of both the sampled $\langle R_g \rangle$ and κ range was required.

All of the variants generated in $\langle R_g \rangle$-evolution runs were collected into one set. This set of variants were clustered into 20 clusters by their $\langle R_g \rangle$-value, from which 20 representative sequences were selected [Appendix G (c)] (see Methods: *Selecting representative sequences from clusters*).

As κ-constrained evolution was quick, 11 runs were submitted targeting $\kappa = [0.0, 0.1 \ ... \ 0.9, 1.0]$. All of the evolutions were afterwards collected into one set, 20 clusters were fitted, and 20 representative

sequences chosen [Appendix G (d)].

In the end, it was decided to include extra variants that were known to sample both $\langle R_g \rangle$- and κ-space simultaneously. This involved fitting the collective $\langle R_g \rangle$-evolution set to 50 clusters based both on variant $\langle R_g \rangle$ and κ, from which 50 representative sequences were chosen [Appendix G (b)].

Thus, from runs of the evolution algorithm, 90 artificial variants of the H1.0 CTD were generated.

3.4 Simulating the H1.0 CTD–ProTα interaction

The binding affinity of the complex of the H1.0 CTD and ProTα was assessed using two-chain simulations of both the wild type peptides as well as with representative H1.0 CTD variants from the simulated sequence evolution.

3.4.1 Evaluting wild-type H1.0 CTD-ProTα binding

Borgia et al. [15] utilised single-molecule FRET to measure a set of dissociation constants K_d for the full length H1.0-ProTα complex to establish the binding affinity over a range of ionic strengths. These *in vitro* results, though based on the full-length H1.0 protein, will provide the benchmark for assessing whether or not the applied simulation setup can capture the H1.0 CTD-ProTα interaction.

The wild type H1.0 CTD and ProTα were simulated together in a 25 nm box, corresponding to a concentration of 66 mM of each species (see Methods: *Two-chain simulations*). While initially placed 10 nm apart, the two peptide chains quickly associate and bind strongly. So strongly in fact that it was very rare, even for long trajectories, to sample unbinding events. Even though the interaction is highly dynamic it is also extremely high-affinity.

This presented a challenge for calculating K_d by traditional means, which often involves counting the number of frames in the trajectory where the two chains are bound and unbound respectively. With few to no unbinding events, the K_d would either be impossible to calculate or be far from converged. Instead, with inspiration from similar MD simulations in Borgia et al. [15], the K_d was calculated using a free-energy approach that evaluates the interaction energy between the two peptides as a function of their distance (see Methods: *Dimer dissociation constant*). This only required an energy landscape with the peptide chain distance as the collective variable.

Such a landscape could be robustly mapped out with a set of frames from even a single unbinding event [Fig. 7B]. However, despite the efficacy of the method, not all simulations sampled a single binding event. The vast majority of the interactions between the H1.0 CTD and ProTα are charged-based, and thus salt-dependent. Therefore, observing an unbinding event was severely affected by whether or not the ionic strength was high enough to weaken the electrostatic interactions that afford the complex its strong affinity.

In the cases where binding events were sampled, however, calculated K_ds were underestimated [Fig. 7C], and affinity thus overestimated. Good agreement with *in vitro* data could, however, in all cases be obtained by scaling down the effective potential mean force $pmf_{eff}(r)$ by a unitless factor of 2.48 (see Methods: *Dimer dissociation constant*). Because of the empirical efficacy of this scaling to achieve agreement with experimental data, all K_d calculations were adjusted as such.

The *in vitro* and *in silico* K_ds can be found in [Table 2 and Fig. 7C]. In cases without an unbinding event, the K_d was, as expected, far off from experimental data. Thus, the simulation setup captures the affinity of the complex well, as long as unbinding is observed, and the $pmf_{eff}(r)$ is scaled.

The simulations also capture structural changes in both peptides upon complex formation [Appendix H]. Specifically, both peptides seem to undergo a compaction in size and become slightly less prolate

Ionic strength	**In vitro** K_d	**In silico** K_d	**In silico** Adjusted K_d	**In silico** Unbinding sampled
165 mM	2 pM	141 fM	787 nM	No
180 mM	37 pM	7 fM	242 nM	No
205 mM	1 nM	17 fM	291 nM	No
240 mM	25 nM	4 aM	8 nM	Yes
290 mM	230 nM	14 pM	213 nM	Yes
330 mM	1 µM	2 fM	1 µM	Yes
340 mM	4 µM	6 fM	2 µM	Yes

Table 2 Experimental K_ds from Borgia et al. [15] for wild type H1.0-ProTα binding against adjusted K_ds calculated from coarse-grained simulations of H1.0 CTD and ProTα peptides. Note that unbinding of the two peptide chains did not always occur in the sampling time (10 µM), which incurs an incomplete sampling of the interaction free energy landscape used to calculate the K_d (see Methods: *Dimer dissociation constant*). Errors for *in vitro* K_ds are illustrated in Fig. 7.

and asymmetric, compacting more towards, but not quite into, a common ampholytic globule. In accordance with Borgia et al.[15], most H1.0 CTD contacts with ProTα occur in its central charge-rich region [Appendix I].

3.4.2 Investigating variant H1.0 CTD-ProTα binding

With the credibility of the simulations to capture binding affinity assured and with artificial variants prepared, the influence of sequence variation on the H1.0 CTD-ProTα interaction could be assessed.

Alongside the wild type a total of 90 artificial variants were tested, which represented sequences sampled from the $\langle R_g \rangle$-κ-space [Fig. 8A]. Due to the observed sampling issues at low ionic strengths, all variants were run at 290 mM ionic strength, at which, almost all simulations sampled at least one unbinding event.

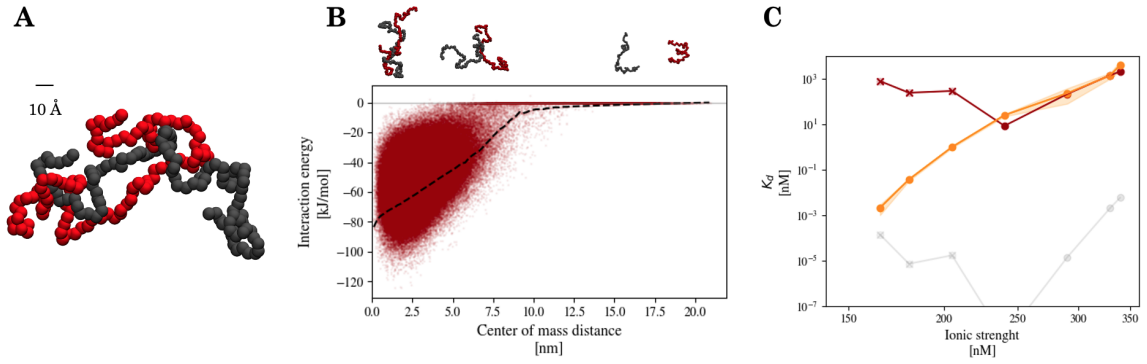


Figure 7 The binding of wild-type H1.0 CTD and ProTα. **A** A representative orthographic view of the binding of the coarse grained H1.0 CTD (red) and ProTα (black) peptides. **B** The energy landscape used to calculate K_d (Ionic strenght of 290 mM). Each red dot represents a frame in the trajectory. The dashed line represents the effective potential mean force $pmf_{eff}(r)$ (see Methods: *Dimer dissociation constant*). **C** The salt-dependence of K_ds, both *in vitro* (yellow), non-adjusted *in silico* (grey), and adjusted *in silico* (red). Error bars of *in vitro* K_ds is denoted as shaded area, and *in silico* data with no sampled unbinding event has a × marker.

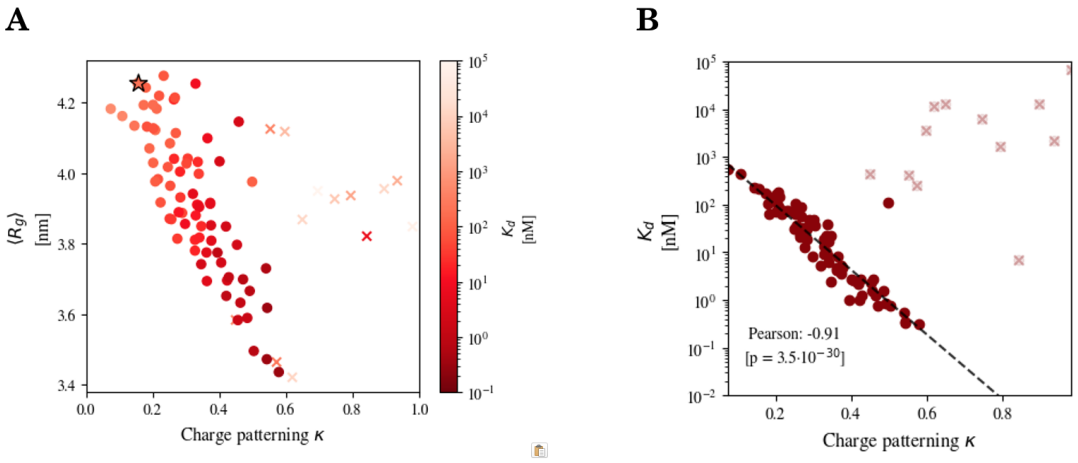


Figure 8 The binding of artificial H1.0 CTD variants and ProTα. **A** The $\langle R_g \rangle$- and κ-value of artificial variants coloured by their K_d. The human wild type is included and has a $\star$ marker. Data with no sampled unbinding event has a $\times$ marker, and was not included in correlation. **B** The correlation between variant κ and K_d. Correlation of K_d with other measures can be found in Appendix J

Resulting binding simulations demonstrated a correlation between K_d and the measures κ, $\langle R_g \rangle$, Δ, S, and ν to a varying degree [Appendix J]. The fact that all of these measures seem to correlate with K_d to some degree is not surprising, as there generally is a high multicollinearity between them [Appendix J].

The measure which most significantly correlated with K_d, however, was κ [Fig. 8B]. High κ-valued sequences achieved lower K_ds and thus had a higher affinity for ProTα. Noticeably, the patterning of charge in sequences with the same composition could affect K_d over three orders of magnitude [Fig. 8B]. Due to the high multicollinearity between the measures however, it is not straightforward to distinguish the causative effect on K_d that each measure has independently of each other.

4 Discussion

This study set out to assess the importance of charge patterning and structural properties for the binding of the H1.0 CTD to its chaperone ProTα. Throughout the work to prepare this assessment, insights on the H1.0 CTD surfaced.

4.1 A highly expanded H1.0 CTD is conserved, though longer chains compact slightly

Initial investigation of the human H1.0 CTD found it to be a highly expanded chain in regards to its scaling exponent ν, even for IDP standards. This is in accordance with previous work, which found that highly expanded IDRs are significantly enriched in histones [17]. Additionally, this high scaling exponent was conserved across orthologs.

Several reasons may explain why this highly-expanded-chain property of the H1.0 is so conserved. For one, it may be a coincidental consequence of its high amount of charged residues that may otherwise be required for DNA-binding. As demonstrated, a random shuffling of the sequence lead to a relatively low κ, which is correlated with high ν [16, 17].

On the other hand, the expanded and well-solvated random coil may be beneficial for exposing itself to its binding partners, whether DNA or other proteins. It could be speculated that since the CTD is known to recognise many binding partners specifically while at times being partly buried in chromatosome structures, chain compaction may risk burying certain motifs within chromatin.

Another interesting feature of the H1.0 CTD polymer scaling was the observation of a slight plateau in $\langle R_g \rangle$ at a certain sequence length amongst orthologs, accompanied by a drop in scaling exponent values. This sort of upper constraint of IDR dimensions is a phenomena that has recently become of interest in contemporary studies [17, 29, 30], and which may represent a general conservation phenomena in IDR conservation.

4.2 The relatively high sequence identity of the H1.0 CTD could be caused by H1-specific and non-specific sequence constraints

While contemporary literature highlights the lack of conservation in IDR sequences [1, 2], this study found a somewhat high degree of conserved sequence identity across H1.0 CTD orthologs.

One factor that is important to note in this regard, is that the definition of what constitutes the intrinsically disordered CTD may differ. In this study, MobiDB [21] disorder predictions were used to achieve a consistent definition of the CTD. One could just as easily utilise other predictors, or, alternatively define the intrinsically disordered CTD as starting where annotations for the folded H15 domain ends.

Nonetheless, the relatively high sequence identity may be a biological consequence of the fact that the H1.0 CTD has many important functional roles that require a conserved primary sequence. For one, as already mentioned, the CTD is known have several binding partners, to specifically conserve consensus sequence sites for regulatory post-translational modifications, and to have conserved motifs for DNA-binding. Other non-histone specific factors may also play a role in the specific distribution of residues, such as the fact that sequence patches with a high density of positively charged residues ($\geq$6 in a 10 AA segment) may cause stalling in ribosomes [31].

4.3 The conserved sequence composition of the H1.0 CTD may be explained by DNA-binding and intrinsic disorder

The sequence composition was well conserved among orthologs, as it is also known to be across paralogs [2–4]. While conserved composition among orthologs may in part be attributed to sequence identity, the fact that composition, and functional roles, is reported to be conserved across paralogs indicates that more than the primary sequence may be at play [2, 4].

Generally, the sequence composition is of low complexity, with only a few types of residues making up the vast majority of it. The vast majority of the composition is Lys, Ala, and Pro, which may be explained both by the promotion of disorder and by DNA-binding.

For one, all three residues are known to be enriched in intrinsically disordered proteins [28]. As pointed out, intrinsic disorder is thought to be an important part of the the CTD functions [2–4], making these disorder-promoting residues beneficial.

Besides that, Pro and especially Lys is reported to be important for DNA-binding [6, 14], which is an important function of the H1.0 CTD [3, 4]. Additionally, phosphate groups from post-translational modifications are reported to drastically affect H1 binding to DNA [13], which explains the apparent need for a conserved deficiency of the negatively charged residues Asp and Glu.

4.4 While $\langle R_g \rangle$-constrained evolution simultaneously samples κ-space, the opposite is not true

The main setbacks faced with the evolution algorithm (in regards to the objectives of [**this**] project) was difficulty in achieving wide sampling of $\langle R_g \rangle$ and κ-space efficiently. This was the case for both the $\langle R_g \rangle$- and the κ-constrained evolution.

In the case of the $\langle R_g \rangle$-evolution, the main issue was long sampling times. The evolution did, however, sample $\langle R_g \rangle$-values in the range 3.4-4.3 nm, as well as κ-values in the range of 0.2 and 0.7. While this is not the entire κ-range of 0.0-1.0, it still covers a substantial part of it.

The κ-constrained evolution, on the other hand, sampled the entire κ-range, but the $\langle R_g \rangle$-values of the resulting representative variants only varied little with $\langle R_g \rangle$ in the range of 3.9-4.1 nm. While $\langle R_g \rangle$ and κ-values have been reported to be well correlated [16], this provides an example of a case, where κ can vary significantly with little $\langle R_g \rangle$-variation in response, and vice versa.

Thus, while the $\langle R_g \rangle$-evolution was significantly slower, it managed to sample a larger space of both structural and charge patterning variants. Neither algorithm was, however, able to efficiently sample the $\langle R_g \rangle$ and κ space simultaneously.

4.5 Experimental affinities of H1.0-ProTα binding are somewhat captured

The affinity of the wild type H1.0-ProTα interaction was highly overestimated *in silico*, in the form of underestimated K_ds.

The coarse-grained simulation model system for the H1.0-ProTα interaction presented involves several approximations in both regards to molecular representations and physical laws. Beyond that, the interaction between full length H1.0 and ProTα has been reduced to simply the CTD of H1.0 and ProTα. These approximations may play at least some part in explaining any discrepancy.

The inability of this simulation framework to capture the interaction is somewhat surprising, seeing as a very similar MD simulation was conducted by Borgia et al. with no need for correction of the pmf_{eff} [15]. There are however some differences between the two simulation setups that may explain some of the discrepancy. For instance, Borgia et al. employed umbrella sampling at low (165 mM) ionic strength, used full length H1.0, and tuned the ϵ-parameter of their Lennard-Jones short range potential to match experimental FRET data. The tuned ϵ was orders of magnitude smaller ($0.001kJ/mol < 0.837kJ/mol$) than the one employed in this study's model. While this may explain why affinity is overestimated to

some degree, short-range Ashbaugh-Hatch interactions did not make up a large part of the H1.0 CTD-ProTα interaction [Appendix I]. It should be noted that the energy term for salt-screened interactions reported in Borgia et al. [15] does not include division by the residue distance, but this is suspected to be an accidental omission.

Another explanation might be unfulfilled sampling of the potential mean force across peptide distances, which would mean that the K_ds are not fully converged. Given the long simulation times (10 µM) though, this seems unlikely.

4.6 H1.0 CTD-ProTα affinity decreases with homogeneous charge patterning

While multicolinearity of κ and the structural measures complicates matters, there was an unequivocal correlation between κ and K_d.

Higher κ-values lead to orders of magnitude lower K_d, and thus higher binding affinity, and vice versa. This is significant, as it establishes that distributing interactions *throughout* the CTD actually decreases affinity. One of the likely causes of this is that ProTα has a central patch that is dense in negatively charged residues. Thus, H1.0 CTD sequences that form positively charged patches can achieve favourable electrostatic binding with many more of its residues simultaneously, compared to if they had been evenly distributed across the sequence.

Intriguingly, it can be seen in [Fig. 8] that the human wild type variant, as marked by a $\star$, is in the lowest κ and highest $\langle R_g \rangle$ corner, even despite wide $\langle R_g \rangle$ and κ sampling with the evolution algorithms. This sequence area appears to be characterised by high K_ds and thus lower affinity. As has been discussed, several factors may explain the low κ and expanded size of the H1.0 CTD, whereof ProTα binding may not necessarily be the deciding one. The placement of the wild type H1.0 CTD sequence in the extreme corner of κ-$\langle R_g \rangle$ space, where affinity is at its lowest, may imply that there probably is not a lot of 'evolutionary wiggleroom' for H1.0 to work with. Seeing as H1.0 presumably has many more vital functional roles to conserve, it may conversely rather be the case that ProTα modulates its charge patterning.

This idea is supported by the fact that ProTαhas some extent of clustering of its negative charges, as reflected in its κ-value of 0.42. This is in spite of the previously mentioned fact that polyelectrolytic IDRs tend to have low κ values [16], presumable to conserve IDR dimensions. Interestingly, ProTα , which is of the same size (111 AA) and which has a slightly higher fraction of charged residues than the H1.0 CTD (58% > 42%), achieves almost the same dimensions as the H1.0 CTD [Appendix H], despite a much higher κ. Thus, presumably, it achieves its dimensions by other means than κ. Instead, given the observed dependence of charge patterning for complex affinity, its relatively high κ-value may reflect a tuning of charge patterning to achieve an optimum K_d for it to act as the chaperone of H1 histones in competition with DNA.

4.7 Further work could assess ProTα charge patternning or investigate other binding partners

While this project achieved to sample many K_ds of artificial H1.0 CTD variants binding to ProTα , not all interactions could be properly sampled for unbinding due to the high affinity of the complex. This was mostly an issue in regards to measuring K_ds at low ionic strengths and at high κ, at which binding was the strongest. One possible way to sample unbinding when unbinding is severely disfavoured, could be through enhanced sampling techniques, like umbrella sampling as exemplified in Borgia et al. [15].

Optimisations of $\langle R_g \rangle$-values in the evolution algorithm runs also had trouble evolving to higher/lower values. While the $\langle R_g \rangle$-maximisation may have had hit an biologically intrinsic upper limit, it was demonstrated that $\langle R_g \rangle$-minimisation only achieved half of the possible $\langle R_g \rangle$-reduction. This presented an issue, as the output of the evolution is used to sample wide ranges of $\langle R_g \rangle$-variants.

The role of charge patterning in determining $\langle R_g \rangle$ is evident, and one issue that the algorithm may be facing is its limitation to only switch two single residues at a time. Allowing the algorithm to switch blocks of residues instead, may speed up the ability of the algorithm to search sequence space. Another possibility may be to assist the search for $\langle R_g \rangle$-values by using either κ or a machine-learning model that predicts $\langle R_g \rangle$ to choose the best candidate among a set of possible sequences before proceeding to simulation. Alternatively, other measures than κ for charge patterning could be considered, like sequence charge decoration (SCD), which, as opposed to κ captures long-range charge patterning, and which is additionally reported to capture other charge patterning effects on structure than κ [17]. Finally, one could consider to add new features to the algorithm, such as constraints to preserve conserved sequence motifs and thus more accurately investigate the possibilities of sequence optimisation in an evolutionary context.

It could be interesting to look further into the H1.0 CTD-ProTα -system. For one, the N-terminal and folded domain of H1.0 could be included in the simulations, but with variant CTDs. This is, however, unlikely to have a large impact, as demonstrated by the ability of the wild type CTD to achieve the affinity of the full length construct [15]. Alternatively, the effect of post-translational modifications, which are important for both for IDP conformational properties [16] and the H1 CTD specifically [3, 13], could be assessed by including representations of these in the coarse-grained structure.

In regards to the speculation of κ tuning in ProTα, it may be insightful to repeat the type of analysis in this study on ProTα instead. Alternatively, one could investigate the importance of composition, which dictates the sequence space that can be sampled, by applying the same procedures to other H1 paralogs.

Lastly, it may also be interesting to investigate other functional partners of H1.0 than ProTα. While DNA may initially seem as an obvious choice, it may not be straightforward to incorporate into the current coarse-grained simulation framework. Another exciting binding partner to investigate would be the linker histone-binding nuclear autoantigenic sperm protein (NASP), which has intrinsically disordered histone binding domains [32].

5 Conclusions

The H1.0 CTD was characterised as being an intrinsically disordered, highly positively charged and highly expanded chain. Its low complexity composition, dominated by Lys, Ala, and Pro, and its homogeneous charge patterning was found to be well-conserved along with its expanded structural profile across orthologs.

Artificial variants of the H1.0 CTD was generated from an ortholog consensus sequence by an evolution algorithm that maintained sequence composition but targeted structural or sequence measures. The algorithm was however challenged by computational limitations and convergence difficulties.

The wild-type H1.0 CTD-ProTαinteraction, while not initially captured by a simulated model, were eventually adjusted by an energy scaling term to be in agreement with experimental data as long as unbinding of the complex was observed. Testing this interaction with artificial variants revealed that homogeneous charge patterning assures a lower affinity of the highly electrostatic complex, presumably due to a central negative patch in ProTα.

Interesting angles for further work includes modifications to the applied evolution algorithm to enhance sampling and convergence, to investigate the effect of post-translational modifications, to perform a similar analysis on ProTα , or to investigate other binding partners of the H1.0 CTD like NASP.

Data and code availability

The data, source code, and notebooks used for performing the work in this project is found in a publicly accessible well-documented GitHub repository (github.com/frederikespersen/OrthoIDP).

All figures and Appendix entries can be found in original resolution in the repository under `~/figures/`.

References

1. Moesa, H. A., Wakabayashi, S., Nakai, K. & Patil, A. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol. BioSyst.* **8,** 3262–3273. http://dx.doi.org/10.1039/C2MB25202C (12 2012).
2. Hansen, J. C., Lu, X., Ross, E. D. & Woody, R. W. Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains. *Journal of Biological Chemistry* **281,** 1853–1856. ISSN: 0021-9258. https://www.sciencedirect.com/science/article/pii/S0021925820707325 (2006).
3. Parseghian, M. H. What is the role of histone H1 heterogeneity? A functional model emerges from a 50 year mystery. *AIMS Biophysics* **2,** 724–772. ISSN: 2377-9098. https://www.aimspress.com/article/doi/10.3934/biophy.2015.4.724 (2015).
4. Caterino, T. L. & Hayes, J. J. Structure of the H1 C-terminal domain and function in chromatin condensation. *Biochemistry and Cell Biology* **89.** PMID: 21326361, 35–44. eprint: https://doi.org/10.1139/O10-024. https://doi.org/10.1139/O10-024 (2011).
5. Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I. & Bai, Y. Emerging roles of linker histones in regulating chromatin structure and function. *Nature Reviews Molecular Cell Biology* **19,** 192–206. ISSN: 1471-0080. https://doi.org/10.1038/nrm.2017.94 (Mar. 2018).
6. Ponte, I., Romero, D., Yero, D., Suau, P. & Roque, A. Complex Evolutionary History of the Mammalian Histone H1.1-H1.5 Gene Family. en. *Mol Biol Evol* **34,** 545–558 (Mar. 2017).
7. Lu, X. & Hansen, J. C. Identification of Specific Functional Subdomains within the Linker Histone H10 C-terminal Domain. *Journal of Biological Chemistry* **279,** 8701–8707. ISSN: 0021-9258. https://www.sciencedirect.com/science/article/pii/S0021925817477684 (2004).
8. Lu, X., Hamkalo, B., Parseghian, M. H. & Hansen, J. C. Chromatin Condensing Functions of the Linker Histone C-Terminal Domain Are Mediated by Specific Amino Acid Composition and Intrinsic Protein Disorder. *Biochemistry* **48,** 164–172. ISSN: 0006-2960. https://doi.org/10.1021/bi801636y (Jan. 2009).
9. Vila, R., Ponte, I., Jiménez, M. A., Rico, M. & Suau, P. A helix-turn motif in the C-terminal domain of histone H1. *Protein Science* **9.** Cited by: 41; All Open Access, Bronze Open Access, Green Open Access, 627–636. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034090246&doi=10.1110%5C%2fps.9.4.627&partnerID=40&md5=b55567a057aaf08d043339d733a9bb7f (2000).
10. Vila, R., Ponte, I., Collado, M., Arrondo, J. R. & Suau, P. Induction of Secondary Structure in a COOH-terminal Peptide of Histone H1 by Interaction with the DNA: AN INFRARED SPECTROSCOPY STUDY. *Journal of Biological Chemistry* **276,** 30898–30903. ISSN: 0021-9258. https://www.sciencedirect.com/science/article/pii/S0021925820802361 (2001).
11. Roque, A., Iloro, I., Ponte, I., Arrondo, J. L. R. & Suau, P. DNA-induced Secondary Structure of the Carboxyl-terminal Domain of Histone H1. *Journal of Biological Chemistry* **280,** 32141–32147. ISSN: 0021-9258. https://www.sciencedirect.com/science/article/pii/S0021925820791866 (2005).

12. Caterino, T. L., Fang, H. & Hayes, J. J. Nucleosome Linker DNA Contacts and Induces Specific Folding of the Intrinsically Disordered H1 Carboxyl-Terminal Domain. *Molecular and Cellular Biology* **31.** PMID: 21464206, 2341–2348. eprint: https://doi.org/10.1128/MCB.05145-11. https://doi.org/10.1128/MCB.05145-11 (2011).

13. Roque, A., Ponte, I. & Suau, P. Post-translational modifications of the intrinsically disordered terminal domains of histone H1: effects on secondary structure and chromatin dynamics. *Chromosoma* **126,** 83–91. ISSN: 1432-0886. https://doi.org/10.1007/s00412-016-0591-8 (Feb. 2017).

14. Churchill, M. E. & Travers, A. A. Protein motifs that recognize structural features of DNA. *Trends in Biochemical Sciences* **16,** 92–97. ISSN: 0968-0004. https://www.sciencedirect.com/science/article/pii/0968000491900403 (1991).

15. Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555,** 61–66. ISSN: 1476-4687. https://doi.org/10.1038/nature25762 (Mar. 2018).

16. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **110,** 13392–13397. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1304749110. https://www.pnas.org/doi/abs/10.1073/pnas.1304749110 (2013).

17. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2023/05/08/2023.05.08.539815.full.pdf. https://www.biorxiv.org/content/early/2023/05/08/2023.05.08.539815 (2023).

18. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proceedings of the National Academy of Sciences* **118.** https://www.pnas.org/doi/abs/10.1073/pnas.2111696118 (2021).

19. Tesei, G. & Lindorff-Larsen, K. Improved Predictions of Phase Behaviour of Intrinsically Disordered Proteins by Tuning the Interaction Range. *bioRxiv.* https://www.biorxiv.org/content/early/2022/07/13/2022.07.09.499434 (2022).

20. Zdobnov, E. M. *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **49,** D389–D393. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D389/35363966/gkaa1009.pdf. https://doi.org/10.1093/nar/gkaa1009 (Nov. 2020).

21. Piovesan, D. *et al.* MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Research* **49,** D361–D367. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D361/35363750/gkaa1058.pdf. https://doi.org/10.1093/nar/gkaa1058 (Nov. 2020).

22. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal* **112,** 16–21. ISSN: 0006-3495. https://doi.org/10.1016/j.bpj.2016.11.3200 (Jan. 2017).

23. Bellay, J. *et al.* Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biology* **12,** R14. ISSN: 1474-760X. https://doi.org/10.1186/gb-2011-12-2-r14 (Feb. 2011).

24. Aronovitz, J.A. & Nelson, D.R. Universal features of polymer shapes. *J. Phys. France* **47,** 1445–1456. https://doi.org/10.1051/jphys:019860047090144500 (1986).

25. Pesce, F. *et al.* Design of intrinsically disordered protein variants with diverse structural properties and fixed amino acid composition (*Unpublished manuscript*).

26. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. arXiv: 1201.0490 [cs.LG] (2018).

27. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17,** 261–272. ISSN: 1548-7105. https://doi.org/10.1038/s41592-019-0686-2 (Mar. 2020).

28. Uversky, V. N. The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins* **1.** PMID: 28516010, e24684. eprint: https://doi.org/10.4161/idp.24684. https://doi.org/10.4161/idp.24684 (2013).

29. González-Foutel, N. S. *et al.* Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nature Structural & Molecular Biology* **29,** 781–790. ISSN: 1545-9985. https://doi.org/10.1038/s41594-022-00811-w (Aug. 2022).

30. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct Prediction of Intrinsically Disordered Protein Conformational Properties From Sequence. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2023/05/29/2023.05.08.539824.full.pdf. https://www.biorxiv.org/content/early/2023/05/29/2023.05.08.539824 (2023).

31. Requião, R. D., de Souza, H. J. A., Rossetto, S., Domitrovic, T. & Palhano, F. L. Increased ribosome density associated to positively charged residues is evident in ribosome profiling experiments performed in the absence of translation inhibitors. *RNA Biology* **13.** PMID: 27064519, 561–568. eprint: https://doi.org/10.1080/15476286.2016.1172755. https://doi.org/10.1080/15476286.2016.1172755 (2016).

32. Richardson, R. T. *et al.* Characterization of the Histone H1-binding Protein, NASP, as a Cell Cycle-regulated Somatic Protein. *Journal of Biological Chemistry* **275,** 30378–30386. ISSN: 0021-9258. https://www.sciencedirect.com/science/article/pii/S0021925818443773 (2000).

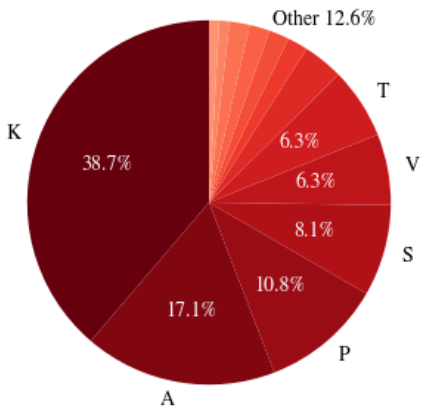
Appendix

A Simulation conditions

System	Topology	Ionic strength [mM]	Boxsize [nm]	Time [µs]
Human H1.0 CTD variants [Table 1, p. 11]	Arbitrary	150	100	1
H1.0 CTD orthologs [Fig. 4, p. 13]	Arbitrary	150	100	1
R_g evolution [Fig. 5, p. 15]	Arbitrary	150	200	0.5
Human H1.0 CTD [To generate topology for two-chain simulations]	Arbitrary	150	100	1
Human ProTα [To generate topology for two-chain simulations]	Arbitrary	150	100	1
Human H1.0 CTD + Human ProTα [Table 2, p. 18]	Compact topologies merged with 10 nm spacing	165 180 210 240 290 330 340	25	10
Evolution variants [To generate topology for two-chain simulations]	Arbitrary	150	100	0.1
Evolution variants + Human ProTα [Fig. 8, p. 19]	Compact topologies merged with 10 nm spacing	290	25	10

Table 3 Simulation-specific conditions. All simulations were run at pH 7 and 298 K.

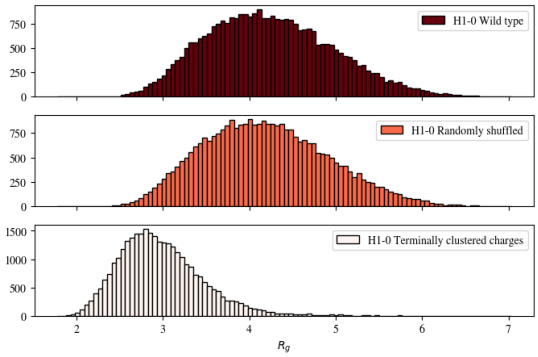
B Amino acid composition and charge distribution in the human H1.0 CTD



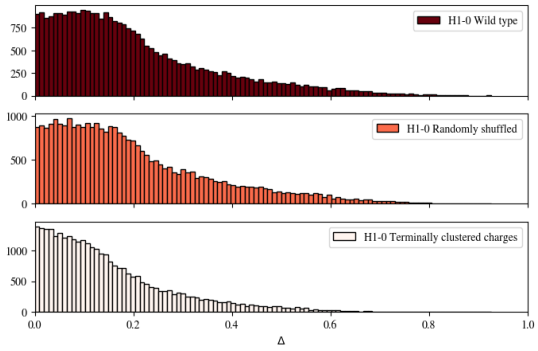
(a) The amino acid content in the wild type CTD of Human H1.0

(b) The patterning of charges in the wild type CTD of Human H1.0

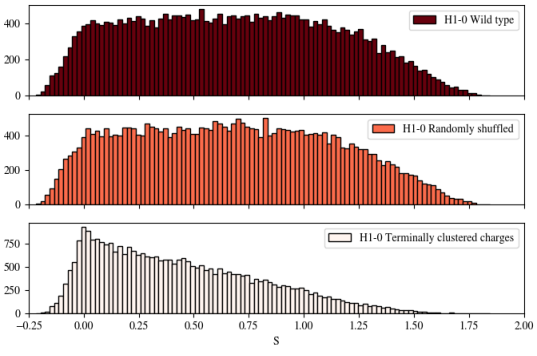
C Structural measures for human H1.0 CTD variants



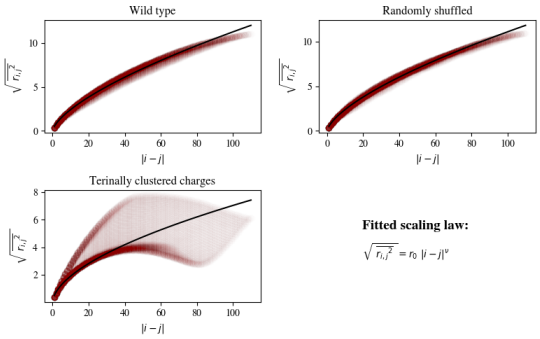
(a) Radius of gyration R_g distribution



(b) Asphericity Δ distribution

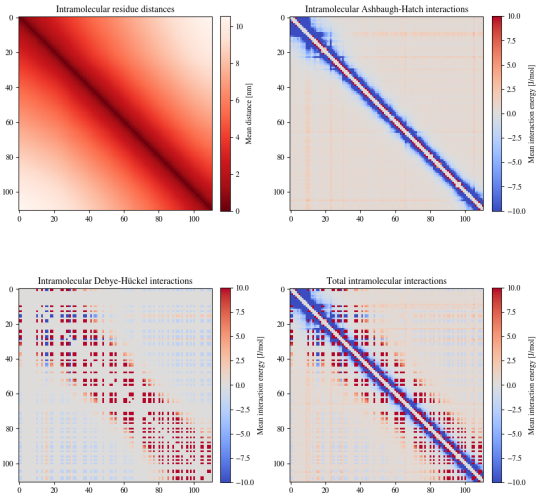


(c) Prolateness S distribution

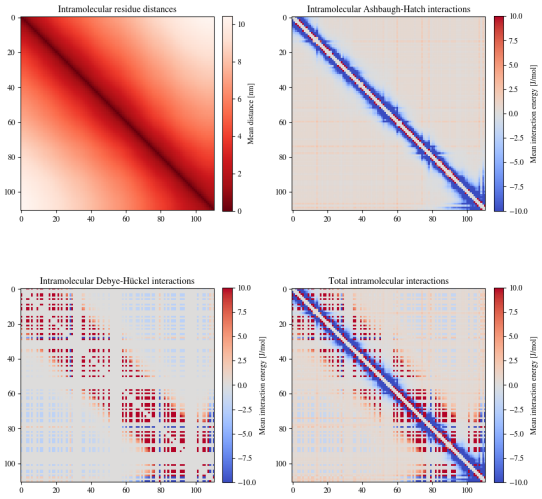


(d) Flory scaling exponent ν fit

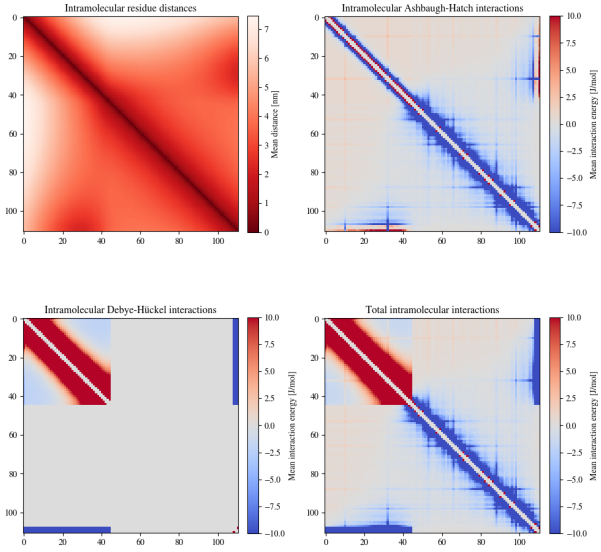
D Contact maps for human H1.0 CTD variants



(a) Wild type

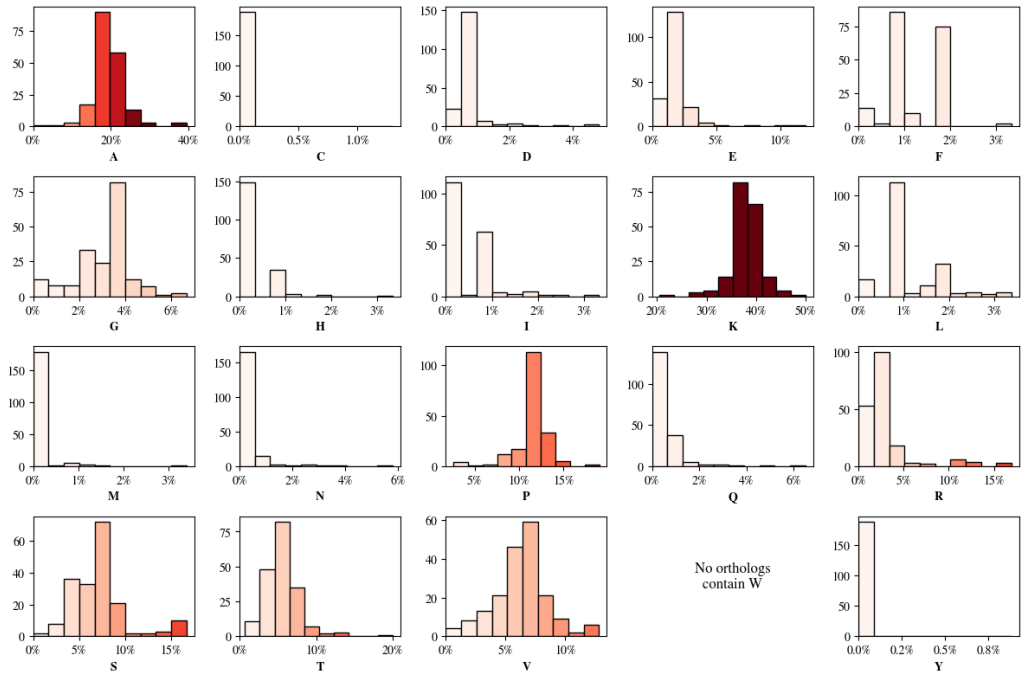


(b) Randomly shuffled

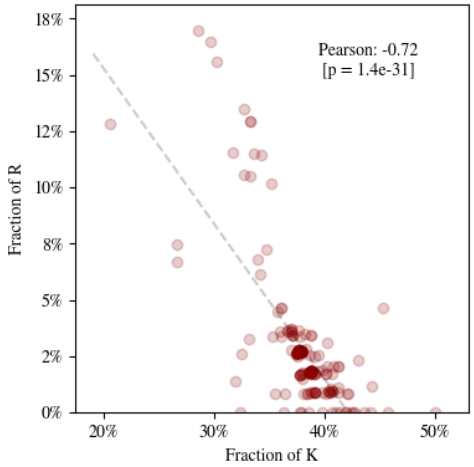


(c) Terminally clustered charges

E Amino acid composition in the CTD of H1.0 orthologs



(a) The distribution of amino acid content in orthologs.



(b) The correlation between the fraction of K and R content in H1.0 CTD orthologs (Pearson correlation, $p = 1.4 \cdot 10^{-31}$).

F Positionwise identity scores for the CTD of H1.0 ortologs

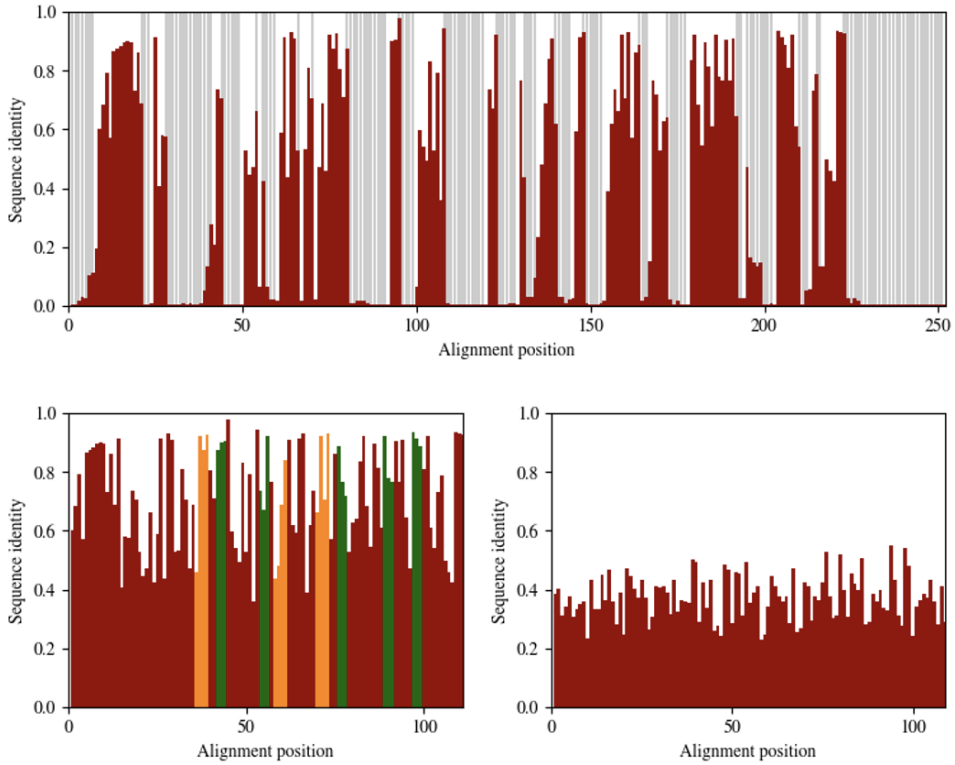
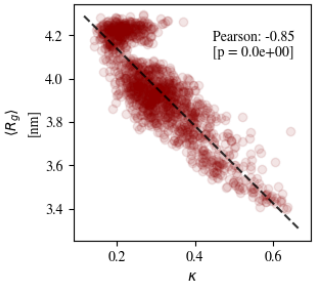
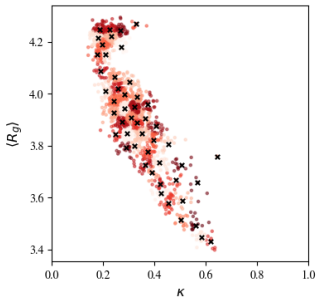


Figure 13 The poitionwise sequence identity in alignments of H1.0 ortholog CTD sequences. **Upper** The positionwise sequence identity in an alignment of wild-type sequences. Grey lines represent positions where a gap '-' was the consensus. **Lower left** The positionwise sequence identity in an alignment of wild-type sequences, where gap-consensus positions were filtered away. (Orange) indicates positions with a conserved cyclin-dependent kinase (CDK) phosphorylation site consensus sequence (S/T-P-X-K/R) [13] and (green) sites with a suspected DNA-binding motif (T-P-K-K or K-P-K) (The first of which also includes CDK-sites and are therefore orange)[9, 14]. **Lower right** The positionwise sequence identity in an alignment of randomly shuffled sequences, where gap-consensus positions were filtered away.

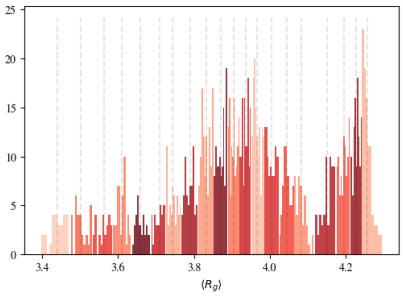
G Generated variants from simulated evolution



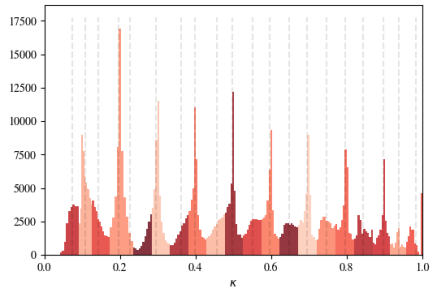
(a) The correlation between κ and $\langle R_g \rangle$ for sequences generated in the $\langle R_g \rangle$-constrained evolution algorithm (Pearson correlation, p-value was too small to compute).



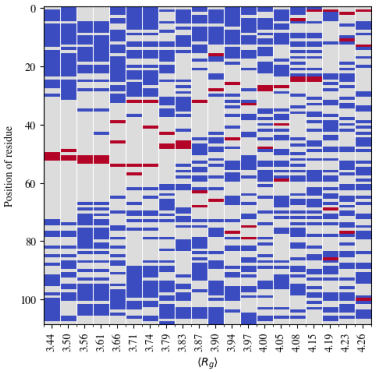
(b) Assigned clusters in the distribution of κ- and $\langle R_g \rangle$-values of the generations generated in the $\langle R_g \rangle$-constrained evolution. $\times$-markers represent cluster centroids.



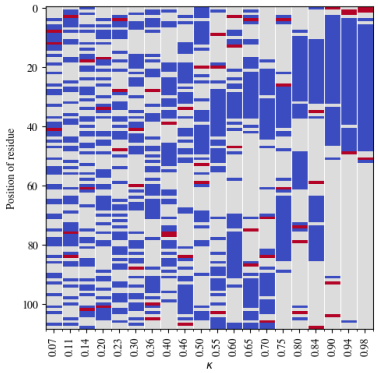
(c) Assigned clusters in the distribution of $\langle R_g \rangle$-values of the generations generated in the $\langle R_g \rangle$-constrained evolution. Dashed lines represent cluster centroids.



(d) Assigned clusters in the distribution of κ-values of the generations generated in the κ-constrained evolution. Dashed lines represent cluster centroids.



(e) The charge patterning in the cluster sequences from the $\langle R_g \rangle$-constrained evolution. The sequence with the $\langle R_g \rangle$-value closest to the cluster centres (see (a)) was chosen as the representative for the cluster



(f) The charge patterning in the cluster sequences from the κ-constrained evolution. The sequence with the κ-value closest to the cluster centres (see (a)) was chosen as the representative for the cluster

H Structural measures of H1.0 CTD and ProTα before and after binding

Peptide	State	$\langle R_g \rangle$ *Size*	ν *Expandedness*	Δ *Symmetry*	S *Shape*
H1.0 CTD	*Unbound*	3.9 nm	0.65	0.26	1.41
	Bound	3.2 nm	0.62	0.24	1.38
ProTα	*Unbound*	3.9 nm	0.65	0.28	1.53
	Bound	3.7 nm	0.63	0.26	1.43

Table 4 Differences in structural measures of the peptides in unbound and bound state, simulated at 290 mM ionic strength.

I Contact maps for wild type H1.0 CTD and ProTα binding

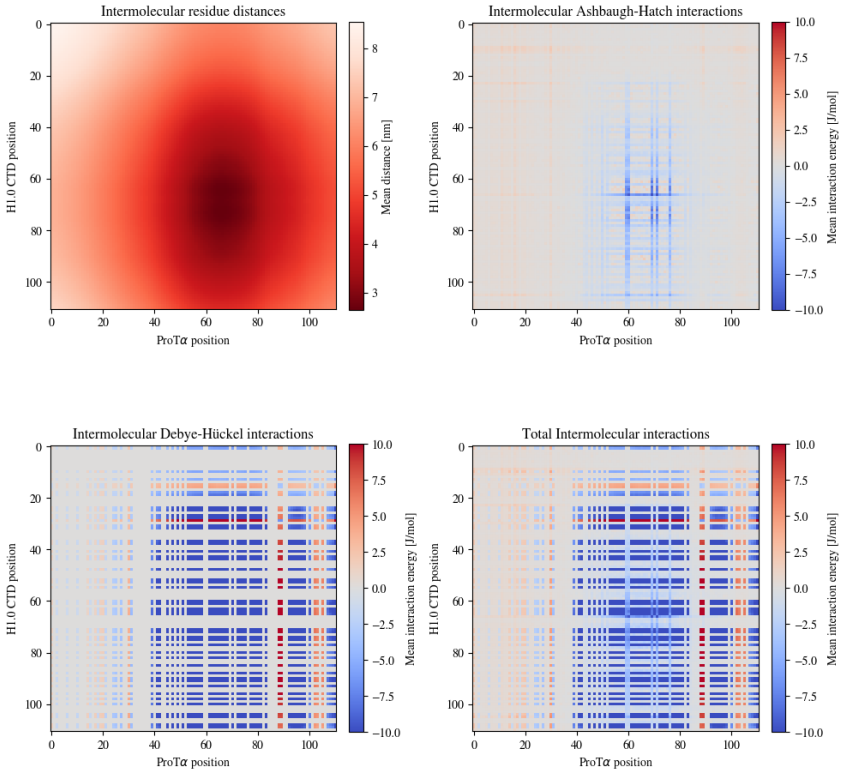
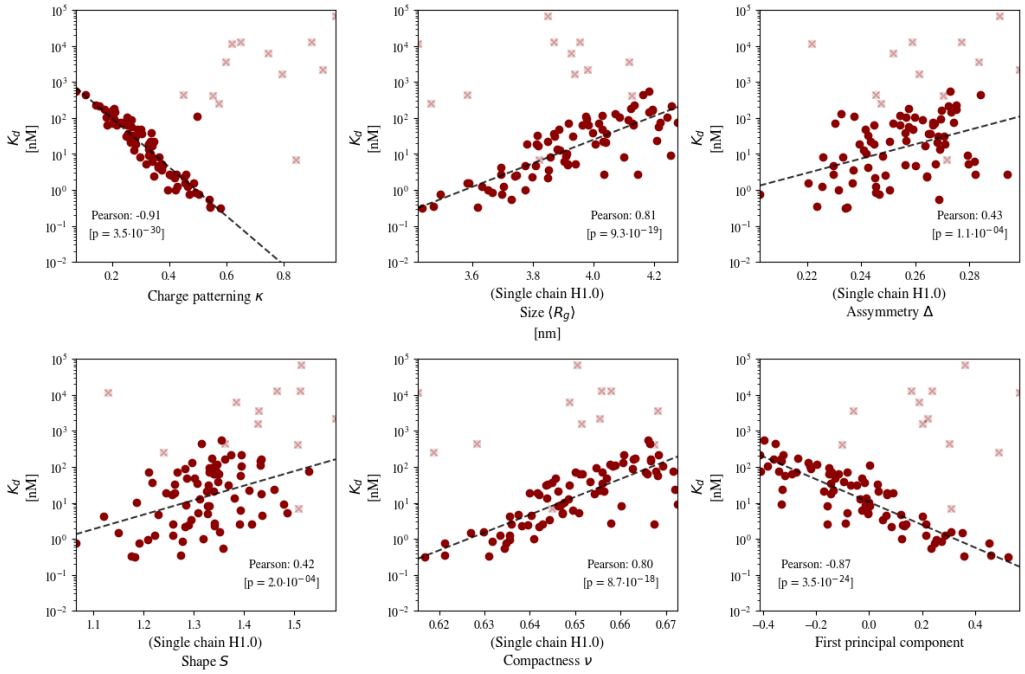
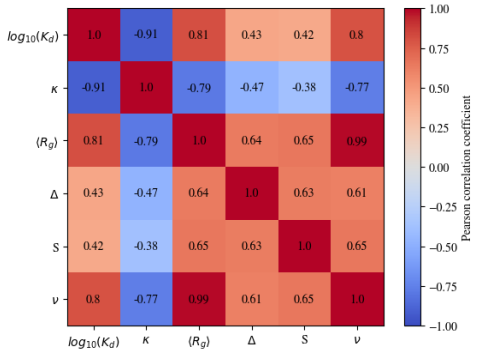


Figure 15 Interchain distance and interaction energy contact maps for bound H1.0 CTD and ProTα at 290 mM ionic strength

J Correlation of variant H1.0 CTD-ProTα K_d with various measures



(a) The correlation of K_d with various sequence and structural measures. Note that all variants have the same amino acid composition. The first principal component is from a principal component analysis including the other five measures: κ, $\langle R_g \rangle$, Δ, S, and ν. K_ds were calculated at 290 mM ionic strength. Data with no sampled unbinding event has a $\times$ marker, and was not included in correlation



(b) The correlation between K_d and various sequence and (H1.0 CTD single chain) structural measures. Note the multicolinearity between measures. K_ds were calculated at 290 mM ionic strength.