Summer Project Report

Protein clustering as a mechanism in gene regulation: a simulation study

Elena Espinosa Miñano - s2205640

Project Supervisor: Dr. Chris Brackley

June - July 2024

Abstract

Polymer physics models were used to conduct molecular dynamics simulations of protein posttranslational modifications (PTMs) with varying protein-chromatin and protein-protein attraction strengths. The effect these two factors have on the formation and dynamics of protein clusters in chromatin was investigated. We found that when proteins could switch "on" and "off" (undergo a PTM) and the protein-protein attraction was $4k_BT$; small, dynamic and short-lived "protein-only clusters" formed without being bound to binding sites along the chromatin. When the protein-protein attraction was lowered to $2k_BT$, this did not occur; however, the size of the largest cluster was found to increase. Further simulations are required to interpret this result among others discussed in the report.

Personal Statement

This summer project gave me the opportunity to gain insight into the research process, see how investigations are done in practice and gain some research experience within academia. I found it extremely interesting to learn about the field of biophysics, see how computational models are applied to this discipline and see the links biophysics has to prior Junior Honours courses like Thermodynamics and Statistical Mechanics.

Doing this project helped me develop my communication skills by having regular meetings with my project supervisor and interacting with my supervisors' research group. I also had the chance to join a meeting with a visiting, experimental biophysics lab and see how interdisciplinary fields work together. Finally, the hands-on computational project helped me gain confidence in coding, using the Linux terminal, CPU clusters and making bash scripts.

All in all, this was an amazing experience which will undoubtedly help me clarify my future career path.

Lay Summary

DNA is a two meter long, tightly packed molecule that lies inside the cell nucleus. It interacts with itself and proteins in order to compact further and to carry out biological functions.

In this project, we created a simple model where DNA (represented as a chain of beads) interacts with nearby beads (proteins). Proteins in our model are attracted with some strength to certain "sticky" parts of the DNA to which they bind to when close. When a protein attaches to two parts of the DNA forming a loop, it prompts other proteins to go to that region and join the initial protein, as there is more DNA to which they can bind. This forms a protein cluster.

Making measurements on these protein clusters we found that when proteins are also attracted to other proteins, protein clusters are able to form without being bound to sticky sites in the DNA.

Contents

1 Introduction

DNA is a highly structured and organised molecule inside the cell nucleus which encodes the genetic information of an organism [1][2]. DNA coils around histone proteins to form nucleosomes, these compact together to form the chromatin fibre which interacts with itself and proteins to form chromatin loops, domains and compartments [2] (see Figure 1).

To shed light on this non-random organisation of the genome and how its structure may be linked to cell function [2]; many polymer physics models (see [3, 4, 5]) have been created in order to simulate the binding of transcription factors (TFs)[1] and other transcriptional machinery to specific regions along the chromatin fibre. These have been successful in helping to understand how chromatin binding proteins can affect and drive genome structure, protein phase separation as well as regulate gene expression [2][6].

For example, it has been observed, firstly in simulations [7, 8], and not long ago in-vitro [9], that proteins can bind to two distant parts along the chromatin fiber to create a loop. This increase in chromatin density at the point of intersection induces other proteins to bind there

[1]TFs: proteins involved in regulating gene transcription [6].

leading to the formation of a protein cluster. This observation was termed by authors in Ref. [7] as the bridging-induced attraction (BIA).

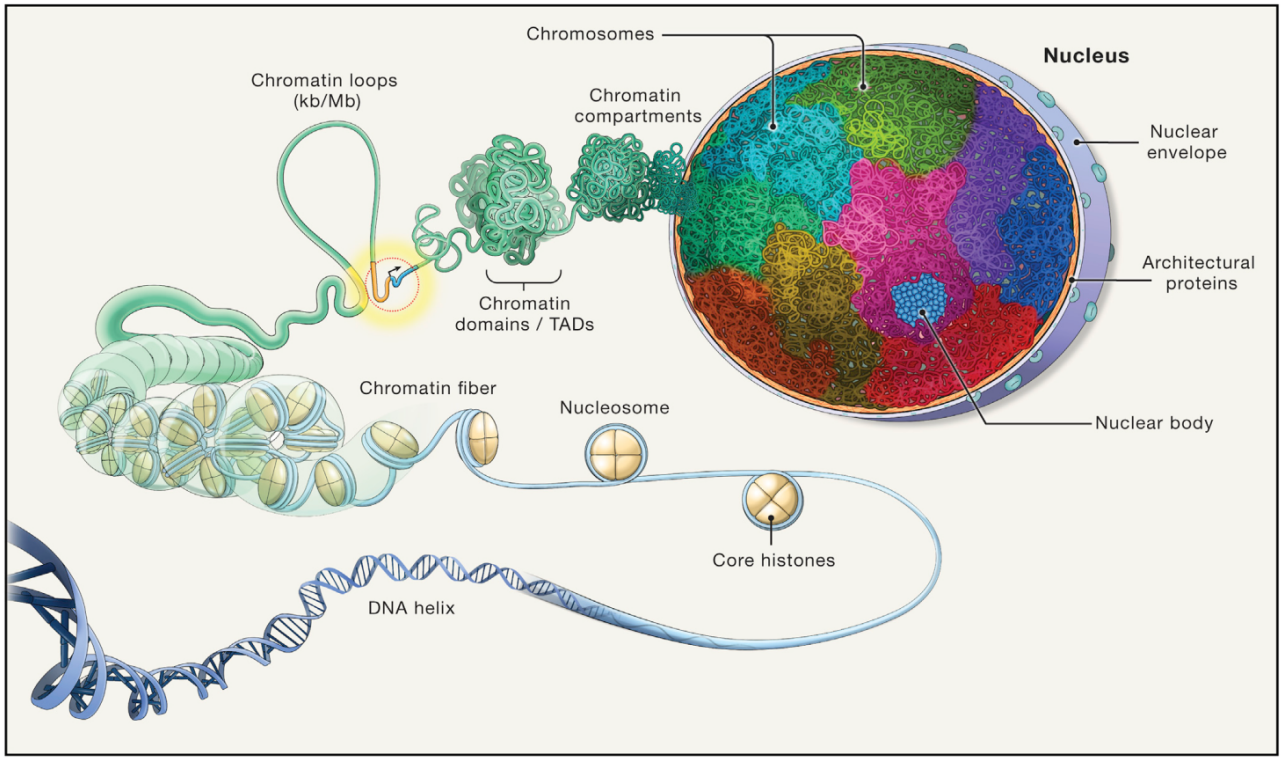


Figure 1: Figure reproduced from Ref. [2] showing how DNA is folded and compacted inside the cell nucleus.

In Ref. [10], the effect of protein post-translational modifications (PTMs)[2] was studied by modelling the chromatin as a polymer chain with specific binding and non-binding sites to which transcription factors (TFs), modelled as diffusing beads, could bind to when active. During the simulation, TFs could probabilistically "switch off" (or "on" if off), representing a protein which had undergone a PTM. When switched "on", proteins could bind to chromatin binding sites; when "off", they could not bind. In this paper, the authors show that protein PTMs affect the formation and dynamics of proteins clusters by limiting their growth.

In Ref. [12], the effect of varying the strength of protein-protein and protein-chromatin interactions was studied, again with a similar chromatin-polymer modelling scheme mentioned above. In this paper, the authors show that when there is *protein-chromatin* attraction, it leads to the BIA; while, when there is *protein-protein* attraction, it leads to the liquid-liquid phase separation (LLPS) of protein clusters. Lastly, they show that when both attraction types are present, the polymer (chromatin) becomes "absorbed" into a large protein droplet.

A logical next step from these two investigations (Ref. [10] [12]) is to combine these studied factors into a single model.

[2]PTMs: chemical modifications that change the properties of proteins affecting their affinity for binding sites [11]. In our models this is simulated by proteins switching "on" or "of" (can or cannot bind to chromatin).

In our project, we aim to achieve this by combining the **switching of proteins** that bind to specific sites along the chromatin with **varying protein-chromatin and protein-protein attraction strengths**. Then, making measurements on our simulation results, we seek to observe how these factors affect the formation and dynamics of protein clusters in chromatin.

Specifically, we will study these effects on a region of the mouse genome around the *Sox2* gene. This gene has been very well studied and is of particular interest to biologists as it encodes for a TF called SOX-2, involved in regulating the expression of genes associated with the development of embryonic stem cells [13]. In Ref. [14], authors observed that without the presence of the *Sox2* super-enhancer (a large enhancer region), clustering of polymerase proteins near the *Sox2* gene did not occur and the gene was not expressed. This suggests that the presence of the many protein binding sites which form the *Sox2* super-enhancer (thought to nucleate a protein cluster leading to liquid-like phase separation) is crucial for the expression of the *Sox2* gene.

In conclusion, we will use polymer physics models to simulate molecule interactions and so provide insight into how protein clusters form and evolve under varying model conditions. In the following sections, we will explain how the chromatin-protein system is modelled, describe the specific models and simulations conducted, and lastly, present the overall findings and conclusions of the study.

2 Methods

In this project, the chromatin fibre is represented as a polymer chain of N spherical beads with diameter σ. As done in Ref. [10], adjacent beads along the chain of polymers are joined together by finite-extensible non-linear (FENE) bonds and the cosine angle potential is added between triplets of beads to give the bending rigidity of the polymer. For the exact form of these potentials see Refs. [10][15].

The Weeks-Chandler-Anderson (WCA) interaction potential [15], also known as the shifted, truncated Lennard Jones (LJ) potential, seen in Eq. 2.1, provides the steric interaction between any pair of beads, ensuring they do not overlap.

$$U_{WCA}(r_{ij}) = \begin{cases} 4\epsilon_{ij}\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] + \epsilon_{ij}, & \text{if } r_{ij} < r_{cut} \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

where, U_{WCA} is the WCA potential, $r_{ij} = |r_i - r_j|$ is the distance between beads i and j, σ is the mean diameter of beads i and j (since all beads have the same diameter this is just the diameter of a bead) and is also our simulation length unit, ϵ_{ij} is interaction energy between beads i and j and $r_{cut} = 2^{1/6}\sigma$.

Transcription factors (TFs) are modelled as diffusing spherical beads of diameter σ. Active TFs (type 4) interact with specific binding sites (types 2 and 3) along the polymer according to the shifted and truncated LJ potential [15]:

$$U_{LJ-TS}(r_{ij}) = \begin{cases} U_{LJ}(r_{ij}) - U_{LJ}(r_{cut}), & \text{if } r_{ij} < r_{cut} \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

where,

$$U_{LJ}(r) = 4\epsilon_{ij}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.3}$$

where, U_{LJ} is the LJ potential, U_{LJ-ST} is the shifted and truncated LJ potential, $r_{cut} = 1.8\sigma$ and the rest of the symbols have been previously defined.

A molecular dynamics simulator known as LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator, [16]) is used to simulate the evolution of the position of beads in the system according to the Langevin equation [17]:

$$m\frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}^2t} = -\boldsymbol{\nabla}U_{total} - \gamma\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} + \sqrt{2\gamma k_B T}\boldsymbol{\eta}(t) \tag{2.4}$$

where, m is the mass of a bead, $\mathbf{x}$ its position, U_{total} is the total potential energy of the system, γ is the friction or drag coefficient, $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$ is the velocity of the bead, $\gamma\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$ is the drag force on the bead due to the solvent (nucleoplasm), k_B is the Boltzmann constant, T is the temperature

and $\boldsymbol{\eta}(t)$ is a noise term representing random forces acting on the bead. LAMMPS outputs the state of the system every 5000 time-steps.

In our simulation, we model the dynamics of a 5 Mbp chromatin region of chromosome 5 in mouse embryonic stem cells. A 5000 bead long polymer is used to represent the chromatin fibre with a single bead corresponding to ∼1 kbp of chromatin. The region of chromatin simulated contains both the locus of the *Sox2* gene as well as its super-enhancer.

Using data from ATAC-seq [18] and CHiP-seq [19] experiments obtained from the ENCODE project [20], specific beads along the polymer are inferred to be binding sites for the TFs (see Table 1 for more details).

These sites, labelled as type 2 and type 3 in Figure 2, represent sections along the chromatin to which proteins have an attraction to with some interaction strength ϵ, specified in Table 1.

The entire chromatin and proteins system is confined to a sphere with a density matching that of chromatin in real cells. A schematic diagram of the general chromatin and proteins model described above can be seen in Figure 2.

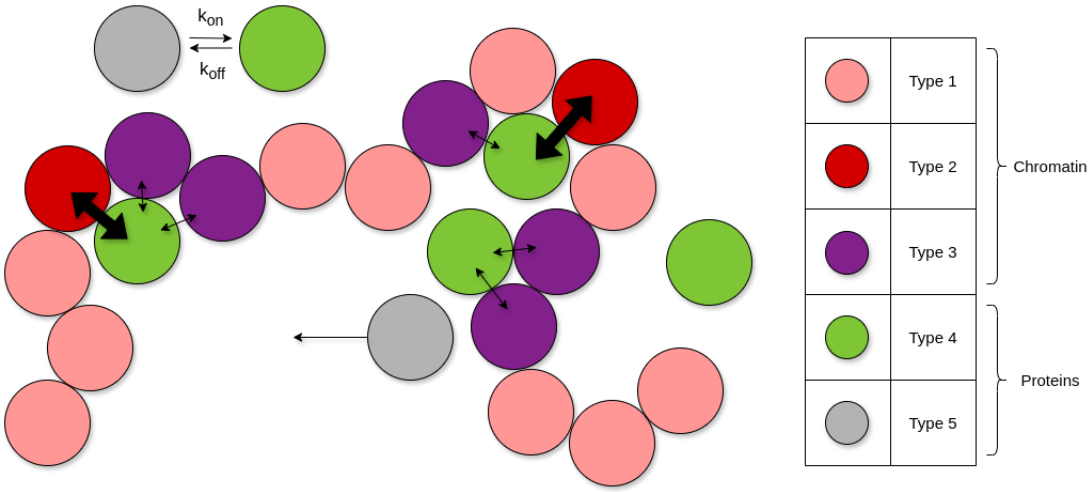


Figure 2: A schematic diagram showing a section of chromatin modelled as a chain of three different bead types (type 1, 2 and 3). Active (type 4) and inactive (type 5) proteins interact with each other and the polymer chain. Bold, double-headed arrows represent the strong attraction active TFs (green) have to highly accessible DNA sites (red) along the chromatin fibre. Thin, double-headed arrows represent the weak attraction active TFs have to enhancer-promoter regions (purple). Single arrow emanating from inactive protein (grey) represents diffusion from a protein cluster once an active TF switches to inactive.

To investigate the effect of protein post-translational modifications (PTMs) [11], as done in Ref. [10], we incorporated "switching proteins" to several models in our simulation (shown in green and grey in Figures 2 and 3). These proteins can probabilistically switch between an "on" and an "off" state at regular time intervals during the simulation (on average every 500 time-units = 5×10^4 simulation time-steps, such that they switch with rate $k_{on} = k_{off} = 2 \times 10^{-5}$ inverse time-steps); when proteins are "on" ("off") their attraction to specific binding

sites (type 2 and 3) is "on" ("off"). Therefore, a protein initially bound to a binding site might stochastically "switch off" during the simulation resulting in that protein diffusing away from the binding site.

	Bead type	Number of beads	Bead type - Type 4 interaction strength (kBT)	Experimental input	Represents:
Chromatin	Type 1	4669	1 - steric	---	Normal DNA
	Type 2	81	8* - strong	ATAC-seq	Highly accessible DNA sites to which active TFs have a high affinity for
	Type 3	250	4* - weak	CHiP-seq	DNA enhancer-promoter regions with high levels of histone modification H3K27ac - weak affinity to active TFs
Proteins	Type 4	300*	1* - steric (default)	---	Active transcription factors (TFs)
	Type 5	300*	1 - steric	---	Inactive transcription factors (TFs)

Table 1: Summary of key data for each type of bead in LAMMPS simulations. The experimental input column gives the name of experiment performed in order to locate the binding sites along the polymer. The right-most column gives the biological representation of each bead. Asterisk * indicates values which are model dependent. See Table 2 and text for more details.

First, we implemented several models to systematically compare the effects of including switching proteins and varying the protein-protein attraction. The model differences are summarised in Table 2.

Models		Interaction strength (kBT)		
Switching OFF	Switching ON	Type 2 - Type 4	Type 3 - Type 4	Type 4 - Type 4
1	5	8	1	1
2	6	8	4	1
3	7A	8	4	4
4	7B	8	4	1-8 / 2
	0	1	1	1

Table 2: Summary of the variation in interaction strength between chromatin binding sites (type 2 and 3) and active proteins (type 4) for the different models. Models 1-4, contain 300 non-switching proteins; models 0 and 5-7 contain 600 switching proteins, so that on average, at every time-step of the simulation there are 300 active proteins in the system, making models 1-4 comparable to 5-7.

Then, we investigated the effect of varying the switching interval between 300 and 700 time-units for Model 7A (contains 600 switching proteins that can bind strongly to type 2, weakly

to type 3 and weakly to each other). Finally, we looked at the effect of varying the number of proteins between 300 and 600 for models 5-7.

All the models and investigations described above were implemented in LAMMPS scripts (available here) and the dynamics of the systems visualised using Visual Molecular Dynamics (VMD, [21]). A snapshot of the system visualised using VMD can be seen in Figure 3.

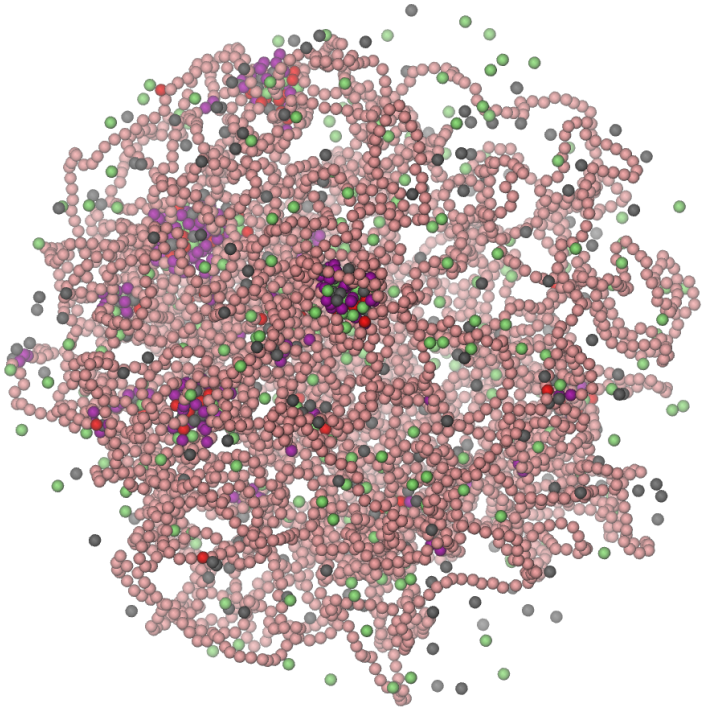


Figure 3: A snapshot of the system for model 6 at a single time-step during the simulation. Different bead colors represent the bead types specified in Figure 2. Cluster of proteins (grey and green) can be seen forming around binding sites (red and purple).

Measurements

Python code was implemented to make measurements on the output files from the simulations (made available here). The main function the code utilises for making further measurements is based on the DBSCAN clustering algorithm [22] (see Appendix A).

To make measurements on the clusters several definitions had to be made.

Importantly, in this project, a cluster was defined as having at least 2 active TFs within the *cluster threshold* distance of 2.3σ (ie. type 4 beads i and j formed a cluster if $r_{ij} < 2.3\sigma$). This threshold distance was chosen so as to allow for a protein cluster to form with a polymer bead in between the TFs.

A protein was considered bound to a polymer, and vice-versa, if these were within the *bound threshold* distance of 1.8σ ($r_{ij} < r_{cut} = 1.8\sigma$).

Once protein clusters were identified and given a specific `cluster_ids`, further measurements on the system included: obtaining the total number of clusters in the system, their mean and largest size, the number of proteins bound to any polymer bead and so on.

3 Results and Discussion

3.1 Determining if the system had reached steady-state

After running the simulations, plots of the measurements versus time-step of the simulation were made in order to see when the system had reached a non-equilibrium steady state. An example plot is seen in Figure 4.

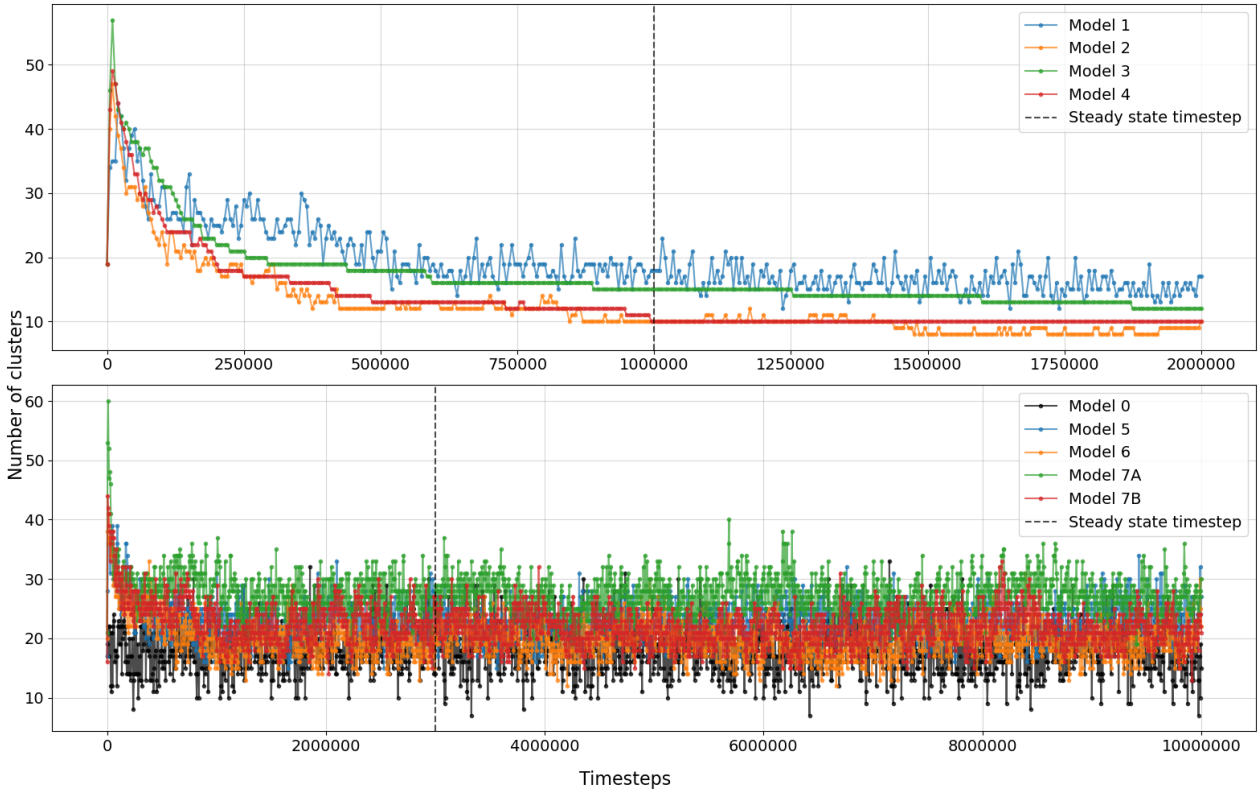


Figure 4: Number of clusters in each model vs. time-steps in simulation. (Top) Models 1-4 (300 non-switching proteins, length of simulation: 2×10^6 time-steps). (Bottom) Models 0, 5-7 (600 switching proteins, length of simulation: 10×10^6 time-steps). Dashed vertical lines, indicate the time-step from which steady state is assumed to be reached. Only values past this time-step are used in obtaining the mean and standard deviation for the measurements, see Figure 5.

As seen in the top subplot in Figure 4, the number of clusters (NoC) in models 1 and 2 decrease steadily until $\sim 1 \times 10^6$ time-steps, from which they seem to oscillate about some stable steady-state value. A different general trend is seen for models 3 and 4 where the NoC seem to decrease in jumps and it is unclear if they reach a steady state within the duration of the simulation. This can be understood in the following way: firstly many small protein clusters form (proteins bind to type 2/3 sites), diffusing proteins join to these clusters due to the bridging-induced-attraction making them more stable, these stable protein clusters then diffuse around until bumping into another cluster and coalescing.

With the protein-protein attraction we expect clusters to grow, stabilise, merge and eventually coarsen into one large cluster. However, as the clusters grow their diffusion slows down. The process of coalescence can be very slow, so simulations would need to be run for much longer in order to confirm this.

The bottom subplot of Figure 4, shows that for models 0, 5-7 steady-state is reached undoubtedly from $\sim 3 \times 10^6$ time-steps. With switching, we expect that clusters will be much more dynamic due to proteins continually joining and leaving clusters and so the system for the models should reach a steady state faster.

Hence, as we suspect that our current simulations for models 3 and 4 are not reaching steady state, we will not look into these further, but instead focus on understanding models 5-7 and comparing them to models 1 and 2. Model 0, will act as a control for models 5-7.

3.2 Model comparison

First of all, we will compare and attempt to interpret the properties of the system for the models in consideration.

An important simplification that was made in obtaining the standard error on the mean (SEM) for our steady-state measurements was to treat all measurements from the critical time-step onwards as independent. However, this cannot be the case as the system doesn't change drastically every 5000 time-steps and so consecutive measurements must be slightly correlated and cannot be truly independent from each other.

Given that: $SEM = \frac{STD}{\sqrt{M}}$, where STD is the standard deviation and M is the number of independent measurements. With our simplification we overestimate M and hence the obtained SEM is actually an underestimate of the "true SEM". In order to obtain this "true SEM" we would need to have calculated the correlation function between our measurements. Due to time constraints, we decided to use the underestimated SEM and treat it as a lower bound for the "true SEM".

Looking at the steady-state bar plots in Figure 5, we notice that despite the beads in model 0 having no attraction to each other, around 18 clusters of mean size ~ 2 form during the simulation. This important result suggests that, due to the chromatin and proteins system being confined to a sphere; clusters of proteins form just due to proteins diffusing past each other (recall that the *cluster threshold* is 2.3σ so proteins do not need to be touching in order to form a cluster). As the proteins forming these "fake" clusters are not bound together, we will call these "transient clusters" to distinguish them from the "real clusters". Another thing to note is that, we expect the number of transient clusters to decrease and their effect to become

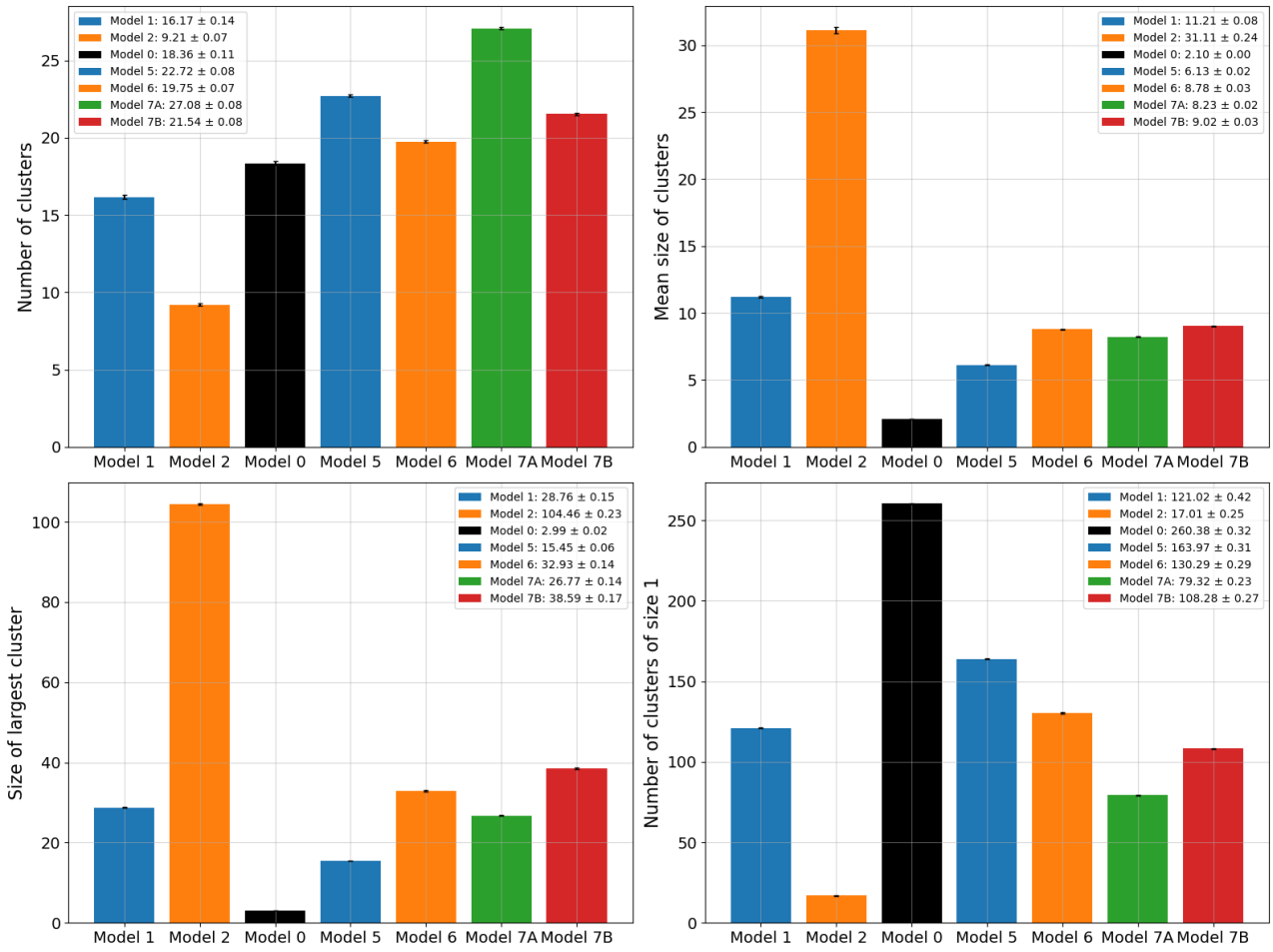


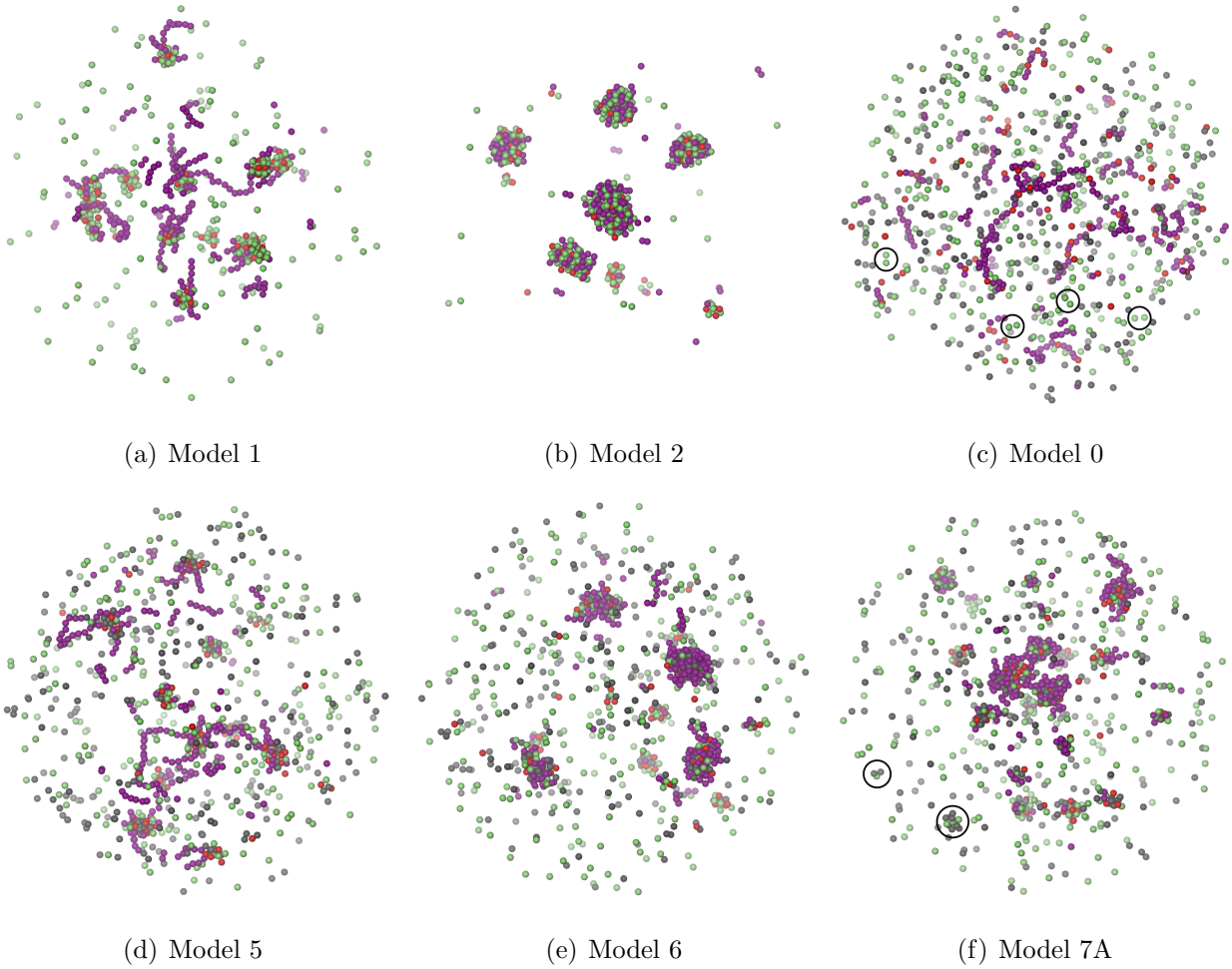
Figure 5: Steady-state subplots of the mean ± 1 SEM for various measurements on clusters forming in the system for the different models described in Section 2. See Appendix B for steady-state plots of other measurements.

negligible as proteins begin to interact with the polymer and with themselves (as in the rest of the models).

Going from model 1 to 2, we see from Figures 5 and 6 that the addition of weakly attractive sites (shown in purple in the VMD snapshots) roughly halves the number of clusters forming in the system and triples their mean size. On average during the simulation the largest cluster in model 2 contains around a third of all type 4 proteins!

This clustering effect has been previously discussed in Ref. [8], in which the authors propose that, due to proteins having a larger "target area" to which they can bind to (type 2 and type 3 beads), it is easier and more probable for a diffusing protein to encounter a "sticky site", bind to it and quickly become stabilised by nearby "very sticky" sites and any surrounding attractive beads. This means that when proteins join clusters it will be very unlikely for thermal fluctuations to kick them out of their stable cluster.

Going from models 1 to 5, and 2 to 6, in Figures 5 and 6; its clear to see that when proteins can switch between on and off (in models 5 and 6), more clusters of smaller size seem to form.



(a) Model 1 (b) Model 2 (c) Model 0

(d) Model 5 (e) Model 6 (f) Model 7A

Figure 6: Subplots show VMD snapshots of the proteins-chromatin system for the investigated models. Normal, type 1 polymer (shown in pink in Figures 2 and 3) is not shown to facilitate seeing protein clusters. (a)-(b) Shows the effect of having type 3 weakly attractive sites. Type 3 beads act to stabilise protein clusters. (c) Shows "transient clusters" circled in black. (d)-(e) Shows the effect of incorporating switching to models 1 and 2 (f) Shows "protein-only" clusters forming when protein-protein attraction is large enough (circled in black).

Again, the effect seen here, has been examined in Ref. [10], where the authors suggest that with switching there is a constant flux of proteins joining and leaving clusters, making these more dynamic but also limiting their size and stopping their growth.

Looking at the cluster size distributions in Figure 7, we observe that with switching there is a continuous spread of sizes instead of clumps of favoured cluster sizes (as for models 1 and 2). Currently, we do not have a good explanation for why this occurs; more simulations on a larger system would be needed to try and understand.

Going from model 6 to 7A (which contains protein switching **and** protein-protein attraction $4k_BT$) we see in Figures 5, 6 and 7 that the number of clusters forming in the system increases, their mean and largest size decreases and the number of size 1 clusters decreases significantly.

A possible interpretation for this observation could be due to an interplay between the

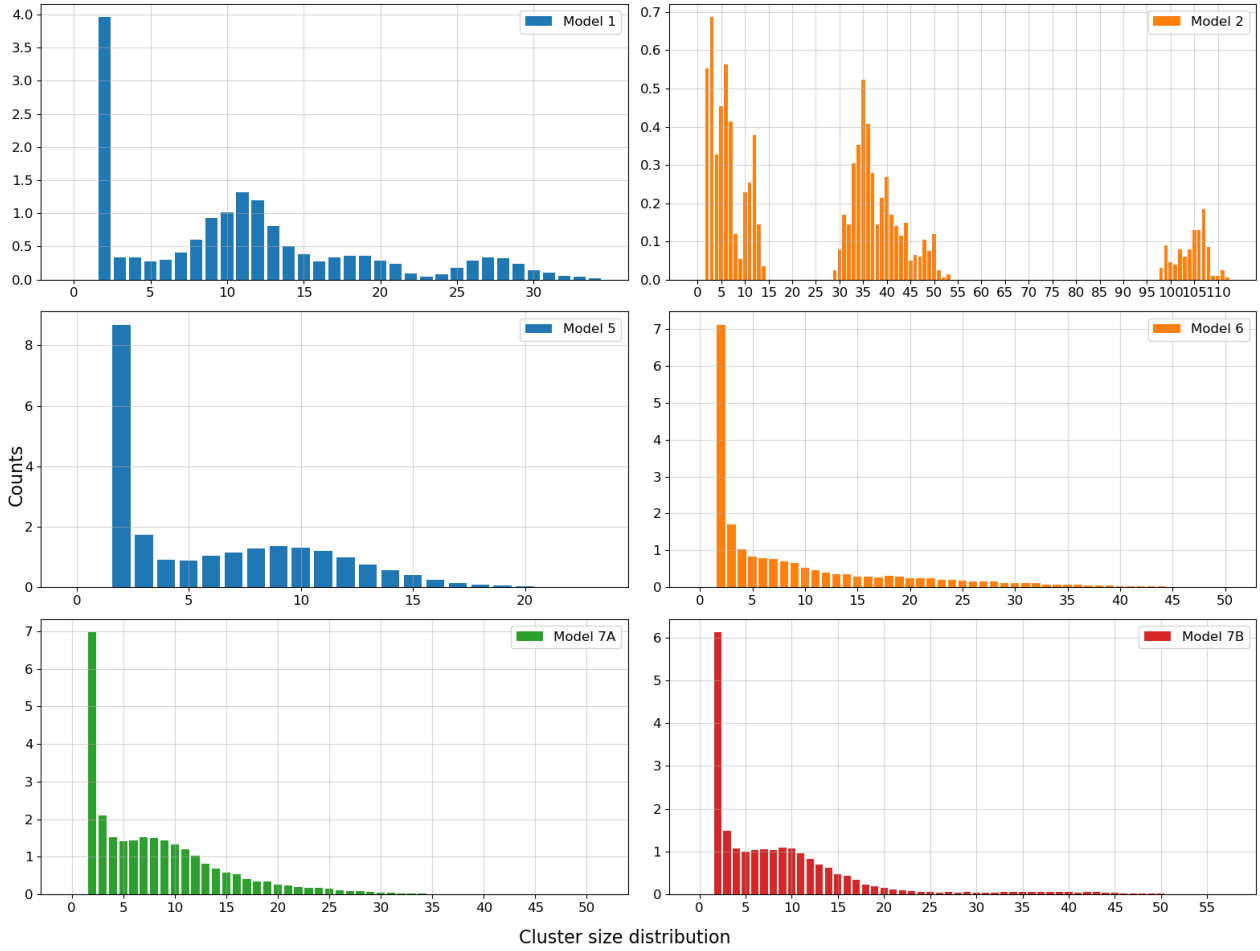


Figure 7: Steady-state histogram subplots showing the cluster size distribution for various models. The height of the bars represent the normalised cumulative number of counts for each cluster size every 5000 time-steps in the simulation.

protein-protein attraction, the confined system and the switching TFs. Diffusing proteins in model 7A are now attracted to a lot more beads ($\sim$ 300 other active proteins) with the same attraction strength as to "weakly sticky" type 3 sites. This could mean that due to the confined system, proteins simply diffusing past each other as observed occurring in model 0 (recall that with no attraction at all $\sim$ 18 transient clusters were able to form), might now bind to each other forming a real "protein-only" cluster. However, because of the random switching, these "protein-only" clusters are expected to be small, very dynamic and short-lived, which seems to agree with the result of a decreased mean cluster size.

The protein-protein attraction in model 7 might also result in the formation of smaller clusters. In model 6, large stretches of sticky sites are needed in order to form large protein clusters. The protein-protein attraction in model 7A might mean that only a few sticky sites are needed, as once proteins bind to these, then they themselves can attract further proteins. This results in proteins being shared more evenly among the binding sites and hence seeing smaller clusters.

Regarding model 7B (like 7A but with lower protein-protein attraction $2k_BT$) we notice that it doesn't form as many clusters as in 7A, but the mean and largest size of clusters is actually bigger. The histogram plot shows that models 7A and 7B have similar distributions, with 7B having the "bump" around cluster size 9 (instead of 7 in 7A) and the "tail" of the distribution reaching bigger cluster sizes.

This might mean that with a smaller protein-protein attraction ($2k_BT$ instead of $4k_BT$), proteins cannot form clusters on their own and so need to bind to the chromatin in order to cluster, leading to larger clusters. But it is not yet clear why with a weaker protein-protein attraction we could expect the size of the largest cluster forming in the system to be bigger.

Further simulations are thus required to investigate and support this proposal.

3.3 Effect of varying the number of proteins in models 5-7

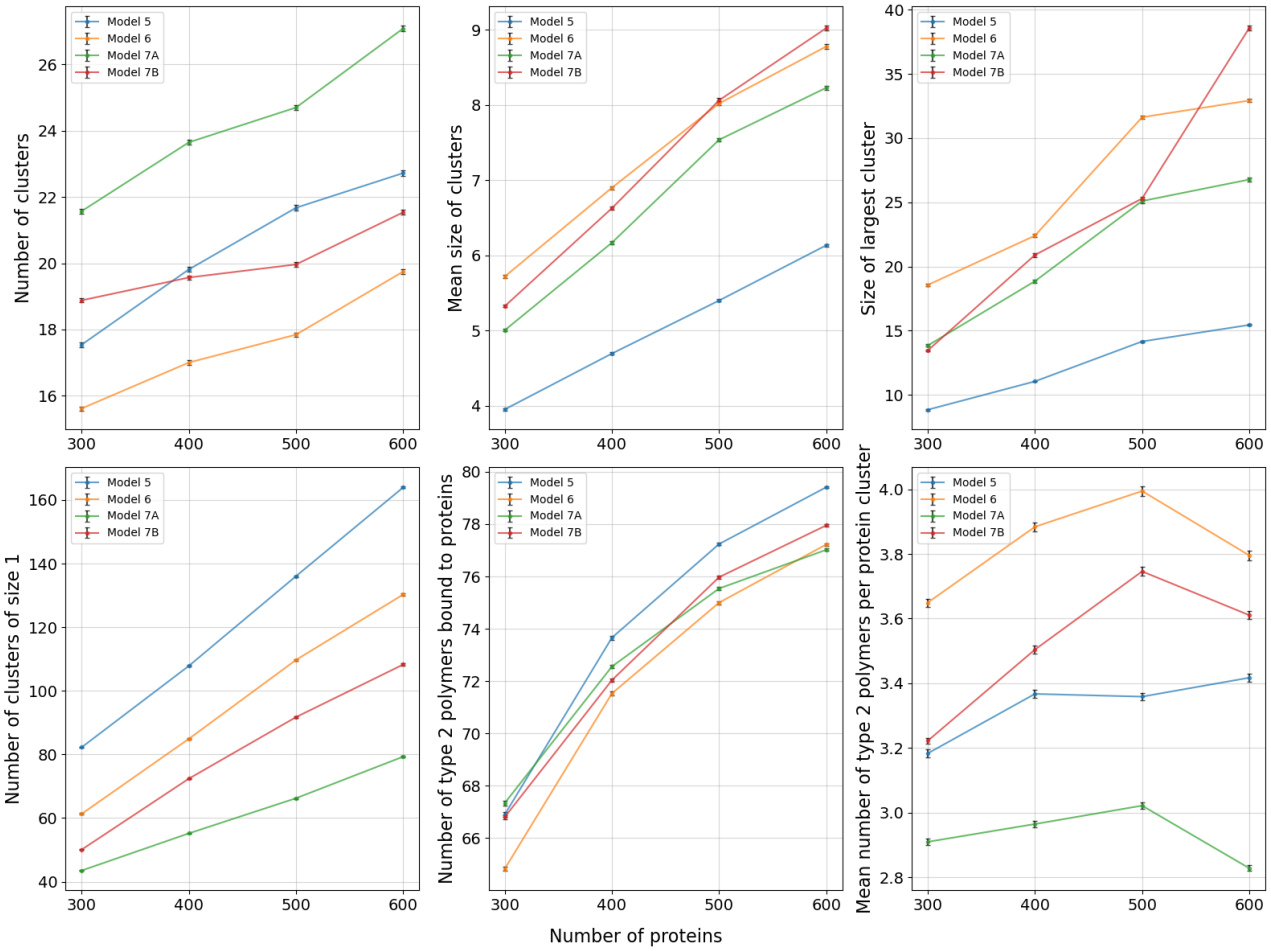


Figure 8: Steady-state subplots of the mean ± 1 SEM for various measurements when varying the number of proteins for the different models. Coloured lines are shown as a guide to the eye.

Figure 8 shows how protein density affects the properties of the system in steady state for the different models. Recall from Table 2 that, model 5 proteins are only attracted strongly

($8k_BT$) to type 2 sites, model 6 proteins are additionally attracted weakly ($4k_BT$) to type 3 sites and proteins in model 7 are also attracted to each other with attraction strength $4k_BT$ (in 7A) and $2k_BT$ (in 7B).

From Figure 8 we see some expected results mentioned in the earlier subsection. The first four subplots show that, as you increase the number of proteins (NoP), the number of clusters (NoC) forming increases, their mean and largest size increases and the number of size 1 clusters increases almost linearly. The differences between the models are consistent with those seen earlier in Figure 5.

We can also see that as the protein density in the system increases, more type 2, "strongly sticky" polymer sites have proteins bound to them. There are only 81 such sites in the system, so with more proteins there is a higher chance of them diffusing past a type 2 bead and binding to it. As the NoP increases we start to see that the "strongly sticky" sites are becoming saturated, as would be expected.

We cannot draw any strong conclusion from the last subplot as it shows no clear trend in how increasing the NoP might affect the mean number of type 2 polymer beads in a protein cluster. However, we notice that there is a clear jump between models 5, 6 and 7A (with 7B between 5 and 6). The fact that on average there are less type 2s in protein clusters for model 7A might point towards the presence of "protein-only" clusters forming in this model as well as the absence of these in model 7B (where protein attraction is not strong enough for proteins to cluster on their own). Doing repeat simulations and averaging over these would help to clarify and confirm this.

3.4 Effect of varying the switching interval in model 7A

Figure 9, shows the effect of increasing the time interval between the time-steps proteins can probabilistically switch "on" or "off" (ie. decreasing the switching rate), specifically for model 7A.

The first four subplots show that as the switching interval increases (proteins remain longer in either an "on" or "off" state) the NoC forming decreases, their mean and largest size increases and the number of proteins not in clusters decreases. This agrees with our expectations that as the switching interval becomes longer, clusters should become less dynamic as the turnover of proteins becomes less frequent allowing for the formation of larger clusters. In the limit of an extremely large switching interval, we expect model 7A to give results similar to model 3 (no switching). More simulations would need to be run in order to confirm this.

The last two subplots show that as the switching interval increases, more proteins are bound

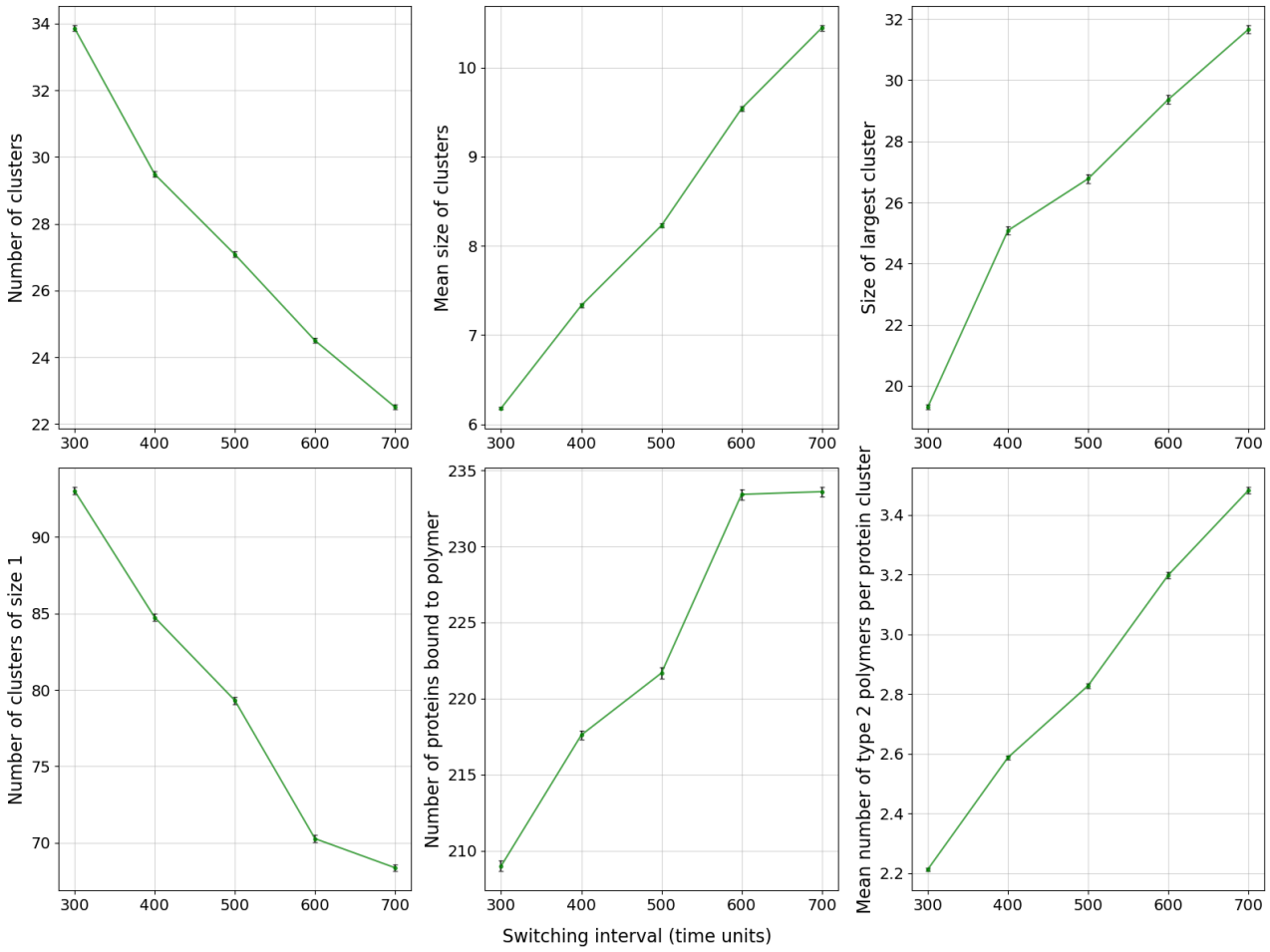


Figure 9: Steady-state subplots of the mean ± 1 SEM for various measurements when varying the switching interval for model 7A. Green lines are shown as a guide to the eye.

to the polymer and protein clusters contain more type 2 polymer beads. The first observation suggests that because proteins remain in the "on" state for longer they are also bound to the polymer for longer. The last observation makes sense considering that the mean and largest cluster size follow a similar increasing trend.

4 Conclusion

To conclude, in this project we set out to investigate the effects of switching proteins binding to specific sites along the chromatin region of the mouse *Sox2* gene while varying the protein-chromatin and protein-protein attraction strengths.

This was achieved by creating polymer physics models with different characteristics and comparing them to interpret the properties of each system. We then looked at the effect of varying the number of proteins in the system and the rate at which proteins could switch "on" and "off" to see if the results agreed with our expectations.

Firstly, models 1 and 2, reaffirmed the effect (discussed previously in Ref. [8]) that the presence of large sections of "weakly sticky" sites (Model 2) act to stabilise proteins binding there and allow them to form large and stable clusters. Then, in the models that contained protein switching (5-7), we observed that the protein clusters forming were more dynamic but could not grow indefinitely and had a limited size (as seen in Ref. [10]).

We came across the surprising formation of "transient clusters" in Model 0 (no attractions) due to the chromatin-proteins system being confined and proteins diffusing past each other.

Lastly, we saw that with switching TFs and a $4k_BT$ protein-protein attraction (Model 7A), dynamic and short-lived "protein-only" clusters formed; while, with the lower protein-protein attraction of $2k_BT$ (Model 7B), these did not appear. Repeat runs of these simulations should be conducted to validate these results and further simulations are required to investigate the implications this protein clustering mechanism might have for the expression of the *Sox2* gene.

During the course of the project we discovered some issues which had an impact on our findings but which due to time constraints, could not be resolved within the length scale of the project. For example, given more time we would have changed the clustering algorithm so that the "transient clusters" where not counted as real clusters in our measurements. This would have been done by having a specific cluster threshold depending on which type of neighbouring atom each protein was bound to. This change would mainly affect the number of clusters and the number of proteins (or type 2) bound to polymer (or proteins) measurements.

To see if the decrease in the fraction of clusters (see Fig. 10 in Appendix B) bound to the polymer for model 7A corresponded to "protein-only" clusters we could have measured the number and the mean size of clusters that were not bound to the polymer.

Another thing we could implement in the future is to add spatial dependence to the switching of proteins. In reality, switching requires the protein to interact or collide with an enzyme. These enzymes are thought to be bound to the chromatin at enhancer sites. So simulations could be modified so that for proteins to switch, proteins need to be at enhancer sites.

Given more time, we also would have done repeat runs of the simulations and run the simulations on a larger system to validate our results and reach stronger conclusions.

5 Acknowledgements

I wish to thank Chris Brackley for his patient guidance, enthusiasm and support throughout the project and in answering all my questions. Thank you to the School of Physics and Astronomy Career Development Summer Scholarships for funding this project. Lastly, I am grateful to my family for their unending encouragement in everything I undertake.

6 References

[1] B. Alberts, A. Johnson, J. Lewis, *et al.* "Molecular Biology of the Cell". 4th edition. New York: Garland Science (2002).

[2] T. Misteli. "The Self-Organizing Genome: Principles of Genome Architecture and Function". *Cell* **183**, 28-45 (2020).

[3] A. Buckle, C. A. Brackley, S. Boyle, *et al.* "Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci". *Molecular Cell* **72**, 786-797 (2018).

[4] S. Bianco, D. G. Lupiáñez, A. M. Chiariello, *et al.* "Polymer physics predicts the effects of structural variants on chromatin architecture". *Nature Genetics* **50**, 662–667 (2018).

[5] M. Chiang, G. Forte, N. Gilbert, *et al.* (2022). "Predictive Polymer Models for 3D Chromosome Organization". In: S. Bicciato and F. Ferrari (eds.) *"Hi-C Data Analysis: Methods in Molecular Biology"*. New York: Humana. **2301**, 267–291.

[6] D. S. Latchman. "Transcription factors: an overview". *International Journal of Experimental Pathology* **74**, 417-422 (1993).

[7] C. A. Brackley, S. Taylor, D. Marenduzzo, *et al.* "Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization". *Proceedings of the National Academy of Sciences of the United States of America* **110(38)**, E3605–E3611 (2013).

[8] C. A. Brackley, J. Johnson, D. Marenduzzo, *et al.* "Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains". *Nucleic Acids Research* **44**, 3503–3512 (2016).

[9] J. K. Ryu, C. Bouchoux, H. W. Liu, *et al.* "Bridging-induced phase separation induced by cohesin SMC protein complexes". *Science Advances* **7**, eabe5905 (2021).

[10] C. A. Brackley, B. Liebchen, D. Michieletto, *et al.* "Ephemeral Protein Binding to DNA Shapes Stable Nuclear Bodies and Chromatin Domains". *Biophysical Journal* **112**, 1085-1093 (2017).

[11] S. Ramazi and J. Zahiri. "Post-translational modifications in proteins: resources, tools and prediction methods". *Database* **2021**, 1-20 (2021).

[12] M. Ancona and C. A. Brackley. "Simulating the chromatin-mediated phase separation of model proteins with multiple domains". *Biophysical Journal* **121**, 2600-2612 (2022).

[13] M. Maruyama, T. Ichisaka, M. Nakagawa, *et al.* "Differential Roles for Sox15 and Sox2 in Transcriptional Control in Mouse Embryonic Stem Cells". *Journal of Biological Chemistry* **280**, 24371-24379 (2005).

[14] M. Du, S. H. Stitzinger, J. H. Spille, *et al.* "Direct observation of a condensate effect on super-enhancer controlled gene bursting". *Cell* **187**, 331-344 (2024).

[15] The ESPResSo project. (2018-2023). *Bonded interactions*. [Online]. Available: https://espressomd.github.io/doc/inter_bonded.html

[16] A. P. Thompson, H. M. Aktulga, S. J. Plimpton, *et al.* "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales". *Computer Physics Communications* **271**, 108171 (2022).

[17] Y. Demirel and V. Gerbaud "Nonequilibrium Thermodynamics - Probabilistic Approach in Thermodynamics". 4th edition. Elsevier (2019).

[18] F. C. Grandi, H. Modi, L. Kampman, *et al.* "Chromatin accessibility profiling by ATAC-seq". *Nature Protocols* **17**, 1518–1552 (2022).

[19] R. Nakato and T. Sakata. "Methods for ChIP-seq analysis: A practical workflow and advanced applications". *Methods* **187**, 44-53 (2021).

[20] The ENCODE Project Consortium "An integrated encyclopedia of DNA elements in the human genome". *Nature* **489**, 57–74 (2012).

[21] W. Humphrey, A. Dalke and K. Schulten. "VMD: Visual molecular dynamics". *Journal of Molecular Graphics* **14**, 33-38 (1996).

[22] M. Ester, H. P. Kriegel, J. Sander, *et al.* "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231 (1996).

A Appendix: DBSCAN function

```
def dbscan(atoms, threshold, cluster_types):
    """ Takes in a list of Atom objects, distance threshold + cluster types (type=4).
        Sets the cluster id for atoms we are not interested in to -2.
        Returns a list of cluster ids where the ith element is the cluster id for ith atom in input list """

    if not atoms: # edge case of empty atom list - return an empty cluster id list
        return []

    cluster_id = 0
    cluster_ids = [-1] * len(atoms)  # initialize cluster ids for each atom; -1 means unclassified

    threshold_2 = threshold**2

    def find_neighbors(atom_index):
        """ Takes in index of an atom and finds the neighbours of that atom.
            Atoms are neighbours if within cluster threshold distance of 2.3.
            Returns a list of neighbours of type=4 for atom index inputted """

        return [i for i, other_atom in enumerate(atoms)
                if i != atom_index and atoms[atom_index].type in cluster_types and other_atom.type in cluster_types and
                atoms[atom_index].sep_2(other_atom) < threshold_2]

    # loops through atoms list
    for i in range(len(atoms)):

        # checks to see if cluster id of ith atom is -1 - if not atom already processed
        if cluster_ids[i] != -1 or atoms[i].type not in cluster_types:

            # checks to see if atom type is target type - if not sets cluster id of that atom to -2 (not of interest)
            if atoms[i].type not in cluster_types:
                cluster_ids[i] = -2
```

```
32
33              continue
34
35          neighbors = find_neighbors(i) # finds neighbours of atom i
36
37          cluster_id += 1  # cluster_id acts like a 'cluster count'
38          cluster_ids[i] = cluster_id
39
40          k = 0
41
42          # for every neighbour of atom i, finds the neighbours of that
    neighbour and adds it to original neighbour list
43          while k < len(neighbors):
44              neighbor_idx = neighbors[k]
45              if cluster_ids[neighbor_idx] == -1:
46                  cluster_ids[neighbor_idx] = cluster_id
47
48                  new_neighbors = find_neighbors(neighbor_idx)
49                  for new_neighbor in new_neighbors:
50                      if cluster_ids[new_neighbor] == -1:
51                          neighbors.append(new_neighbor)
52              k += 1
53
54      no_of_clusters = cluster_id
55
56      return no_of_clusters, cluster_ids
```

B Appendix: Plots

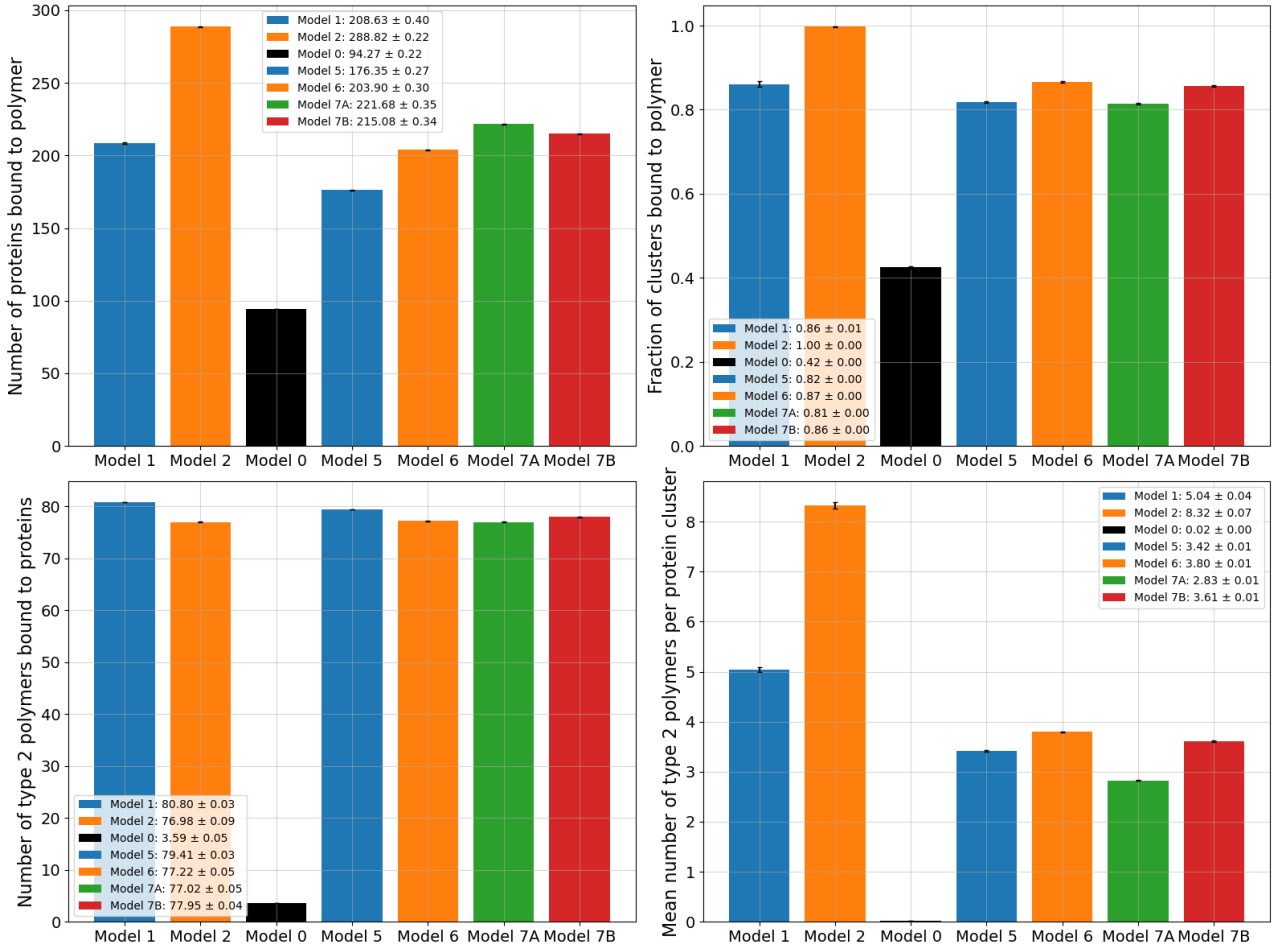


Figure 10: More steady-state subplots of the mean ± 1 SEM for various measurements on clusters forming in the system for the different models described in Section 2.