Loops are geometric catalysts for DNA integration

Cleis Battaglia 1 and Davide Michieletto 1,2,*

1School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK
2MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK
*To whom correspondence should be addressed. Email: davide.michieletto@ed.ac.uk

Abstract
The insertion of DNA elements within genomes underpins both genetic diversity and disease when unregulated. Most of DNA insertions are not random and the physical mechanisms underlying the integration site selection are poorly understood. Here, we perform Molecular Dynamics simulations to study the insertion of DNA elements, such as viral DNA or transposons, into naked DNA or chromatin substrates. More specifically, we explore the role of loops within the polymeric substrate and discover that they act as ‘geometric catalysts’ for DNA integration by reducing the energy barrier for substrate deformation. Additionally, we discover that the 1D pattern and 3D conformation of loops have a marked effect on the distribution of integration sites. Finally, we show that loops may compete with nucleosomes to attract DNA integrations. These results may be tested in vitro and they may help to understand patterns of DNA insertions with implications in genome evolution and engineering.

Graphical abstract

Introduction
Genomes are constantly reshaped and manipulated during transcription, replication and cell division. DNA sequences also need to be reshuffled during eukaryotic meiosis and undergo horizontal exchange in prokaryotes. DNA transposition—the process by which a DNA sequence is either copied or cut and pasted in a different location in the genome—is one of the main factors driving genetic diversity and expansion, especially in plants. Transposons make up about 85% of the maize genome (1,2) and up to about 50% of the human genome (3) and unregulated DNA transposition is associated with the onset of various diseases (4). At the same time, the infection and replication of many viruses, including HIV-1, require the integration of the viral DNA within the host genome (5). Given the large abundance of non-transcribed DNA within the human genome (up to 90% (6)), a random insertion process would most of the time lead to unsuccessful infections. Instead, HIV-1 and other lentiviruses, appear to integrate their DNA non-randomly and in the vicinity of transcriptionally active genes (5,7,8). Additionally, both viral and transposable element insertions often appear clustered (7), e.g. the LINE-1 and Alu repeats (8). These clusters of repeated elements often contribute to the organization and compartmentalization of the genome, thereby impacting the global genome organization (9–15). Vice versa, activation of transcription of transposable elements can cause the unfolding of chromatin and the downstream spatial rearrangement of the genome (16).

Understanding the process of DNA insertion and the biophysical mechanisms driving non-random and clustered integration patterns is an important outstanding question (17,18). While it has been argued that DNA topology and chromatin structure have an impact on the integration site selection (19–22), a systematic quantification of how integration is favoured or disfavoured in certain DNA and chromatin topologies remains an open challenge (23–26).

Recent works have highlighted the importance of spatial and physical constraints in the integration patterns of HIV-1 (5,20). For instance, genes that are located closer to the nuclear envelope are more often integrated by HIV (5).

Received: June 23, 2023. Revised: May 21, 2024. Editorial Decision: May 22, 2024. Accepted: June 11, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
same time, insertion of transposable elements may be limited by the accessibility of transposase into chromatin (18), and indeed this feature is employed in the ‘assay for transposable-accessible chromatin’ (ATAC) to map open regions of the genome (27–29). Additionally, previous simulations have suggested that different physical features may be important at different scales during the process of integration (20). For instance, the HIV-1 pre-integration complex, composed of HIV-1 viral DNA and the multimeric nucleoprotein complex, intra-some, first enters the nuclear envelope and diffuses through the nucleus within the large-scale chromatin mesh (30,31); then, it likely binds and diffuses a chromatin fibre (32) and ultimately attempts to integrate within DNA (33), which requires elastic deformation of the DNA double-helix (34). DNA integration is a multi-scale problem, involving a 3D search in a complex environment (5,30) and a 1D search within a complex free energy landscape (33). For this reason, in order to fully understand DNA integration, it is required to dissect the important contributions at each length-scale (20,21).

In this work, we focus on the contribution of DNA looping, its 3D organization and competition with nucleosomes in attracting DNA integrations. Specifically, we first quantify how the length of DNA loops affects integration site selection; we discover that the shorter the loops the more likely they are to attract integrations, as they lower the energy barrier for substrate deformation. Intriguingly, the rate of integrations within the loops is non-monotonic and displays a minimum: short loops are energetically favourable to integrate but difficult to find, while larger loops are less energetically favourable but more likely to be found by the viral DNA or transposon. We also find that the way loops are organised matters: clustered loops are systematically less integrated than sparse ones. Finally, we show that within a landscape where both loops and nucleosomes are present, loops are significantly more integrated if their length is shorter than the one of nucleosomal DNA. Overall, our findings help to gain a better understanding of the mechanisms driving DNA integration site-selection with potential implications in genome evolution and engineering.

Materials and methods

We model the insertion of a DNA element, e.g. transposons or viral DNA, into a DNA substrate, e.g. human or bacterial genome, using a coarse-grained bead-spring polymer model. Since integration is a relatively rare event that we want to study in isolation, using a coarse-grained model allows us to sample a far larger time and statistics than all-atom or oxDNA models. In our model, DNA is represented as a semi-flexible bead-spring chain polymer composed of N beads of diameter σ, which is set to be 2.5 nm (or 7.35 bp). The dynamics of each bead are determined by a Langevin equation:

\[
d \frac{d^2 r_i}{dt^2} = -\nabla U_i - \gamma_i \frac{d r_i}{dt} + \delta F_i, \tag{1}
\]

where \( r_i \) is the position of the i-th bead, \( m_i \) and \( \gamma_i \) are respectively the mass and the friction coefficient of the i-th bead due to an implicit solvent. Energies are expressed in units of \( k_B T \), where \( k_B \) is the Boltzmann constant and \( T \) is the system’s temperature. Distances are expressed in units of \( \sigma \). Time is expressed in units of the typical time for a bead to diffuse a distance of its size \( \tau_{D} = \sigma^2 / D = 3 \pi \eta \sigma^2 / k_B T \), where \( D \) is the diffusion constant for a bead, and \( \eta \) the viscosity of the implicit solvent. For simplicity, all the particles have the same mass and friction coefficient. The last term in Eq. (1), \( \delta F_i \), is a stochastic force with zero mean \( \langle \delta F_i(t) \rangle = 0 \) and amplitude \( \langle \delta F_i(t) \delta F_j(t') \rangle = 2 k_B T \gamma_i \delta_{ij} \delta(t - t') \), where \( i \) and \( j \) run over particles and \( \alpha \) and \( \beta \) run over the Cartesian components. This choice ensures detailed balance (35). The term containing the gradient \( \nabla U_i \) is the force acting on particle \( i \) due to all the other particles in the system as explained below. Consecutive polymer beads are connected through a harmonic potential:

\[
U_{\text{bend}} = k_{\text{bend}} (r_{i,i+1} - r_0)^2 \tag{2}
\]

where \( k_{\text{bend}} \) (set to 20\( k_B T / \sigma^2 \)) determines the strength of the spring and \( r_0 \) (set to 1.1\( \sigma \)) is the equilibrium bond distance. For the bending rigidity of the DNA through a Kratky–Porod potential between triplets of consecutive beads

\[
U_{\text{bend}} = k_{\text{bend}} \frac{T \ell_p}{\sigma} \left[ 1 - \frac{t_i \cdot t_{i+1}}{|t_i| |t_{i+1}|} \right] \tag{3}
\]

where \( t_i \) is the tangent vector connecting beads \( i - 1 \) to \( i \), and \( \ell_p \) is the persistence length of the polymer which is set to \( \ell_p = 20\sigma = 55 \) nm. Finally, steric hindrance interactions between non-adjacent polymer beads are simulated through an LJ potential, shifted to be zero at \( r_{\text{cut}} = 2.5 \sigma \), as

\[
U_{\text{LJ,cut}} = U_{\text{LJ}}(r_{i,j}) - U_{\text{LJ}}(r_{\text{cut}}) \tag{4}
\]

if \( r < r_{\text{cut}} \) and 0 otherwise, where

\[
U_{\text{LJ}} = 4\epsilon \left( \frac{\sigma}{r_{i,j}} \right)^{12} - \left( \frac{\sigma}{r_{i,j}} \right)^{6}. \tag{5}
\]

Below, we also consider a model where the viral DNA has a single bead with an attractive LJ interaction with the substrate polymer, mimicking the attraction of the intasome to the DNA. For this model, we use an LJ energy \( \epsilon = 4 k_B T \) and a cutoff \( r_c = 1.8 \sigma \). Unless otherwise stated, we will employ the model with purely repulsive interactions. The total potential energy acting on bead \( i \) is the sum of these three contributions. Equation (1) is integrated using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) with a velocity–Verlet scheme (36). We set the integration time step to be \( \Delta t = 0.01 \varepsilon / \gamma \). Our model does not account for the torsional rigidity of the DNA, and we plan to include it in the future.

Using the model described above we simulate two chains, one circular representing viral DNA or a transposable element, and the other linear representing the target, or substrate DNA. Both polymers diffuse in a box and we attempt the integration of the circular polymer into the linear chain using a modified version of the 'double-bridging' algorithm implemented in LAMMPS as fix bond/swap (37). This code allows us to perform 'reconnection' moves and swap the bonds in the system that are connecting beads that are closer than \( r_c = 2 \sigma \) in 3D (see Figure 1). The reconnections are attempted every 1 LAMMPS step. We implemented some modifications to this fix that allow us to perform recombination moves between viral and host polymers (inter-chain reconnections) while avoiding intra-chain (or self) reconnections (20,38,39).

The reconnection moves are also weighted by a Metropolis test that assigns a probability of successful swap depending on the energy difference \( \Delta U \) between the old and the new configuration before and after the swap. This probability is 1 if \( \Delta U < 0 \) and \( p = e^{-\Delta U / k_B T} \) if \( \Delta U \geq 0 \). We also
modified this part of the code to perform polymer reconnections that bypass the Metropolis test thus allowing non-equilibrium integration. These modified codes can be found at https://git.ecdf.ed.ac.uk/taplab.

The simulations are performed as follows. We prepare an initial polymer configuration as a random walk and impose the loop by setting a harmonic bond (with strength increasing from 1 to 20 \( k_B T \) during the equilibration) in between two beads at 1D distance \( l \). For each of the >1000 replicas, we let the system equilibrate for at least \( 3 \times 10^5 \) LAMMPS steps. At the end of the equilibration, we allow the integration to happen. We stop the simulation after 10^7 timesteps, regardless of whether the integration has happened or not. We output the bond list every 10^6 steps and from that, we reconstruct the topology of the polymers and can detect when and where the integration has happened. Using this information we then build a histogram of where the integration sites are along the polymer and can thus quantify how many are within and outside the looped regions.

Results
Integration in DNA loops
DNA loops are thought to be abundant and biologically important in both eukaryotic and prokaryotic genomes (40–44). Beyond classic enhancer-promoter interactions (45), several new mechanisms of loop formation have been identified recently, for instance those mediated by CTCF (43) (∼0.1–1 Mb), cohesin (46,47) and condensin (48–50) (∼1–10 kb). In bacteria, structural-maintenance-of-chromosome (SMC) proteins (51,52) and nucleoid associated proteins (53) also play a major role in forming DNA loops in vivo (54) (30–420 kb). Additionally, integrase can itself deform the substrate in such a way as to create an extremely small (∼5–10 bp) DNA loop within the nucleosome (55). Finally, investigation of DNA integration on looped substrates is also possible in vitro (33,40,56).

In analogy with previous work, here we argue that short DNA loops may attract more integrations due to the fact that they have a lower energy barrier for elastic deformations (20). On the other hand, short DNA loops are energetically costly (46,57,58). In order to quantify how frequently short DNA loops may appear in the human genome, we compute the 1D distance between nearest-neighbouring active regulatory elements, such as enhancers and promoters, as listed in the GenHancher database (59). Figure 2A shows that about 8% of these regulatory elements are closer than 1 kb apart and 3% of them closer than 500 bp. While only a fraction of them will be looped at any one time, these figures show that there is a considerable amount of short (<1 kb) potential loops in the human genome. On top of this, we note that the integrase itself can compete with nucleosomes to create an extremely short loop of DNA in the range of 5–10 bp (55). Likewise, we expect a similar, if not greater proportion of short loops between regulatory elements in bacteria, such as the lac repressor (60) or deformations of the DNA substrate for instance created by nucleoid-associated proteins such as IHF (53,61).

Motivated by this, we aim to characterize how the 1D loop length affects the insertion of the circular DNA into the target substrate DNA. To do this, we perform simulations in which a loop of length \( \ell \) is stabilized in a polymer of length \( N = 100 \) beads ≥735 bp introducing a harmonic spring in between two beads. We then perform 1500 independent simulations to stochastically sample the probability distribution, \( p(x) \), of integrating at position \( x \) along the polymer. In Figure 2D we show the profile of integrations, i.e. \( p(x) \), over the simulated polymer. One can appreciate that the case in which there is a short loop (80 bp) in the polymer yields a distribution \( p(x) \) peaked within the loop, whereas the case with a longer loop (448 bp) yields a constant distribution compatible with the random \( p_{rand} = 1/N \) distribution. The presence of weak (\( \epsilon = 4k_B T \)) unspecific attractive interaction between the viral DNA and the substrate yields two peaks corresponding to the base of the substrate loop (Figure 2D). We argue that this is due to the locally higher density of target beads, thus maximising the attractive interaction between substrate and intasome.

To better quantify the enhancement of integrations within the looped region compared with the ones outside it, we compute the probability of integration per bead inside the loop as \( p_{in} = l_{in}/(l_{tot}) \), where \( l_{in} \) is the number of integrations that occur inside the loop and \( l_{tot} \) the total number of integrations across our simulations. If the insertions were random we would expect \( l_{in}/l_{tot} = \ell/N \), recovering a random integration probability per bead \( p_{rand} = 1/N \) (see Figure 2E). Importantly, as seen in Figure 2E, the presence of a weak attractive interaction does not significantly affect the integration probability as a function of loop length.

By analysing our simulations, we find that \( p_{in} \) is significantly larger than the random probability when the polymer has a short loop (<200 bp) and that \( p_{in} \) decays to 1/N for large loop lengths. This is also in line with the experimental finding of favoured integration in DNA molecules carrying torsional stress (17). We understand this as follows: Short loops store a significant amount of energy in the form of bending. This implies that when the viral loop integrates into the pre-looped substrate, it can immediately lower the (free) energy of the system because the loop is then longer and the local curvature smaller. Indeed, the elastic energy paid for the looping of an elastic rod of length \( \ell \) and persistence length \( l_p \) is

\[
\frac{F}{k_B T} = \frac{\epsilon l_p}{\ell}
\]

with \( \epsilon \) a numerical factor depending on the shape of the loop (62). Assuming the integration process to be an Arrhenius, energy activated, process we model it through a Metropolis Monte Carlo scheme, and thus accept each integration attempt with a probability that is given by \( p_{MC} = \exp\left( -\Delta U/k_B T \right) \) when \( \Delta U \geq 0 \) and 1 otherwise, where \( \Delta U = U(\text{after}) - U(\text{before}) \). Importantly, \( U(\text{after}) \) and \( U(\text{before}) \) are the total energies of the system just after and just before the bond swap that leads to integration (see Methods). In
Figure 2. (A) Sketch of our search procedure for nearest 1D regulatory elements in the genome. (B) Histogram of 1D distances between nearest-neighbor pairs of regulatory elements, i.e. enhancer and promoters, within the human genome: 3% of these are below 500 bp and 8% below 1 kb. (C) Sketch of the simulation setup, where we investigate the integration probability in loops of different sizes. (D) Histogram of integration sites probability along the polymer for loops of different sizes (80 bp in green and 448 bp in grey). In pink, we show the histogram for a 448 bp loop where the viral DNA is performing facilitated diffusion (fd). (E) Probability of integration inside the looped region as a function of loop length, comparing equilibrium (green) and non-equilibrium (purple) algorithms. The red curve shows the case in equilibrium where the viral DNA undergoes facilitated diffusion (fd). The black solid line indicates what we would expect for random integrations.

Figure 3. (A) Total angle energy measured during the course of two simulations, one with an integration event within a loop of length 51 bp (blue trace) and one with an integration in the non-looped region (green trace). (B) Plot of $-\Delta U_{\text{angle}}$ showing the difference between one timestep before and one timestep after an integration event within a loop as a function of the loop length. The solid line is a fit with $\Delta U_{\text{angle}} = a/\ell + b/\ell - b/\ell$. Note that short loops display a positive change in $-\Delta U_{\text{angle}}$, i.e. a decrease in free energy, whilst large loops display a negative change, i.e. an increase in free energy, at the instant of integration. (C) Scaling of the fraction of time the viral DNA spends in the three-dimensional proximity of the looped section as a function of loop length. The solid line is a guide for the eye for a linear scaling. (D) Fraction of integrations within the looped region. The solid line is a fit with Eq. (8) with $a$ and $b$ as free parameters; the values obtained are $a = 594$ bp and $b = 356$ bp.

Figure 3A, we show that an integration event within a loop (indicated with an arrow) lowers the angular energy of the system (blue trace). On the contrary, for integrations in non-looped substrates, the angular energy difference is within thermal fluctuations (green trace). We note that we expect components of the total energy, for instance the bond energy, to increase during the integration step, as the bonds are necessarily stretched, yet this contribution does not depend on the loop length. We can estimate the angular contribution of the total energy difference before/after integration within a loop as $\Delta U_{\text{angle}} = (U_{\text{angle}}(\text{after}) - U_{\text{angle}}(\text{before}))/k_B T \approx a/\ell + l_{\text{HIV}}/b/\ell$ where $a$ and $b$ are numerical constants with units of length whose values are dependent on the shape of the loop before and after integration. We test this approximation by measuring the energy change during an integration event and we plot the difference $\Delta U$ in Figure 3B which indeed displays the expected behaviour. In turn, the probability of integration within a loop can be written as

$$P(\text{found} | \text{in})P(\text{in}) = P(\text{in} | \text{found})P(\text{found})$$

and where we can take $P(\text{found} | \text{in}) = 1$. In this equation, $P(\text{found})$ is the probability of a viral DNA finding a loop, and $P(\text{in})$ is the probability of integration in the loop. In turn, we can write that the probability that there is an integration within the loop, conditional to the fact that the viral DNA has found the loop is $P(\text{int} | \text{found}) = \exp(-\Delta U/k_B T)$, the same as $\exp(-\Delta U_{\text{angle}}/k_B T) \approx \exp(-a/\ell + l_{\text{HIV}}/b/\ell)$, where we neglected loop-length independent contributions. In other words, the probability that an integration attempt is accepted, on a configuration where the viral DNA is in contact with the substrate loop, is equal to the Boltzmann factor. At the same time, the probability that the viral DNA finds the looped segment within the simulation box is $P(\text{found}) = (1/6)\pi \sigma \ell V/\ell$, where $V$ is the volume of the box. In simulations, we can test this scaling by tracking the fraction of time that the viral loop spends in 3D proximity of the DNA loop as a function of different loop sizes; as shown in Figure 3C, this quantity is to a good extent linearly increasing with the size of the loop. The same argument holds for integrations outside of the loop, where $P(\text{outside}) = \exp(-\Delta U_{\text{angle}}/k_B T) \sim O(1)$, as for non-
looped regions (or very large loops) we expect a constant that is not dependent on the loop length (see Figure 3A). Additionally, the probability of finding a non-looped region is \( P(\text{found}) \sim (N - \ell) \). Thus, we can write that the fraction of viral loops integrated within the looped substrate is then expected to be

\[
 f_{in} = \frac{P(\text{in})}{P(\text{in}) + P(\text{out})} = \frac{\ell e^{-a/\ell} + b/\ell}{\ell e^{-a/\ell} + b/\ell + (N - \ell)} .
\]

By plotting the fraction of integrations inside the looped region \( f_{in} = p_{in} \ell \) we see that the data follows extremely well the prediction of Eq. (8) (see Figure 3D). Intriguingly, our findings imply that there is an optimum length to reduce the amount of integrations within the looped segment. As one can appreciate from Figure 3D, the minimum corresponds to a loop that is long enough to possess a small free energy gain upon integration, yet short enough to be difficult to find. We also note that the minimum corresponds to a size comparable to the one of a nucleosomal DNA (including the linker DNA).

Integration in clustered and sparse loops

We now ask what happens to the distribution of integrations when there are many loops along the substrate. In particular, we are interested in understanding what are the cooperative effects that appear when the loops are sparse and uniformly distributed along the substrate or when they are clustered in a short segment of the polymer (Figure 4A). To investigate this question, we, therefore, perform simulations with many loops (all of the same length \( \ell = 80 \) bp) formed along a polymer of 10.3 kb in either sparse or clustered 1D arrangement (Figure 4A).

Interestingly, we observe that the 3D conformations of the polymers are rather distinct, with the latter (clustered) case being more swollen and with the looped region occupying a smaller volume fraction than in the former (sparse) case (see snapshots in Figure 4A and radius of gyration in Figure 4C). We argue that the radius of gyration decreases for a sparse distribution of loops because the loops act as kinks, effectively increasing the DNA’s flexibility. On the other hand, clustered loops do not affect the overall flexibility of the chain.

To understand if this distinct 3D organization due to the 1D arrangement affects the overall integration probability, we perform simulations with varying numbers of loops while preserving constant loop size (Figure 4D). In line with what we discovered in the previous section, we observe that the sparse loops attract more integrations overall. Again, we argue that this is due to the smaller 3D volume fraction occupied by the 1D clustered loop arrangement.

The consequence of the different 3D organizations is reflected also in the distribution of integration events. As shown in Figure 4E and F, we observe that in the sparse case the integrations are uniformly distributed among all the loops in the polymer, however, in the case of clustered loops, the ones at the two ends of the region hosting the loops display more integrations than the inner ones (Figure 4C). We argue that this is due to a ‘screening’ effect, whereby the marginal loops are typically the ones more exposed while the more internal ones are more tucked in within the clustered looped region and therefore less accessible to integration because of excluded volume and steric hindrance of the outer loops.

Integration in looped euchromatin

We conclude this work by investigating the integration within a substrate that displays both loops and nucleosomes. Nucleosomes are assembled as in (29) where we have previously shown that DNA wrapped around nucleosomes attract more integrations due to their strongly bent states, in line with experiments (17,23,32,55). As we previously showed in (20) that open chromatin fibres, compared to more compacted ones, displayed the highest enhancement of nucleosomal DNA integration, we particularly focus on a euchromatic substrate. We model euchromatin by assigning a short-ranged attractive interaction (\( \epsilon = 4k_BT \)) between individual nucleosome core particles (NCPs) (represented by spheres of size \( a_b = 3\sigma = 7.5 \) nm) and selected segments of the polymer representing the nucleosomal DNA. The nucleosome–nucleosome and nucleosome–DNA interactions are otherwise of steric hindrance. We let the DNA and NCPs diffuse within the simulation box, allowing the self-assembly of the chromatinized DNA as seen in Figure 5A,C). Each block of nucleosomal DNA is 20 beads or \( \sim 147 \) bp and is separated from the adjacent NCP by linker DNA (10 beads or \( \sim 74 \) bp) segments. Notice that this model has been employed in existing literature to describe in vitro chromatin reconstitution (63). Here we ask what happens if in a segment of chromatin there is a competition, or a synergy, between looped and nucleosomal DNA, and whether the two would enhance even more the integration in looped euchromatin.

To test this, we perform simulations of a polymer of length 10.3 kb wrapped around \( \sim 30 \) NCPs (where each block of nucleosomal DNA is 20 beads or \( \sim 147 \) bp) and displaying a central looped region with varying loop lengths. Specifically, we compare two systems: one with loops of length \( \ell = 80 \) bp (11 beads), and one with loops of length \( \ell = 154 \) bp (21 beads) (see snapshots in Figure 5A, C). We observe that shorter loops enhance integration within the looped region rather than into the nucleosomal DNA. Instead, having loops of approximately the same size as the nucleosomal DNA blocks leads to a more uniformly distributed integration profile, preserving only the preference of ‘bent’ DNA regions over the rest of the polymer (Figure 5B, D). We note that the spikes in the integration profile correspond to the location of a nucleosome or a loop. In line with what we discovered in the previous section, we again observe in the profile of integration probability, that the interior loops and nucleosomes are screened by the more external ones (i.e. \( P(s) \) is systematically smaller in the middle with respect to the edge of the polymer).

Finally, we highlight that recent cryo-EM structural data revealed that the intasome–nucleosome structure creates a short, highly bent DNA loop by shifting the nucleosome out of registry (55). Such a short (5–10 bp) loop would, in our model, favour even more the integration within this extremely bent loop-on-nucleosome feature.

Discussion

How geometric features of DNA and chromatin may affect the integration site selection of retroviral DNA or transposable elements is not fully understood. While it is widely accepted that the integration site selection is not random, the underpinning physical, geometric, and topological elements driving the selection are still largely unexplored.
In this work, we used computer simulations to explore the role of loops in DNA and chromatin substrates in determining DNA integration site-selection. We use a model of reconnecting polymers in equilibrium, hence iso-energetic and satisfying detailed balance, that has already been successful at capturing the distribution of HIV integration site-selection in the human genome (20). First, we discover that integrations are favoured in looped regions due to the release of bending energy (Figure 3). This is in line with experimental observations showing that integration is favoured in nucleosomes (23) and supercoiled substrates (17,33). We argue that, in spite of the fact that the microscopic action of strand exchange is iso-energetic, the release of bending energy acts as a geometric catalyst for the integration site selection by lowering the activation barrier of substrate deformation.

Intriguingly, we observed a trade-off between the free energy gained and the probability of finding such loops. More specifically, integration in small loops releases a larger amount of free energy compared to larger ones, but they are more difficult to find than larger loops. In our simulations, this competition manifests itself with a non-monotonic behaviour of the integration rate displaying a minimum at around 200 bp (Figure 3).

We have also explored the effect of having many loops on the same substrate. Interestingly, the 3D conformation of the polymer strongly depends on whether the loops are sparse, i.e. uniformly distributed along the chain, or are clustered in 1D. In the latter case, the looped region occupies a smaller volume fraction and is self-screened and, as a consequence, overall less integrated than a similar substrate with sparse loops. This many-loop system also displays a screening effect whereby 1D clustered loops display a non-uniform integration probability, that is largest for the loops that are most external to the 1D cluster (Figure 4).

Finally, we perform simulations of a chromatin substrate and discover that loops attract more integrations, as long as they are shorter than the nucleosomal DNA segments (Figure 5). Thus, our simulations suggest that loops-on-nucleosomes, such as the ones generated by the intasome-
nucleosome complex (55), are perhaps the most effective at attracting integrations.

We note that our model neglects electrostatic interactions, the effect of histone tails or other chaperone proteins (such as LEDGF/p57). Whilst more complete models of chromatin and DNA may be considered in the future (63–66), in this work we explicitly focused on the elastic contribution of a chromatin polymer model on the process of integration. We argue that our results shed light on the role of loops as geometric catalysts in determining the integration site selection in vitro and in vivo and could be tested by designing artificial loops either in vitro using, for instance, bivalent dCas9 (67) or in vitro by using DNA origami (68) or generic bridge proteins (46).

Data availability
Source codes are available at https://git.ecdf.ed.ac.uk/tablap.

Acknowledgements
D.M. is a Royal Society University Research Fellow and is supported by the ERC (Starting Grant, Topologically Active Polymers, Ref. 947918). The authors thank Yair Posada for helpful discussions throughout the project. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission. Source codes are available at https://git.ecdf.ed.ac.uk/tablap.

Funding
European Research Council [947918]. Funding for open access charge: ERC [Starting Grant, Topologically Active Polymers, Ref. 947918].

Conflict of interest statement
None declared.

References
kilobase resolution reveals principles of chromatin looping. **Cell**, 159, 1665–1680.


