This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

An asymmetric mesoscopic model for single bulges in RNA

Erik de Oliveira Martins[1, a)] and Gerald Weber[1, b)]

Departamento de Física, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte-MG, Brazil

(Dated: 4 September 2017)

Simple one-dimensional DNA or RNA mesoscopic models are of interest for their computational efficiency while retaining the key elements of the molecular interactions. However, they only deal with perfectly formed DNA or RNA double helices and consider the intra-strand interactions to be the same on both strands. This makes it difficult to describe highly asymmetric structures such as bulges and loops, and for instance prevents the application of mesoscopic models to determine RNA secondary structures. Here we derived the conditions for the Peyrard-Bishop mesoscopic model to overcome these limitations and applied it to the calculation of single bulges, the smallest and simplest of these asymmetric structures. We found that these theoretical conditions can indeed be applied to any situation where stacking asymmetric needs to be considered. The full set of parameters for group I RNA bulges was determined from experimental melting temperatures using an optimization procedure and we also calculated average opening profiles for several RNA sequences. We found that guanosine bulges show the strongest perturbation on their neighboring base pairs, considerably reducing the on-site interactions of their neighboring base pairs.

PACS numbers: 87.14.gk, 87.15.A-, 87.15.B-

Keywords: Peyrard-Bishop model, RNA bulges, hydrogen bonds, stacking interaction

[a)]Electronic mail: erikfisica@gmail.com

[b)]Electronic mail: gweberbh@gmail.com

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

I. INTRODUCTION

Single RNA bulges are unpaired bases that are responsible for important perturbations to the double helix. They occur naturally and are frequently seen in ribossomal RNA[1] but may equally be synthesized by matching strands of unequal length such that one or more bases are left unpaired. Bulges are thought to be responsible for recognition sites by acting as molecular handles.[2] Given the importance of bulges, it would be of interest to use mesoscopic models, such as those belonging to the class of Peyrard-Bishop (PB) models,[3] to describe their thermodynamic properties. For instance, this would enable us to apply this type of model to bioinformatics applications such as secondary structure predictions. Mesoscopic models have an advantage over more complicated approaches such as coarse-grain models[4] which are computationally too expensive to be extensively used for large scale bioinformatics applications. The simpler PB-type models would be efficient enough for this type of application and are of interest as they describe the molecular interactions which nearest-neighbor (NN) models for instance do not.[5] However, PB models are still largely unable to deal with bulges and loops. Here, we will derive the conditions that need to be met by the model Hamiltonian to overcome some of these limitations. We then apply these conditions to the simplest and smallest perturbation which are single bulges flanked by Watson-Crick base pairs in RNA.

RNA bulges are usually classified in terms of the neighboring bases and location in the helix.[6] Group I and II bulges are flanked by CG or AU base pairs while group III and IV bulges may have at least one GU base pair as neighbors. Group I and III bulges have no repeated neighboring bases and therefore their location is well defined, see Fig. S1. In contrast, group II and IV have repeated neighbors and the exact location of the bulge in the sequence is ambiguous.

From a structural point of view, bulges are found stacked into the helix (intercalated), away from the helix (flipped out) or bent towards one of the grooves (side-by-side).[2,7–9] This perturbation to the double helix causes a decrease of the denaturation temperatures.[10] In particular, White and Draper[11,12] studied the intercalation effects of single bulges in RNA and showed that the inclusion of bulges has a dramatic effect on the conformation possibilities of the helix, confirming similar observations for DNA bulges.[13] Adenine bulges flanked by mismatched GA base pairs were crystallized in HIV-1 RNA by Ennifar *et al.*[14] and

studied with x-ray diffraction revealing a significant asymmetry in the deep grove. In some cases bulges were found to assume two different conformations, for instance single cytosine bulges were found to form two different types of extra-helical triples.[15] For uridine bulges, it was found from x-ray diffraction that they flip out and protrude into the minor groove.[16] Similarly to RNA, DNA bulges were found to loop out[17] and destabilize the helix.[18–20] Popenda, Adamiak, and Gdaniec[9] investigated structural differences between a regular RNA duplex and one containing an adenosine Group II bulge. They found that the distortion caused by the bulge propagates through the whole structure.

Barthel and Zacharias[21] performed a molecular dynamics simulation for a specific sequence containing a single adenosine or uridine bulge and found that flipping out of the major groove was less favorable than flipping out of the minor groove. They also found that the extrahelical bulges were stabilized by electrostatic interactions while stacked bulges had van der Waals and nonpolar contributions. Molecular dynamics was also used to study single adenosine bulges under pressure.[22] Flexibility and bending were simulated with coarse-grained models.[23]

The thermodynamic stability of single RNA bulges were analyzed using gradient gel electrophoresis[6] and melting temperature experiments.[1,24–26] One important conclusion emerging from these studies is that the NN model is not well suited to handle the structural perturbation to the helix.[27] This follows similar conclusions for DNA bulges.[20,28] Melting temperatures for single bulges in RNA hairpins[29–31] as well as for longer bulges[32] further highlight these shortcommings. Therefore, there is a motivation to go beyond the nearest-neighbor model but which are still computationally efficient to handle a large number of sequences. Mesoscopic approaches, such as the Peyrard-Bishop (PB) model[3], are an interesting choice as they allow to describe the stacking interaction separately from the hydrogen bonds.

The PB model is an important statistical model used to study mechanical and thermodynamic properties of DNA and RNA molecules.[3,33] An advantage of the PB model is its intuitive way to describe interactions of hydrogen bonds and stacking interactions in double strands by a Hamiltonian. The PB model is under active development and is applied to a number of different physical situations involving oligonucleotides. Some recent examples for DNA are its use to study bubble length distribution,[34] charge transport[35,36] over-stretching transition,[37] and flexibility in circular DNA.[38] One interesting property of the model is that

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

it is easily adaptable to describe further interactions such as the influence of solvent.[39,40] Another crucial aspect of the PB model is that it can be parameterized from melting temperatures and the resulting parameters can be easily interpreted in terms of stacking interactions and hydrogen bonds.[5,41] This parametrization is also remarkably sensitive, for instance we were able to model the ends of a DNA sequence independently showing an important reduction of hydrogen bond strength for low salt concentration.[42] Note that parametrization can also be achieved via molecular dynamics.[43] For our purpose, the PB model has an important advantage over other PB-derived models, such as the Dauxois-Peyrard-Bishop (DPB) anharmonic model[44] or the harmonic model with added solvent potential,[39] as it requires fewer parameters making the optimization much easier. This is particularly important when there is a reduced number of melting temperatures.

However, to apply the PB model to bulges we need to overcome some important limitations of the model. Perhaps the most important approximation is that the PB model considers the stacking interaction to be symmetrical, that is, it considers that the stacking is the same for both strands. For bulges this is certainly not true, whether they are intercalated or flipped out, the stacking on the bulge site will be very different for each strand. Here we modify the PB to account for asymmetric stacking with some very surprising results which are of importance for normal double helices as well. The second difficulty posed by bulges is that one strand is at least one nucleotide shorter than the other while the PB model considers only double strands of equal length. This limitation is easier to overcome by simply adding a pseudo-base pairing with the bulge with no hydrogen bond in a similar approach as used for mismatches.[45]

Having addressed the stacking asymmetry of the PB model, we were able to follow a parametrization procedure[5] for single group I RNA bulges using the experimental melting temperature from Refs. 1,24–26. We have not included type II bulges in our analysis as there is an uncertainty of the location of the bulge position. Bulges neighboring GU (groups III and IV) were also presently not addressed as GU adopts a highly variable conformation, even when surrounded by canonical base pairs.[33]

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

II. MATERIALS AND METHODS

A. The asymmetric stacking of the Peyrard-Bishop model

In the Peyrard-Bishop (PB) model, the Hamiltonian is written in terms of two coordinates u_n and v_n, one for each strand, along the same direction perpendicular to the helix longitudinal axis,[3]

$$H = \sum_n \frac{1}{2}m(\dot{u_n}^2 + \dot{v_n}^2) + \frac{1}{2}k(u_n - u_{n-1})^2 + \frac{1}{2}k(v_n - v_{n-1})^2 + V(u_n - v_n), \tag{1}$$

where k is the stacking constant, V is a potential representing the hydrogen bonds and m is mass of the bases. It is implied that k is in fact $k_{n-1,n}$ and V is V_n, but we will omit these subscripts to ease the notation as they will not change the results. The effective potential representing the on-site interaction is given by a Morse potential[46]

$$V(u_n - v_n) = D(e^{-\frac{1}{\lambda\sqrt{2}}(u_n - v_n)} - 1)^2, \tag{2}$$

where D is the main parameter representing the strength of the potential and λ controls the potential width.

Crucially, Eq. 1 considers the stacking interaction as symmetric, using a uniform harmonic coupling k. A uniform stacking prevents us from applying the model to situations where there is a important strand asymmetry. Here, we introduce an asymmetric stacking k_u and k_v, corresponding to the displacements u_n and v_n. This approach still allows us to use the same change of variables as for the strand-symmetric model[3]

$$x_n = \frac{1}{\sqrt{2}}(u_n + v_n), \quad y_n = \frac{1}{\sqrt{2}}(u_n - v_n), \tag{3}$$

such that

$$u_n = \frac{1}{\sqrt{2}}(x_n + y_n), \quad v_n = \frac{1}{\sqrt{2}}(x_n - y_n), \tag{4}$$

and the factorization in regard to the momentum also still applies,

$$Z = (2\pi m k_B T)^N Z_{xy}. \tag{5}$$

However, unlike the symmetric PB model, we cannot immediately decouple the x and y terms of the partition function, instead we end up with cross terms,

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

$$Z_{xy} = \prod_{n=1}^{N} \int dx_n dy_n \exp\Bigg[-\beta \Bigg\{ \frac{1}{2} k_a \left[(x_n - x_{n-1})^2 + (y_n - y_{n-1})^2 \right] + k_s (x_n - x_{n-1})(y_n - y_{n-1}) + V(y_n) \Bigg\} \Bigg], \tag{6}$$

where

$$k_a = \frac{1}{2}(k_u + k_v), \quad k_s = \frac{1}{2}(k_u - k_v). \tag{7}$$

Nevertheless, it is still possible to factorize the partition function with some algebraic effort. We rewrite the function partition Z_{xy} in the form

$$Z_{xy} = \prod_{n=1}^{N} \int dy_n \exp\left\{ -\beta \left[\frac{1}{2}\gamma (y_n - y_{n-1})^2 + V(y_n) \right] \right\} \times \int dx_n \exp\left\{ -\beta \left[\sqrt{\frac{k_a}{2}} (x_n - x_{n-1}) + \frac{k_s}{\sqrt{2k_a}} (y_n - y_{n-1}) \right]^2 \right\}, \tag{8}$$

where we used the definition

$$\gamma = \frac{k_a^2 - k_s^2}{k_a}. \tag{9}$$

We introduce another change of variables

$$r_n = \sqrt{\frac{k_a}{2}} (x_n - x_{n-1}) + \frac{k_s}{\sqrt{2k_a}} (y_n - y_{n-1}) \tag{10}$$

with

$$dr_n = \sqrt{\frac{k_a}{2}} dx_n, \tag{11}$$

which reduces the second groups of integrals in Eq. (8) to

$$\left(\frac{2}{k_a} \right)^{\frac{N}{2}} \prod_{n=1}^{N} \int dr_n e^{-\beta r_n^2} = \left(\frac{2\pi}{\beta k_a} \right)^{\frac{N}{2}}. \tag{12}$$

The partition function Eq. (8) can now be factorized

$$Z_x = \left(\frac{2\pi}{\beta k_a} \right)^{\frac{N}{2}}, \tag{13}$$

and

$$Z_y = \prod_{n=1}^{N} \int dy_n \exp\left\{ -\beta \left[\frac{1}{2}\gamma (y_n - y_{n-1})^2 + V(y_n) \right] \right\}. \tag{14}$$

The important result here is that the factorized partition functions Z_x and Z_y are functionally identical to the partition function of the symmetric model.[3] The only difference is that the

original symmetric k is now replaced by an effective stacking constant γ in Z_y

$$\frac{1}{\gamma} = \frac{1}{2}\left(\frac{1}{k_u} + \frac{1}{k_v}\right). \tag{15}$$

This result is valid for any interstrand potential V, but only for harmonic stacking potentials. It cannot be applied to anharmonic stacking as proposed by Dauxois, Peyrard, and Bishop [44] where this type of decoupling is not possible. In this case the Z_{xy} partition function would have to be solved numerically as a two-dimensional Fredholm equation which is out of the scope of this work. One should be aware that the anharmonic or DPB model displays a sharp increase in the average opening mainly due to a numerical artifact,[39,47] and parameter optimizations for this model have not produced better results than the harmonic PB model.[5]

B. Bulge implementation and notation

The PB model cannot account for strands of different sizes but we bypass this difficulty by introducing a pseudo-base at the opposite position of the bulge as shown schematically in Fig. 1. We use the character X, representing the pseudo-base, placed on the shorter opposite strand visually filling-in the vacancy, as in the following example of an adenine bulge

$$\begin{array}{l} 5'\text{-CCAUUACUACC-}3' \\ 3'\text{-GGUAAGAUGG-}5' \end{array} \rightarrow \begin{array}{l} 5'\text{-CCAUUA\underline{C}UACC-}3' \\ 3'\text{-GGUAA\underline{X}GAUGG-}5' \end{array}$$

We will refer to the pseudo-base-pair formed by bulges using the generic notation BX with B=A, C, G or U. In our example that would be a AX bulge-pseudo-base.

The Morse potential of the bulge-pseudo-base will be set to zero as there is no interstrand interaction as shown schematically in Fig. 1. However, we cannot do the same for the stacking interaction k between the pseudo-base X and is neighbors. Effectively, as we are inserting the pseudo-base X between two bases, the stacking interaction will be split in two parts. To understand this, lets use the strand notation of the Hamiltonian Eq. (1) and assume for instance that the bulge is located at the nth position on the u-strand and the X base at the nth site on the v-strand. The stacking interaction of the neighboring bases at $n-1$ and $n+1$, $k_{v;n-1,n}$ and $k_{v;n,n+1}$ will represent what would otherwise be $k_{v;n-1,n+1}$ bridging the bulge, see Fig. 1.

In our optimization scheme, described in section II D, we will allow for the canonical base pairs BP to assume different parameters when they are neighbors of a bulge B. To represent

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

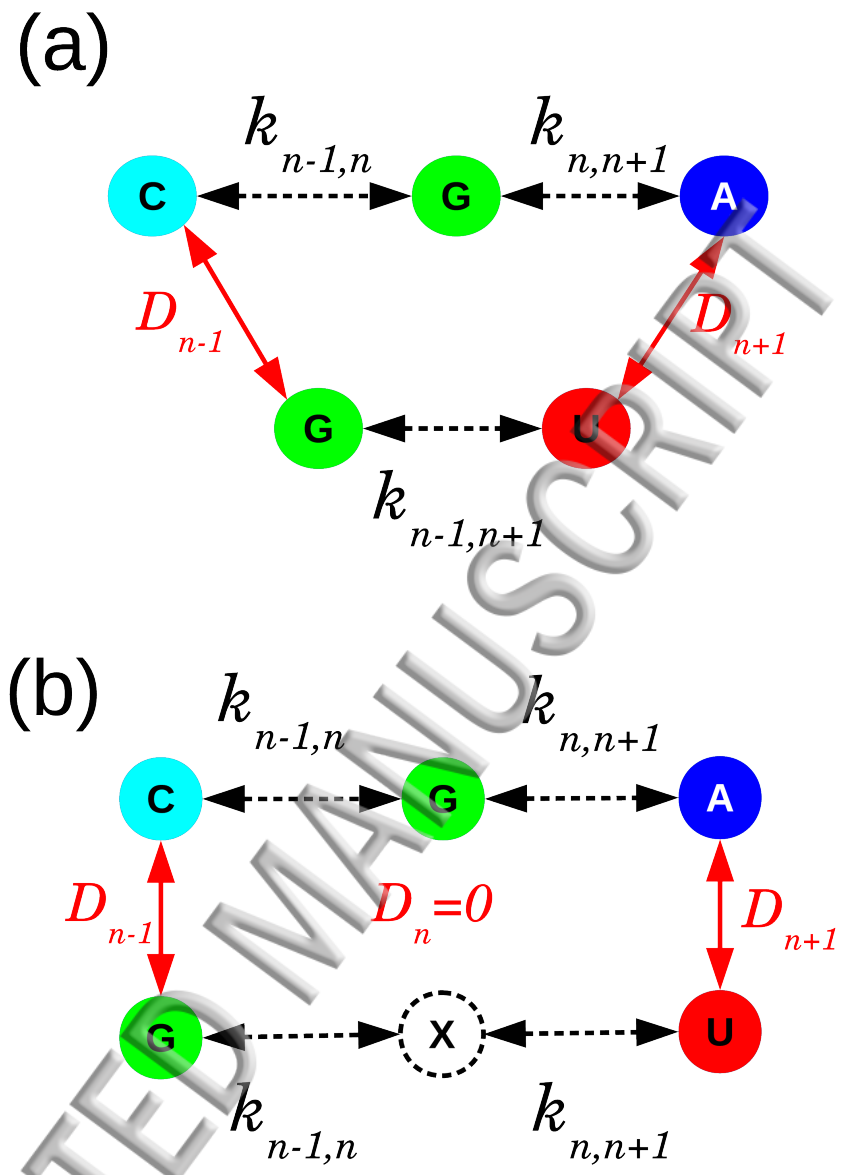


FIG. 1. Bulge implementation. Panel (a) shows an example of a G bulge flanked by CG and AU base pairs. Panel (b) shows the introduction of a pseudo-base X with a zero Morse Potential opposite to the G base.

this unambiguously we write this as BP^B. For instance a CG base pair next to a AX bulge will be written as CG^A, similarly as adopted in Ref. 33. Therefore, a CG^A is a base pair with variable parameters next to a bulge A, while CG without the superscript has constant parameters and is not flanking a bulge. Note that we do not distinguish on which side of the CG^A is located, doing so would require an additional index and would double the number of parameters to be minimized.

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

A consequence of this notation is that we will deal with three types of nearest-neighbours.

a. Fixed nearest-neighbors (FNN) are formed by two AU or CG base pairs away from any bulge. These will be indicated in the form BPpBP. For instance AUpCG is a AU followed by and CG base pair, neither being flanked by a bulge. FNNs will not change their parameters during the minimization procedure.

b. Variable nearest-neighbors (VNN) are formed by two AU or CG base pairs, one of which is flanking a bulge. This is written either as $BP^{B}pBP$ or as $BPpBP^{B}$ depending on which base pair BP is next to a bulge B. VNNs will vary their parameters during the minimization.

c. Bulge nearest-neighbors (BNN) are formed by a bulge-pseudo-base of type BX and either a AU or CG base pair. This will be written generically as $BP^{B}pBX$ or $BXpBP^{B}$. In our example, at the beginning of this section, that would be a AX bulge with $UA^{A}pAX$ nearest-neighbor at the left-hand side and $AXpCG^{A}$ at the right-hand side.

Our modeling strategy is to allow the bulge-flanking base-pairs, of type BP^{B}, to vary their Morse potential D, while the remaining CG and AU base pairs will keep their Morse potential constant. For example, CG^{A} indicates that this particular CG base pair will have its hydrogen bonds altered due to the presence of an adenine bulge. We will also vary the the stacking interactions of VNNs as well as of BNNs, only FNNs will remain fixed.

C. Melting temperature data set

We collected a total of 80 sequences with group I single bulges from Refs. 1,24–26, of which 18 are adenine (AX), 12 cytosine (CX), 24 guanosine (GX) and 26 uridine bulges (UX). Note that Kent *et al.*[26] initially described their bulges as being of group III, however their free energy analysis confirmed that they are in fact group I bulges.

The melting temperatures were recalculated to a strand concentration of 200 μM from their respective total enthalpy and entropy variations, ΔH and ΔS.[1,24–26] This is done for consistency with our previous calculations for CG and AU RNA base pairs.[41] The complete list of sequences and their adjusted melting temperatures are shown in supplementary table S1–S4.

D. Optimization of the model parameters

We used the technique of thermal equivalence to optimize the parameters of the potentials presented.[48] The bulge parameters were optimized by minimizing the square differences between experimental melting temperature T_i, and predicted melting temperature T_i' of all sequences[5]

$$\chi^2 = \sum_{i=1}^{N} \left(T_i' - T_i\right)^2 . \tag{16}$$

We will also refer to the average melting temperature deviation ΔT

$$\langle \Delta T \rangle = \frac{1}{N} \sum_{i=1}^{N} |T_i' - T_i| . \tag{17}$$

The predicted melting temperature T_i' is calculated from its melting index τ_i.[48] We expand the classical partition function, Eq. (14), into into non-diagonal terms of the transfer integral matrix.[49] The adimensional melting index τ_i for each sequence i is the order largest term of this expansion and is a measure of how much the sequence differs from a poly-CG reference sequence. This index is the correlated to the measured temperatures T_i using linear regression[50]

$$T_i' = f_0 + f_1 \tau_i, \tag{18}$$

where f_0 and f_1 are the regression coefficients that are recalculated every time the model parameters change.

It was not possible to use a length-dependent regression as previously used for DNA[5] and RNA[41] due to insufficient variations of sequence lengths in the melting temperature dataset. However, it was possible to perform separate regressions for each type of bulge. Further details concerning the model implementation are described in Refs. 48 and 49, and a free software implementation is reported in Ref. 51.

E. Initial model parameters and minimization procedure

The minimization of Eq. (16) is carried out separately for the four types of bulges, that is, melting temperatures of adenine, cytosine, guanosine and uracil bulges were separated and their parameters optimized independently. During the minimization we varied the Morse potential D of all bulge-flanking base pairs, of type BP^{B}, and the stacking interaction k of

VNNs and BNNs. For the canonical base pairs and FNNs we used the parameters previously calculated for RNA.[41] As the bulge is simulated by a vacant site X at the opposite strand, we use a zero Morse potential, $D = 0$, representing the absent interstrand interaction as shown in Fig. 1.

We performed two rounds of minimizations, independently for each type of bulge:

Initial minimization. The initial stacking interaction k of VNNs is set to a random value within $\pm 20\%$ of their corresponding FNNs. The same is done with the Morse potential D of bulge-flanking base pairs BP^{B}. For BNNs we use $k = 2.5$ meV/nm^2 as initial value[5] and for the FNNs we used the RNA parameters calculated in Ref. 41. We repeated this procedure 200 times for each bulge type, which takes around 4000 hours of processing time on 1.6 GHz processors. The resulting value of each parameter is obtained by averaging over all results provided by the minimization.

Final minimization. We use the averaged results from the previous round as new initial parameters. But now the experimental data set is randomly altered by small amounts corresponding to the experimental error of the temperature measurements.[5] Since the melting temperatures were derived from different sources we assumed an experimental error of 1.3 °C which is the same used for the calculation of canonical RNA.[41] This allowed us to estimate the influence of experimental error on our optimized parameters.[5] Again, all parameters are allowed to vary during this minimization to reach the smallest possible value of χ^2. This procedure was repeated again 200 times and the final results presented here are averaged values over all minimizations.

F. Availability

The bulge parameters calculated in this work were included in version 3.1 of our TfReg software,[51] and can be found at `http://tinyurl.com/tfregufmg`.

III. RESULTS AND DISCUSSION

The first result of this work is to show that the stacking asymmetry is of little relevance for the partition function of the PB Hamiltonian. Equation (14) shows that the Z_y asymmetric partition function is formally identical to is symmetric counterpart.[3] The Z_x

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

partition function Eq. (13) will be numerically different, however as it is always factored out in subsequent calculations this is of no practical consequence. The stacking constant k can now be interpreted as an harmonic average of two asymmetric constants k_u and k_v, as shown in Eq. (15). This is physically similar to an equivalent elastic constant of two springs in series, which is indeed the case as both harmonic potentials are written along the same direction. This result has consequences beyond its application to bulges as it helps explain why the stacking asymmetry so far has not been a problem for the application of the PB model. From a practical point of view it also means that we can apply the PB without modification to any situation that involves asymmetric stacking, we only need to be aware that the resulting stacking interaction is in fact an average of two intra-strand stacking factors.

The analysis of the asymmetric stacking Hamiltonian means that we can fully apply the existing framework of the PB model and perform parameter optimization in the same way as for the symmetric model.[5] Essentially, one only needs to be aware that the resulting stacking interaction is in fact an effective parameter describing the interaction of both strands, see Eq. (15). The PB model was developed for strands of equal length, yet bulges arise due to the fact that one strand is shorter than the other. We compensate for this by introducing a pseudo-base X which does not interact with the bulge, see Fig. 1. This non-interaction is represented by setting the Morse potential between bulge and pseudo-base to zero. For the intra-strand stacking representing the shorter strand which bridges the bulge, this will be split into two components. The resulting stacking interaction, at the bulge region, will therefore be an effective parameter representing both strands as well as the bulge region as discussed in section II B.

Following the procedure used in our work for canonical RNA,[41] the minimization was carried out in two main rounds. The first one using random initial parameters and the second varying the melting temperatures (section II C) to evaluate the influence of the experimental uncertainty on our results, see section II E. The error bars in the following figures represent the uncertainty evaluation of the second minimization round. The final average difference between measured and predicted temperatures, ΔT Eq. (17), was 0.44 °C for AX, 0.32 for CX, 1.42 °C for GC and 1.44 °C for UX, see also supplementary Tab. S5 for χ^2. The low ΔT may be due to the small amount of sequences available for AX and CX. On the other hand the larger ΔT could be due to the uncertainty regarding the true identity of the bulges measured by Kent *et al.*[26] as noted in section II C. Either way, these

values indicate that the minimization progressed well despite the difficulties represented by the regression scheme of Eq. (18). Note that although our results were optimized for the PB model, they can be used for the DPB model with very little change of ΔT. For instance ΔT increases from 0.44 °C to 0.45 °C for AX using $\rho = 2.0$ and $\alpha = 0.35$ Å^{-1}, see also supplementary Tab. S5.

In Fig. 2 we show the averaged Morse parameters D for the base-pairs in the vicinity of a bulge, type BP$^{\mathrm{B}}$. They are ordered to increasing difference to the canonical unperturbed base-pairs. G bulges clearly stand out with very low Morse potentials which would suggest an important perturbation or even a disruption of the neighboring AU and CG base pairs. For U, A and C bulges there appears to be a stabilization of the CG base-pairs while AU base-pairs appear to be less stable, except for C bulges. This increased on-site stabilization may not be entirely due to the base-pair hydrogen bonding, considering that the Morse potential is an effective potential it may also account for influences of the bulged base towards its neighboring bases. In other words, some of the increased Morse potentials may be due to some interstrand interaction of the bulge.

While the results for the Morse potential indicate an important influence of the GX bulge, one can not understand the influence of the bulge solely from the on-site interaction. The model Hamiltonian, Eq. (1), shows that there is an energetic balance between the on-site Morse potential and the harmonic nearest-neighbor potential, that is, a loss of stabilization due to hydrogen bonding could be compensated by the stacking interaction. Figure 3 shows the harmonic constants k for bulge perturbed BNNs ordered to increasing difference δ to their equivalent unperturbed FNN (shown as black squares),

$$\delta = |k_{\mathrm{unperturbed}} - k_{\mathrm{perturbed}}|. \tag{19}$$

Fig. 3b shows that also for the stacking interaction there is a strong variation for GX bulges, yet here also CX bulges also show important deviation from the equivalent FNNs. Figure 4 shows the stacking interaction for VNNs, that is canonical NNs next to a bulge. For most VNNs there is very little change compared for their equivalent FNNs. Note that for VNNs we did not use as initial k value that of their FNNs, therefore in the case where the VNNs converged to the same average value as the FNNs this was not due to a biased minimization. As these VNNs do not directly involve a bulge, even the largest variations are much smaller that those observed for the BNNs of Fig. 3b. The majority of VNNs with large deviations are

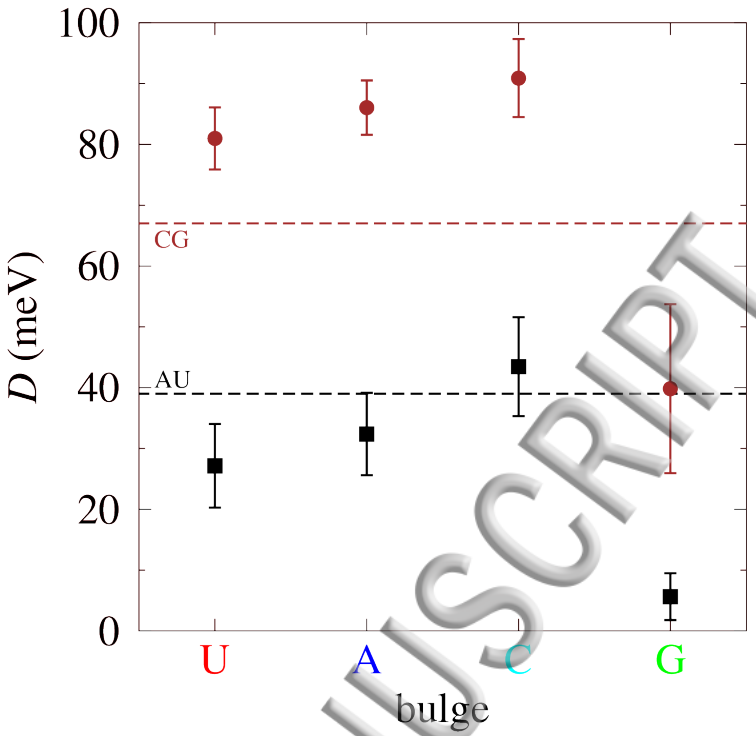


FIG. 2. Average Morse potential D of AU^B (black boxes) and CG^B (brown bullets) base pairs neighboring a bulge B, with B=A, C, G or U. The dashed lines are the corresponding Morse potential of unperturbed base pairs AU and CG.[41] Bulges are shown ordered by total squared difference to the unperturbed Morse potential, with U being the bulge that caused the smallest, and G the largest perturbation to base pairs. The color coding of the horizontal labels is the same used for the remaining figures.

for CX and GX bulges similar to what was observed for BNNs, yet intriguingly the largest variation of all is for the UX related $AUpUA^U$ which corresponds to the least perturbed Morse potential.

A. Opening profiles

The results of inter-strand and stacking interaction shown in the previous section point to an intricate interplay between these two components of the PB Hamiltonian. Therefore, it is not straightforward to infer the effect of the bulge simply by inspecting those parameters. Fortunately, the PB model allows the calculation of the average opening profiles by computing the expected value for y_n from the partition function of Eq. (14). This has the

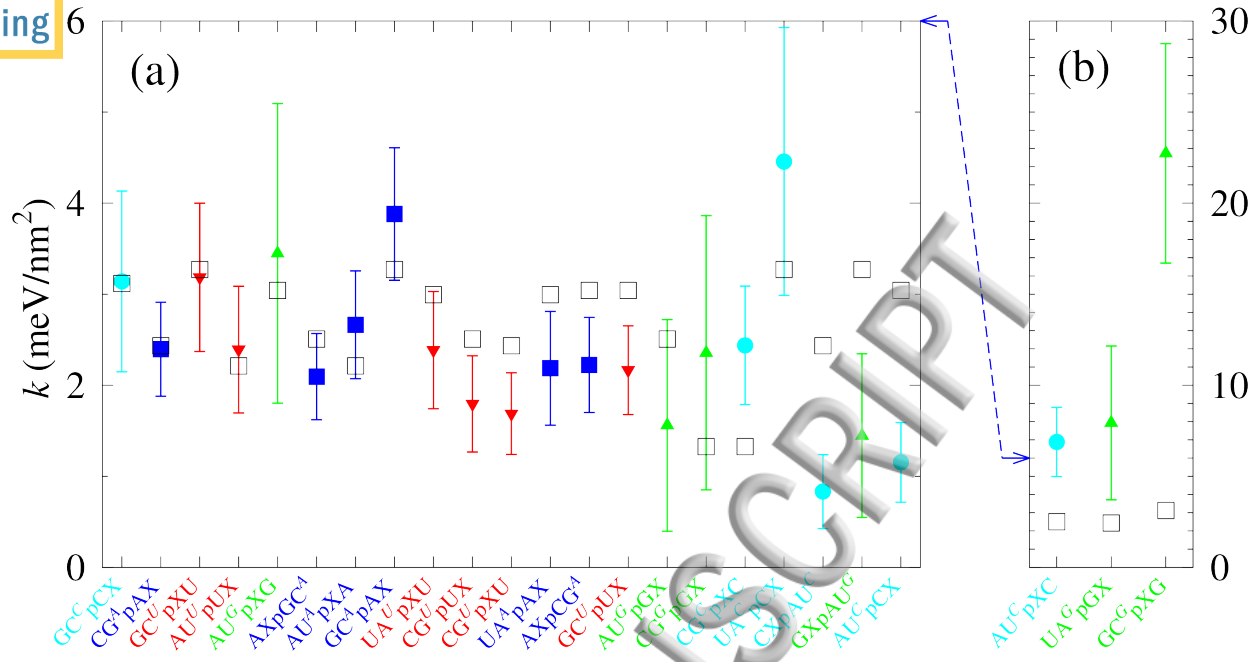


FIG. 3. Average stacking constants k of BNNs ordered to increasing difference δ, Eq. (19). The stacking constants with the largest difference were separated in panel (b) with a different vertical scale as in panel (a). The correspondence of scale between panels (a) and (b) is shown by the dashed blue line. Color coding follows Fig. 2. Black squares are the stacking constants of the corresponding unperturbed FNNs from Ref. 41.

advantage of showing the combined effect of all these parameters as well as the non-linear effect of the model Hamiltonian which may extend over the whole sequence. However, we need to caution that when analyzing these profiles one needs to be aware that, due to a limitation of the model, the reference temperature in the partition functions needs to be set much lower than the actual melting temperatures, here we used 180 K.[51] Therefore the opening profiles should only be seen as a qualitative representation of the helix stability.[51]

In Fig. 5 we show some examples of opening profiles for several types of bulges. For comparison, we also show what we call the corresponding parental sequences, that is, the equivalent sequence with the bulge position filled in by a real complementary base. The sequence of Fig. 5a was adapted from Fig. 1a of Ref. 2, which shows an intercalated AX bulge, by adding extra CG base pairs at the ends. It is interesting to note that the strand opening induced by the bulge is of the same order as that of its terminals. In other words, it is an important perturbation yet comparable to other common structures in the RNA sequence.

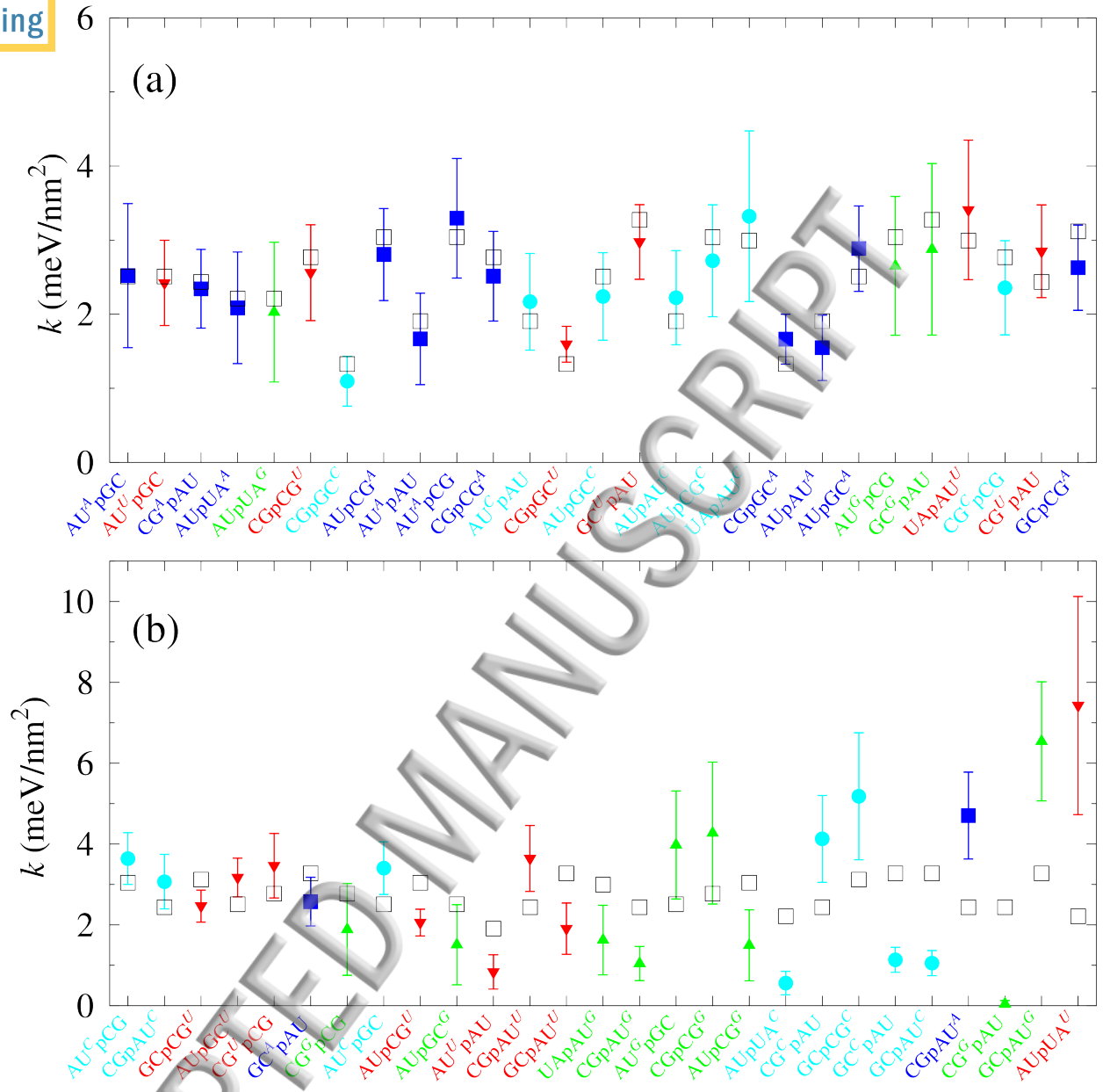


FIG. 4. Average stacking constants k of VNNs ordered to increasing difference δ, Eq. (19). Note that the vertical scale of panels (a) and (b) differ. Color coding follows Fig. 2. Black squares are the stacking constants of the corresponding FNNs from Ref. 41.

However, the bulge perturbation extends over two base-pair positions, towards the left-hand side, due to the presence of a UA base pair. This is markedly different from the situation presented in Fig. 5b, showing two sequences from Popenda *et al.*[52], where the bulge is surrounded by two GC base-pairs to each side. The sequences of Fig. 5b is one of very few NMR studies we are aware of for short RNA in solution with group I bulges. Another

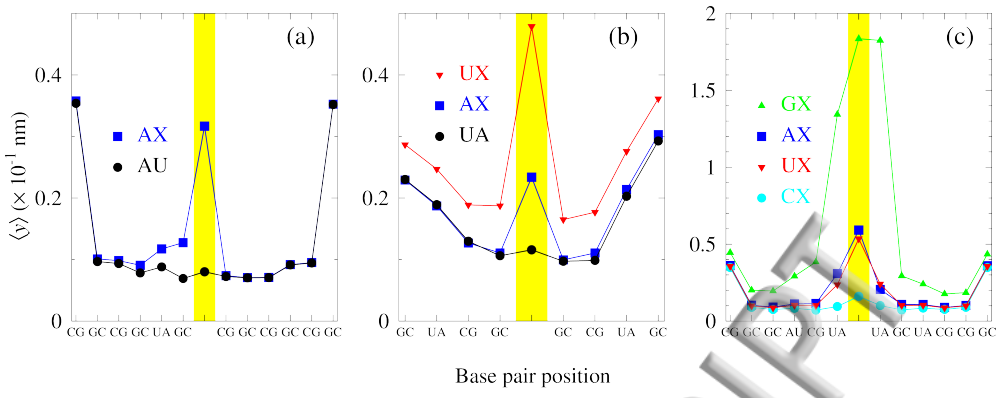


FIG. 5. Average opening profiles for several example sequences. The position corresponding to the bulge is highlighted by a yellow area for UX (red down triangles), AX (blue boxes), CX (cyan bullets) and GX (green up triangles). The corresponding parental sequences are shown as black bullets. Sequences are adapted from (a) Ref. 2 and (b) Ref. 52, while (c) is an arbitrary sequence. Note that the vertical scale of panel (c) differs from (a) and (b). Color coding follows Fig. 2.

sequence studied with NMR by van den Hoogen *et al.*[7] is almost identical to the one of Fig. 5b and is shown in supplementary Fig. S2. Figure 5c shows an arbitrary sequence with all four bulges, where the GX bulge clearly stands out for its very strong perturbation, especially to its neighboring base-pairs.

Given the examples of opening profiles presented in Fig. 5, it would be tempting to try a correlation of the average openings with structural data. For instance, considering the sequence of Fig. 5b, the NMR data reports a side-by-side formation for AX and a looped-out conformation for UX.[52] This seems to be consistent with the opening profile, that is, a small opening for AX and a much larger opening for UX. While this agreement is encouraging, there is not enough structural data for short sequences with single bulges in solution to extend this analysis. We are aware of very few NMR structural studies[7,52] that match the experimental conditions to that of the melting temperature measurements from which we derive our results. We have not considered the more abundant structural data on ribosomal RNAs as it is evident that bulges are very sensitive the proximity of loops and proteins which are present in these cases, that is, we would be comparing exceedingly different experimental conditions.

IV. CONCLUSIONS

We derived the conditions for which the Peyrard-Bishop mesoscopic model can describe highly asymmetric RNA strands. These conditions hold for any model with a harmonic stacking potential which includes the solvent potential Hamiltonian.[39] This was applied successfully to model the melting temperatures of group I RNA bulges and we showed that the resulting parameters can be used also for other PB-derived models such as the DPB model.[44] We showed that guanosine bulges display the largest difference in model parameters when compared to Watson-Crick base pairs resulting in strong perturbations to their neighboring base pairs. These results make it possible to extend this type of approach to other types of bulges in RNA as well as in DNA.

SUPPLEMENTARY MATERIAL

Figure S1 illustrating group I and group II bulges. Figure S2 showing additional examples of average opening profiles with U-bulges. Tables S1–S4 with all sequences, experimental and predicted temperatures used in this work. Table S5 with fitting quality parameters for the PB and the PBD model.

ACKNOWLEDGMENTS

The authors acknowledge the agencies that funded this research: CNPq, Fapemig and Capes.

REFERENCES

[1]B. M. Znosko, S. B. Silvestri, H. Volkman, B. Boswell, and M. J. Serra, "Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges," Biochemistry **41**, 10406–10417 (2002).

[2]T. Hermann and D. J. Patel, "RNA bulges as architectural and recognition motifs," Structure **8**, R47–R54 (2000).

[3]M. Peyrard and A. R. Bishop, "Statistical mechanics of a nonlinear model for DNA denaturation," Phys. Rev. Lett. **62**, 2755–2757 (1989).

[4]P. Šulc, F. Romano, T. E. Ouldridge, J. P. Doye, and A. A. Louis, "A nucleotide-level coarse-grained model of RNA," J. Chem. Phys. **140**, 06B614_1 (2014).

[5]G. Weber, J. W. Essex, and C. Neylon, "Probing the microscopic flexibility of DNA from melting temperatures," Nat. Phys. **5**, 769–773 (2009).

[6]J. Zhu and R. M. Wartell, "The effect of base sequence on the stability of RNA and DNA single base bulges," Biochemistry **38**, 15986–15993 (1999).

[7]Y. T. van den Hoogen, A. A. van Beuzekom, E. de Vroom, G. A. van der Marel, J. H. van Boom, and C. Altona, "Bulge-out structures in the single-stranded trimer AUA and in the duplex (CUGGUGCGG).(CCGCCCAG). a model-building and NMR study," Nucleic Acids Res. **16**, 5013–5030 (1988).

[8]W. A. Hastings, Y. G. Yingling, G. S. Chirikjian, and B. A. Shapiro, "Structural and dynamical classification of RNA single-base bulges for nanostructure design," Journal of Computational and Theoretical Nanoscience **3**, 63–77 (2006).

[9]L. Popenda, R. W. Adamiak, and Z. Gdaniec, "Bulged adenosine influence on the RNA duplex conformation in solution," Biochemistry **47**, 5059–5067 (2008).

[10]J. W. Nelson, F. H. Martin, and I. Tinoco, "DNA and RNA oligomer thermodynamics: The effect of mismatched bases on double-helix stability," Biopolymers **20**, 2509–2531 (1981).

[11]S. A. White and D. E. Draper, "Single base bulges in small RNA hairpins enhance ethidium binding and promote an allosteric transition," Nucleic Acids Research **15**, 4049–4064 (1987).

[12]S. A. White and D. E. Draper, "Effects of single-base bulges on intercalator binding to small RNA and DNA hairpins and a ribosomal RNA fragment," Biochemistry **28**, 1892–1897 (1989).

[13]J. W. Nelson and I. Tinoco Jr, "Ethidium ion binds more strongly to a DNA double helix with a bulged cytosine than to a regular double helix," Biochemistry **24**, 6416–6421 (1985).

[14]E. Ennifar, M. Yusupov, P. Walter, R. Marquet, B. Ehresmann, C. Ehresmann, and P. Dumas, "The crystal structure of the dimerization initiation site of genomic HIV-1 RNA reveals an extended duplex with two adenine bulges," Structure **7**, 1439–1449 (1999).

[15]Y. Xiong and M. Sundaralingam, "Two crystal forms of helix II of xenopus laevis 5S rRNA with a cytosine bulge," RNA **6**, 1316–1324 (2000).

[16]Y. Xiong, J. Deng, C. Sudarsanakumar, and M. Sundaralingam, "Crystal structure of an RNA duplex r(gugucgcac)$_2$ with uridine bulges," Journal of Molecular Biology **313**, 573–582 (2001).

[17]L. Joshua-Tor, D. Rabinovich, H. Hope, F. Frolow, E. Appella, and J. L. Sussman, "The three-dimensional structure of a DNA duplex containing looped-out bases," Nature **334**, 82–84 (1988).

[18]D. A. LeBlanc and K. M. Morden, "Thermodynamic characterization of deoxyribooligonucleotide duplexes containing bulges," Biochemistry **30**, 4042–4047 (1991).

[19]H. Beltz, J. Azoulay, S. Bernacchi, J.-P. Clamme, D. Ficheux, B. Roques, J.-L. Darlix, and Y. Mély, "Impact of the terminal bulges of HIV-1 cTAR DNA on its stability and the destabilizing activity of the nucleocapsid protein NCp7," Journal of Molecular Biology **328**, 95–108 (2003).

[20]F. Tanaka, A. Kameda, M. Yamamoto, and A. Ohuchi, "Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop," Biochemistry **43**, 7143–7150 (2004).

[21]A. Barthel and M. Zacharias, "Conformational transitions in RNA single uridine and adenosine bulge structures: a molecular dynamics free energy simulation study," Biophysical Journal **90**, 2450–2462 (2006).

[22]P. Kumar, J. Lehmann, and A. Libchaber, "Kinetics of bulge bases in small RNAs and the effect of pressure on it," PloS One **7**, e42052 (2012).

[23]J. S. Schreck, T. E. Ouldridge, F. Romano, A. A. Louis, and J. P. Doye, "Characterizing the bending and flexibility induced by bulges in DNA duplexes," J. Chem. Phys. **142**, 165101 (2015).

[24]J. M. Blose, M. L. Manni, K. A. Klapec, Y. Stranger-Jones, A. C. Zyra, V. Sim, C. A. Griffith, J. D. Long, and M. J. Serra, "Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops," Biochemistry **46**, 15123–15135 (2007).

[25]M. D. McCann, G. F. Lim, M. L. Manni, J. Estes, K. A. Klapec, G. D. Frattini, R. J. Knarr, J. L. Gratton, and M. J. Serra, "Non-nearest-neighbor dependence of the stability for RNA group II single-nucleotide bulge loops," RNA **17**, 108–119 (2011).

[26]J. L. Kent, M. D. McCann, D. Phillips, B. L. Panaro, G. F. Lim, and M. J. Serra, "Non-nearest-neighbor dependence of stability for group III RNA single nucleotide bulge loops,"

RNA **20**, 825–834 (2014).

[27]C. E. Longfellow, R. Kierzek, and D. H. Turner, "Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides," Biochemistry **29**, 278–285 (1990).

[28]C. A. S. A. Minetti, D. P. Remeta, R. Dickstein, and K. J. Breslauer, "Energetic signatures of single base bulges: thermodynamic consequences and biological implications," Nucleic Acids Res. **38**, 97–116 (2010).

[29]D. R. Groebe and O. C. Uhlenbeck, "Thermal stability of RNA hairpins containing a four-membered loop and a bulge nucleotide," Biochemistry **28**, 742–747 (1989).

[30]G. F. Lim, G. E. Merz, M. D. McCann, J. M. Gruskiewicz, and M. J. Serra, "Stability of single-nucleotide bulge loops embedded in a GAAA RNA hairpin stem," RNA **18**, 807–814 (2012).

[31]A. L. Dishler, E. L. McMichael, and M. J. Serra, "Determination of the secondary structure of group II bulge loops using the fluorescent probe 2-aminopurine," RNA **21**, 975–984 (2015).

[32]J. C. Tomcho, M. R. Tillman, and B. M. Znosko, "Improved model for predicting the free energy contribution of dinucleotide bulges to RNA duplex stability," Biochemistry **54**, 5290–5296 (2015).

[33]T. D. Amarante and G. Weber, "Evaluating hydrogen bonds and base stackings of single, tandem and terminal GU in RNA mismatches with a mesoscopic model," J. Chem. Inf. Model. **56**, 101–109 (2016), http://dx.doi.org/10.1021/acs.jcim.5b00571.

[34]G. Kalosakas and S. Ares, "Dependence on temperature and guanine-cytosine content of bubble length distributions in DNA," J. Chem. Phys. **130**, 235104 (2009).

[35]G. Kalosakas, "Charge transport in DNA: Dependence of diffusion coefficient on temperature and electron-phonon coupling constant," Phys. Rev. E **84**, 051905 (2011).

[36]N. N. Oiwa, C. E. Cordeiro, and D. W. Heermann, "The electronic behavior of zinc-finger protein binding sites in the context of the DNA extended ladder model," Frontiers in Physics **4**, 13 (2016).

[37]A. E. B. Pupo, F. Falo, and A. Fiasconaro, "DNA overstretching transition induced by melting in a dynamical mesoscopic model," J. Chem. Phys. **139**, 095101 (2013).

[38]M. Zoli, "J-factors of short DNA molecules," J. Chem. Phys. **144**, 214104 (2016).

[39]G. Weber, "Sharp DNA denaturation due to solvent interaction," Europhys. Lett. **73**, 806–811 (2006).

[40]A. Singh and N. Singh, "Phase diagram of mechanically stretched DNA: The salt effect," Physica A: Statistical Mechanics and its Applications **392**, 2052–2059 (2013).

[41]G. Weber, "Mesoscopic model parametrization of hydrogen bonds and stacking interactions of RNA from melting temperatures," Nucleic Acids Res. **41**, e30 (2013).

[42]I. Ferreira, T. D. Amarante, and G. Weber, "DNA terminal base pairs have weaker hydrogen bonds especially for AT under low salt concentration," J. Chem. Phys. **143**, 175101 (2015).

[43]B. Alexandrov, V. Gelev, Y. Monisova, L. Alexandrov, A. Bishop, K. Rasmussen, and A. Usheva, "A nonlinear dynamic model of DNA with a sequence-dependent stacking term," Nucleic Acids Res. **37**, 2405 (2009).

[44]T. Dauxois, M. Peyrard, and A. R. Bishop, "Entropy-driven DNA denaturation," Phys. Rev. E **47**, R44–R47 (1993).

[45]N. Singh and Y. Singh, "Effect of defects on thermal denaturation of DNA oligomers," Phys. Rev. E **64**, 042901 (2001).

[46]P. M. Morse, "Diatomic molecules according to the wave mechanics. II. Vibrational levels," Physical Review **34**, 57 (1929).

[47]Y.-L. Zhang, W.-M. Zheng, J.-X. Liu, and Y. Z. Chen, "Theory of DNA melting based on the Peyrard-Bishop model," Phys. Rev. E **56**, 7100–7115 (1997).

[48]G. Weber, N. Haslam, N. Whiteford, A. Prügel-Bennett, J. W. Essex, and C. Neylon, "Thermal equivalence of DNA duplexes without melting temperature calculation," Nat. Phys. **2**, 55–59 (2006).

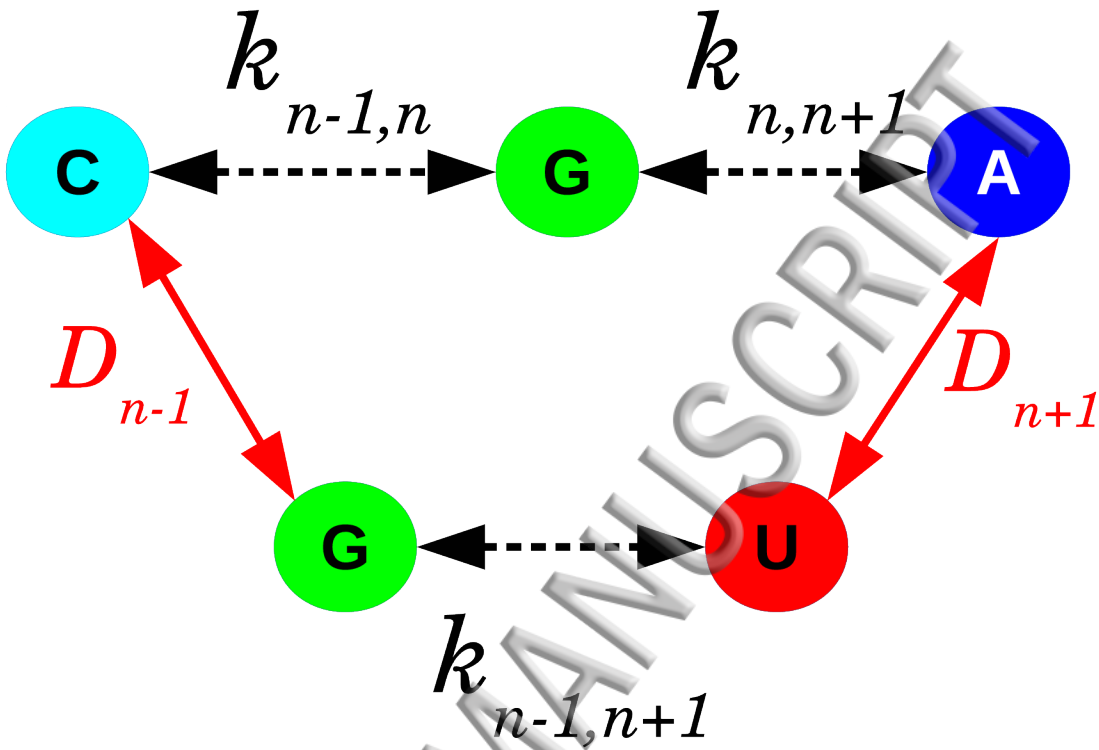
[49]G. Weber, N. Haslam, J. W. Essex, and C. Neylon, "Thermal equivalence of DNA duplexes for probe design," J. Phys.: Condens. Matter **21**, 034106 (2009).

[50]R. V. Maximiano and G. Weber, "Deoxyinosine mismatch parameters calculated with a mesoscopic model result in uniform hydrogen bonding and strongly variable stacking interactions," Chem. Phys. Lett. **631–632**, 87–91 (2015).

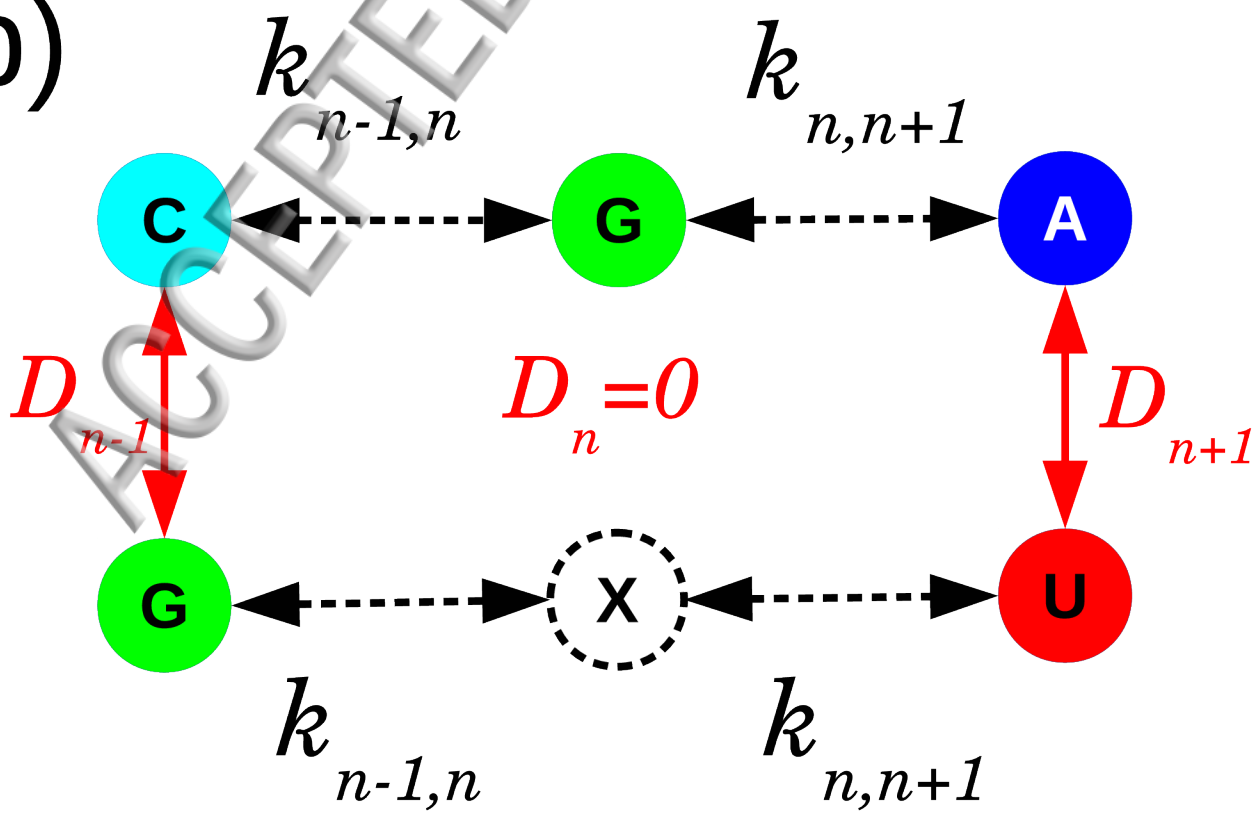
[51]G. Weber, "TfReg: Calculating DNA and RNA melting temperatures and opening profiles with mesoscopic models," Bioinformatics **29**, 1345–1347 (2013).

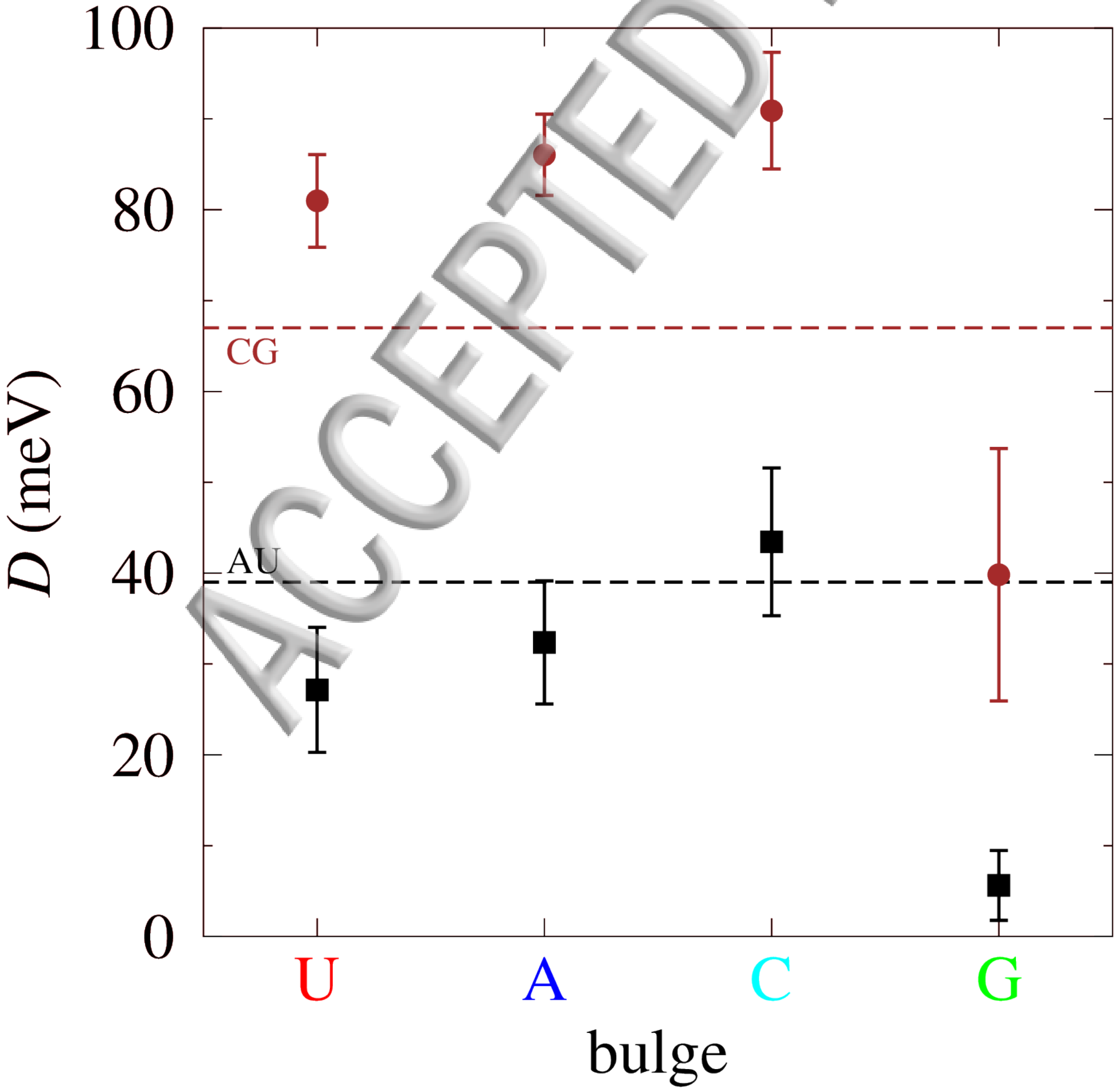
[52]L. Popenda, L. Bielecki, Z. Gdaniec, and R. W. Adamiak, "Structure and dynamics of adenosine bulged RNA duplex reveals formation of the dinucleotide platform in the C: GA triple," Arkivoc **3**, 130–144 (2009).

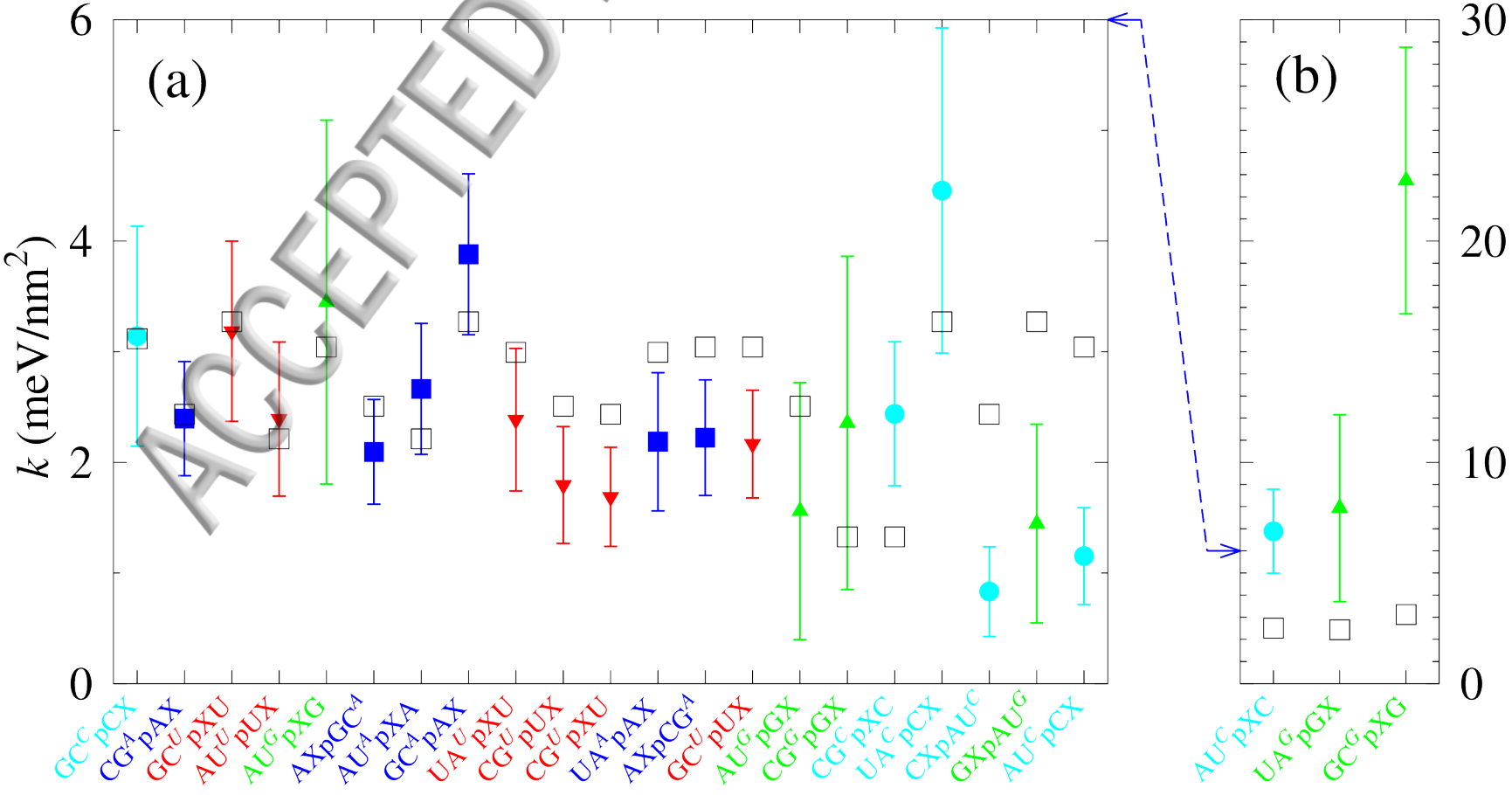
(a)

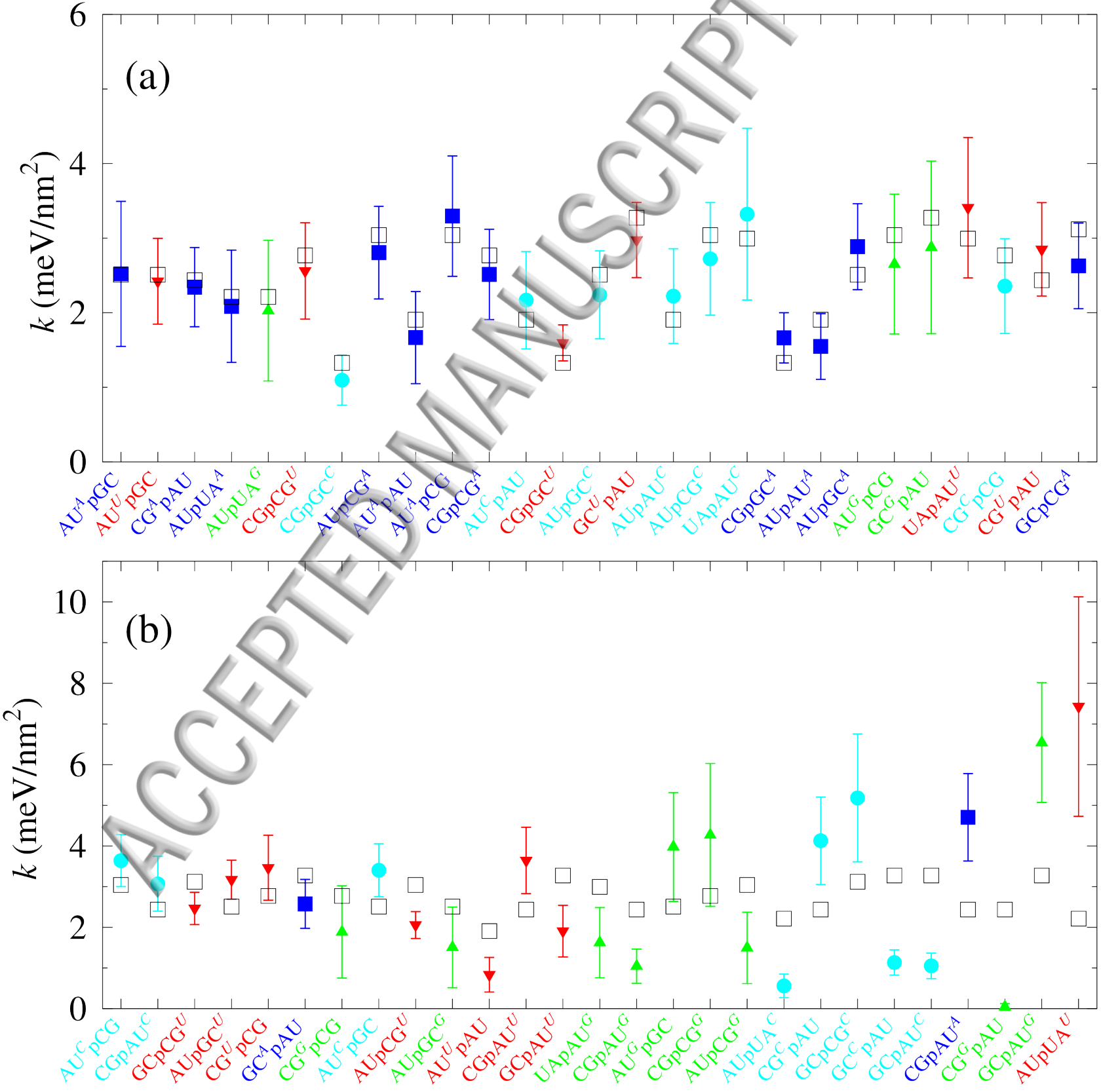


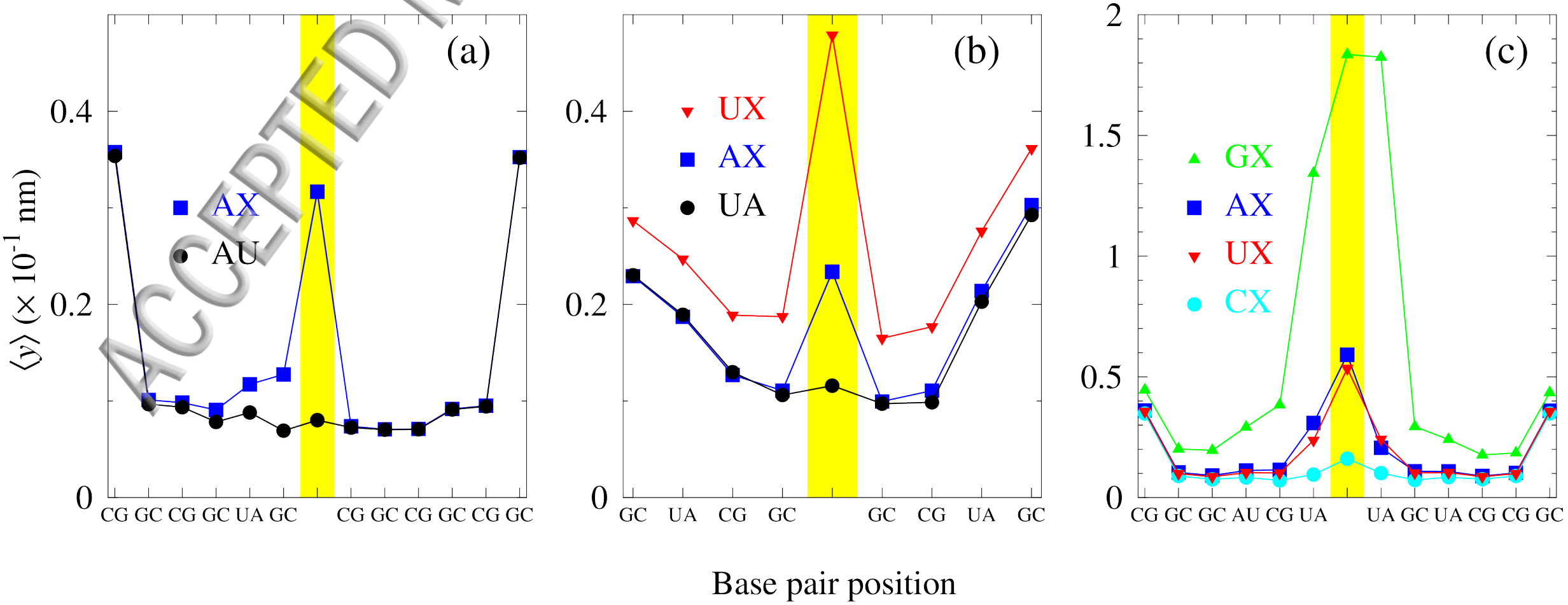
(b)






(a)
(b)
k (meV/nm^2)
0
2
4
6
0
10
20
30
GC^{C}pCX
CG^{A}pAX
GC^{U}pXU
AU^{U}pUX
AU^{G}pXG
AXpGC^A
AU^{A}pXA
GC^{A}pAX
UA^{U}pXU
CG^{U}pUX
CG^{U}pXU
UA^{A}pAX
AXpCG^A
GC^{U}pUX
AU^{G}pGX
CG^{G}pGX
CG^{C}pXC
UA^{C}pCX
CXpAU^C
GXpAU^G
AU^{C}pCX
AU^{C}pXC
UA^{G}pGX
GC^{G}pXG




(a)
(b)
(c)
AX
AU
UX
AX
UA
GX
AX
UX
CX
⟨y⟩ (× 10^-1 nm)
0
0.2
0.4
0.5
1
1.5
2
CG GC CG GC UA GC CG GC CG GC CG GC
GC UA CG GC GC CG UA GC
CG GC GC AU CG UA UA GC UA CG CG GC
Base pair position