PHYSICAL REVIEW E **103**, 042503 (2021)

Statistical physics of DNA hybridization

Carlos A. Plata,[1,2] Stefano Marni,[3] Amos Maritan,[1] Tommaso Bellini,[3] and Samir Suweis[1]
[1]*Dipartimento di Fisica "G. Galilei," INFN, Università di Padova, 35131 Padova, Italy*
[2]*Université Paris-Saclay, CNRS, LPTMS, 91405 Orsay, France*
[3]*Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano, Segrate, MI I-20090, Italy*

(Received 10 November 2020; accepted 23 March 2021; published 19 April 2021)

Deoxyribonucleic acid (DNA) hybridization is at the heart of countless biological and biotechnological processes. Its theoretical modeling played a crucial role, since it has enabled extracting the relevant thermodynamic parameters from systematic measurements of DNA melting curves. In this article, we propose a framework based on statistical physics to describe DNA hybridization and melting in an arbitrary mixture of DNA strands. In particular, we are able to analytically derive closed expressions of the system partition functions for any number N of strings and explicitly calculate them in two paradigmatic situations: (i) a system made of self-complementary sequences and (ii) a system comprising two mutually complementary sequences. We derive the melting curve in the thermodynamic limit ($N \to \infty$) of our description, which provides a full justification for the extra entropic contribution that in classic hybridization modeling was required to correctly describe within the same framework the melting of sequences either self-complementary or not. We thus provide a thorough study comprising limit cases and alternative approaches showing how our framework can give a comprehensive view of hybridization and melting phenomena.

DOI: 10.1103/PhysRevE.103.042503

I. INTRODUCTION

The selective interaction and pair formation of nucleic acid polymers and oligomers is the basic mechanism enabling gene coding and replication. It also at the core of a wealth of other biological processes, such as gene regulation and secondary structuring of Ribonucleic acid (RNA), of biotechnologies, such as Polymerase chain reaction and Systematic evolution of ligands by exponential enrichment, and of Deoxyribonucleic acid (DNA)-based nanotechnologies and DNA origami. These interactions are based, for the largest part, on the Watson-Crick pairing of complementary bases [1]. DNA and RNA pairing has a limited interval of stability. When submitted to enough stress, either physical (temperature, competing forces) or chemical (solvent composition) the double helix denaturates. In particular, thermal denaturation, leading to the unfolding of a single strand or to the splitting of a duplex, is of paramount importance in technologies, including amplification, screening, and sequencing [2–6]. Therein, tailored nucleic acids sequences usually work as probes unveiling biological information of a sample or mediating for amplification during polymerase chain reaction.

Thermal denaturation of duplexes provides an easy but crucial access into DNA and RNA thermodynamics. Melting curves, typically represented as plots of the fraction of paired DNA oligomers vs. temperature (T), can be experimentally accessed either via UV absorbance, based on the so-called hypochromicity effect [7–11], or by measuring the fluorescent emission of environment-sensitive DNA or RNA-binding fluorophores. In both cases, measurements directly yield, after suitable normalization, the melting curve. Melting curves are mainly characterized by their characteristic temperature, the so-called melting temperature T_m, and by the sharpness of the bound-unbound transition, both depending on the energetics of pair interactions. Many statistical physics models [12–21], based for instance on the Ising model or Haminoltonian mechanics, have aimed at the calculation of partition function of the system to describe such interactions. Nevertheless, the most accepted and used model describing the interaction of nucleic acid duplexes is essentially different in its approach. This is the nearest-neighbor (NN) model, which is a macroscopic the NN model based on reaction equilibrium.

Originally introduced by Tinoco *et al.* in order to study the thermostability of RNA [22,23], the NN model has developed into a detailed protocol enabling to predict T_m of strands of arbitrary length and sequence. The model has been recently applied even under circumstances of molecular crowding [24]. According to the NN model, the free energy involved in the pairing of two strands depends on the specific nucleotide sequence and can be obtained as a sum of contributions stemming not just from individual base pairs, but from couples of nearest base pairs. Such a splitting incorporates the notion that the duplex stability is largely depending on stacking forces, acting between neighboring molecular planes of paired bases. The computation of melting temperature T_m in the NN model requires knowing the specific contributions to the free energy conveyed by each couple of nearest-neighbor paired bases, which depends on the identity and location of the four nucleobases involved in such a segment of double helical structure. Conversely, the knowledge of the T_m of a wide number duplexes of variable length and sequence enables the determination of the thermodynamic parameters associated with each type of couples of nearest base pairs [25–34].

The extraction of such thermodynamic parameters necessarily involves a comparison between experimental and theoretical melting curves. The latter is typically based on a two-state transition model, which neglects any state intermediate between the intact duplex and the fully melted strands [35–37]. The theoretical derivation of the melting curve stands on the calculation of either the system partition function or of the equilibrium constant in the balance equation of the two-state model. The functional shape of the melting curve slightly varies when considering complementary or self-complementary sequences. However, in order to use the same model for describing both kind of sequences, an extra symmetric correction is introduced in self-complementary sequences [27,29,30,33,34,36]. Specifically, the resulting symmetric correction is such that $\Delta S_{\text{sym}} = -1.4$ cal K^{-1} $\text{mol}^{-1} \simeq -R\ln 2$ [33]. This entropic contribution can be thought of the twofold rotational symmetry in self-complementary duplexes [38]. Nevertheless, in the traditional approach, this is obtained as an *a posteriori* explanation after fitting a set of mixed data including both kind of sequences.

In this work we provide an alternative modeling framework to describe DNA hybridization. Specifically, our aim is twofold. On the one hand, we want to exploit the parallelism between the problem of DNA melting and the process of dissociation and recombination of diatomic molecules. From the physical perspective, they are equivalent problems and the latter has been very well studied in the past [39,40]. Specifically, solutions for the problem involving diatomic molecules have been obtained with exact mathematical treatment even for systems with a low number of components, away from the thermodynamic limit. On the other hand, we want to shed some light on the origin of the symmetric correction for self-complementary sequences, which is normally used in NN model computations. As we will show, in our approach there is no need for such extra entropic contribution, it will spontaneously emerge due to the degeneracy of the system, when one focus on the entropic contribution coming solely from the interaction between the strands. In other words, our work actually provides a full justification for the origin of the symmetry factor.

To achieve our goals, we put forward a clear formulation of the melting curve based on the partition function calculation. We present exact results for arbitrary DNA mixtures that, afterwards, are thoroughly analyzed in the thermodynamic limit for systems with experimental relevance.

II. MODEL

A. Partition function for a self-complementary mixture

We consider a system made by N identical self-complementary oligomers enclosed in a total volume V at temperature T. The system is characterized thus by a concentration $c = N/V$. Oligomers can be either free or paired, resulting in a duplex. Free oligomers give a contribution G_f to the global Gibbs free energy, whereas each duplex contributes with G_p. Henceforth, the free-energy difference due to the inner degrees of freedom is $\Delta\tilde{G} = \Delta\tilde{H} - T\Delta\tilde{S} = G_p - 2G_f$. There is an extra entropic cost of the pairing, since paired oligomers cannot explore freely the full phase space. Specifically, we consider that pairing interaction has a finite range characterized by a volume $\tilde{V}$, which formally means that if oligomers i and j are paired, and then the relative position vector $\vec{r}_i - \vec{r}_j$ of their centers of mass is enclosed in a volume $\tilde{V}$.

Since the system is in contact with a thermal bath at temperature T, the canonical distribution provides us with the probability $p(n)$ of having n duplexes formed in the system,

$$p(n) \propto f(n)\exp\{-\beta[(N-2n)G_f + nG_p]\} \times \int_{\mathcal{D}} d\vec{r}_1\cdots d\vec{r}_N d\vec{p}_1\cdots d\vec{p}_N \exp\left(-\beta\sum_{i=1}^{N}\frac{\vec{p}_i\cdot\vec{p}_i}{2m}\right), \tag{1}$$

where $\beta = (k_B T)^{-1}$, k_B is the Boltzmann constant (which has to be substituted by the gas constant R when the free energy is expressed in molar quantities), $\mathcal{D}$ is the domain of the $6N$-dimensional available phase space, and $f(n)$ is the number of possible different configurations with exactly n duplexes formed in our system (with a total number of oligomers equal to N). Note that the thermodynamic potential of interest is the Gibbs free energy, which has been previously defined, since we consider that the number of oligomers, the pressure, and the temperature is kept constant. Moreover, in this expression, we take into account the degeneracy of the state through $f(n)$, whereas the integral and the first exponential account for the phase space explored by the centers of mass of the oligomers and the inner degrees of freedom, respectively. In the integral of Eq. (1), there are $N-2n$ free oligomers and $2n$ paired oligomers. Therefore,

$$\int_{\mathcal{D}} d\vec{r}_1\cdots d\vec{r}_N d\vec{p}_1\cdots d\vec{p}_N \exp\left(-\beta\sum_{i=1}^{N}\frac{\vec{p}_i\cdot\vec{p}_i}{2m}\right) = \left(\frac{2\pi m}{\beta}\right)^{\frac{3N}{2}} V^N\left(\frac{\tilde{V}}{V}\right)^n. \tag{2}$$

We thus get that the probability of having n couples of paired oligomers is

$$p(n) = \frac{1}{Z}f(n)\left[\frac{\tilde{V}}{V}\exp(-\beta\Delta\tilde{G})\right]^n, \tag{3}$$

where we have absorbed conveniently some constant factors with respect to n in the partition function Z. It comes in handy to get rid the appearance of $\tilde{V}$ from our expressions. To do that, we write it as an additive entropic contribution, defining the weighted Boltzmann factor

$$\zeta = [c]\exp(-\beta\Delta G), \tag{4}$$

where $[c]$ is the number concentration measured in mol/l and $\Delta G = \Delta H - T\Delta S$ is the global free energy, which includes the effect of $\tilde{V}$, with $\Delta H = \Delta\tilde{H}$ and $\Delta S = \Delta\tilde{S} + k_B\ln(N_A[\tilde{V}])$ with N_A being Avogadro's number (the square brackets in $\tilde{V}$ just indicates that it is consistently measured in liters). Note that, usually, experimental measurements of the entropy of a duplex hardly distinguishes between the different contributions to the total entropy change, which makes our choice especially convenient. Importantly, the choice of units for the concentration and the volume has to be consistent. We have chosen the liter since this is the unit traditionally used in literature and consequently it leads to the values of ΔS reported in literature. We have freedom to redefine a different factor before the exponential in Eq. (4) introducing an extra correcting contributions in ΔS, or changing its initialization value, but we keep our formalism consistent with the conventional approach.

Using the Boltzmann factor introduced above, we obtain finally that the probability of observing n duplexes, or, equivalently, the probability of being in the hybridization state n, is

$$p(n) = \frac{1}{Z} f(n) \left(\frac{\zeta}{N} \right)^n, \tag{5}$$

where the partition function guarantees the correct normalization, that is,

$$Z = \sum_n f(n) \left(\frac{\zeta}{N} \right)^n. \tag{6}$$

The sum above is defined over all possible values of n, which are the integers such that $0 \leqslant 2n \leqslant N$.

At last, we need to write the degeneracy $f(n)$ of an hybridization state. There are several ways to carry out this combinatorics calculation. We reach the result

$$f(n) = \frac{N!}{2^n (N - 2n)! n!}, \tag{7}$$

where we have taken into account directly the different ways of relabelling equivalent states. The numerator stands by the number of ways of ordering all the strands. The factors in the denominator takes into account the different symmetries of a hybridization state under exchanging equivalent oligomers (relabelling). One can exchange oligomers within each duplex, all the free oligomers, and all the duplexes, which correspond to the factors 2^n, $(N - 2n)!$ and $n!$ in the denominator. The factor 2^n in the denominator takes into account the symmetry of the pairing between homologous sequences. We highlight that our ΔS does not take into account any reduction due to the symmetry, carrying just information of the entropy difference because of the interaction between the strands. It is indeed this $1/2$ factor appearing explicitly in the degeneracy which reflects the aforementioned symmetry reduction.

The partition function contains all the relevant statistical information which is useful to derive the expected values for the hybridization state. Specifically one can find straightforward that the average number of duplexes equals

$$\langle n \rangle = \zeta \frac{\partial \ln Z}{\partial \zeta}. \tag{8}$$

Later, we will perform explicit computation of both the partition function and the average number of duplexes. Nevertheless, we focus first on studying a general formalism for analyzing arbitrary mixtures of oligomers.

B. Partition function for an arbitrary mixture

In this subsection, we consider the most general system made by hybridizing oligomers. Specifically, we consider S different sequences labeled from 1 to S. We call the number of oligomers with the sequence i present in our system N_i. Therefore, $N = \sum_{i=1}^{S} N_i$ is the total number of oligomers in the mixture. Hence, the number concentration of the ith sequence is $c_i = N_i/V$, whereas again the total concentration of oligomers is $c = N/V$.

We consider that, in principle, any duplex formation (complementary or not) may occur. In particular, we call the global free-energy change ΔG_{ij}, including the entropic contribution stemming from the corresponding $\tilde{V}$, in a duplex formation comprising oligomers with sequences i and j. Consistently, we can define the Boltzmann factors

$$\zeta_{ij} = [c] \exp(-\beta \Delta G_{ij}), \tag{9}$$

analogous to Eq. (4). Of course, in a real mixture, some attachments may be very unlikely, which means that such duplex has a very high value of ΔG or, equivalently, a very low value of ζ.

In this general case, the hybridization state is not characterized just by one number. On the contrary, we need a vector $\vec{n}$ containing the number of all formed duplexes n_{ij} comprising oligomers with sequences i and j in our system. Note that $n_{ij} = n_{ji}$ by definition. In spite of this dimensional difference, the probability of having a certain hybridization state $p(\vec{n})$ is provided by the canonical distribution

$$p(\vec{n}) = \frac{1}{Z} f(\vec{n}) \prod_{i=1}^{S} \prod_{j=i}^{S} \left(\frac{\zeta_{ij}}{N} \right)^{n_{ij}}, \tag{10}$$

with the partition function

$$Z = \sum_{\vec{n}} f(\vec{n}) \prod_{i=1}^{S} \prod_{j=i}^{S} \left(\frac{\zeta_{ij}}{N} \right)^{n_{ij}}. \tag{11}$$

On this occasion, the sum over $\vec{n}$ is defined over all possible values of n_{ij} that are integers and holding the set of inequalities $0 \leqslant 2n_{ii} + \sum_{j \neq i} n_{ij} \leqslant N_i, \forall i$. Following the same reasoning that leaded to Eq. (7), the degeneracy $f(\vec{n})$ can be explicitly written as

$$f(\vec{n}) = \prod_{i=1}^{S} \frac{N_i!}{2^{n_{ii}} (N_i - 2n_{ii} - \sum_{j \neq i} n_{ij})!} \prod_{j=i}^{S} \frac{1}{n_{ij}!}. \tag{12}$$

Again, the partition function contains all the relevant statistical information from which we can profit to derive expected values. In particular, the average number of duplexes comprising oligomers with sequences i and j is given by

$$\langle n_{ij} \rangle = \zeta_{ij} \frac{\partial \ln Z}{\partial \zeta_{ij}}. \tag{13}$$

Note that all the theory converges consistently to the one developed in the previous subsection for self-complementary mixtures, when considering $S = 1$ and with the identification $n_{11} = n$.

We expect that a high level of hetereogenity in the mixture will lead to greater difficulties in the computation of the partition function and the successive physical quantities of interest. Remarkably, beside the self-complementary situation with $S = 1$, there is another case of particular interest that can be calculated for $S = 2$: when two sequences are complementary. In this instance, in fact, the chances of having an energetically convenient hybridization between two identical oligomers is very unlikely and thus we can consider $\zeta_{11} = \zeta_{22} = 0$. Therefore, the only possible attachment occurs between complementary oligomers and the hybridization vector becomes a scalar $n_{12} \equiv n$. In this case, we thus re-obtain the one-dimensional theory given by Eq. (6), but with the degeneracy being

$$f(n) = \frac{N_1!N_2!}{(N_1 - n)!(N_2 - n)!n!}. \tag{14}$$

Of course, here we do not find any symmetric factor 2^n, implying an entropy reduction.

III. RESULTS

From now on, we will focus on the two cases introduced above. Namely, the self-complementary (SC) system and the complementary couple (CC), neglecting self-hybridization in the latter. First, let us define the melting curve as the fraction of free oligomers corresponding to the less populated sequence. Therefore, the melting curve is normalized with the maximum number of duplexes. Taking into account that the maximum number of duplexes for the SC or CC cases are $N/2$ (we assume an even N for simplicity) or $N_m = \min(N_1, N_2)$, respectively, we get

$$M_c^{(\mathrm{SC})} = 1 - \frac{2\langle n\rangle}{N}, \quad M_c^{(\mathrm{CC})} = 1 - \frac{\langle n\rangle}{N_m}, \tag{15}$$

where we have introduced explicitly in the notation if we are referring to SC or to CC. Hence, once the system is defined, that is, the oligomer concentration c and the enthalpic ΔH and entropic ΔS changes of duplex formation are known, the melting curve M_c is just a function of the temperature. On the one hand, for very low temperatures the state with minimum enthalpy, which corresponds with the maximum number of duplexes and then $M_c \sim 0$, is promoted. On the other hand, entropy is favored for high temperatures and high values of the melting curve $M_c \sim 1$ are expected.

In the following, we compute the melting curve for our two SC and CC mixtures of oligomers. We provide the results both exactly (for finite systems) and in the thermodynamic limit.

A. Exact results for finite systems

Herein, we compute exactly the partition function and the melting curve for SC and CC systems. In order to obtain the partition function, we just need to carry out the sum in Eq. (6) with the corresponding degeneracy [either Eq. (7) for the SC case or Eq. (14) for the CC case]. Taking into account that the upper bounds of the sum are, respectively $N/2$ and $N_m = \min(N_1, N_2)$, and defining conveniently $N_M = \max(N_1, N_2)$, we get

$$Z^{(\mathrm{SC})} = \left(-\frac{2\zeta}{N}\right)^{\frac{N}{2}} U\left(-\frac{N}{2}, \frac{1}{2}, -\frac{N}{2\zeta}\right), \tag{16a}$$

$$Z^{(\mathrm{CC})} = \left(-\frac{\zeta}{N_m + N_M}\right)^{N_m} \times U\left(-N_m, 1 - N_m + N_M, -\frac{N_m + N_M}{\zeta}\right), \tag{16b}$$

where $U(a, b, x)$ is the confluent hypergeometric function of the second kind also known as Tricomi's function [41]. Performing the derivative in Eq. (8) and introducing it in Eq. (15) yields

$$M_c^{(\mathrm{SC})} = -\frac{NU\left(1 - \frac{N}{2}, \frac{3}{2}, -\frac{N}{2\zeta}\right)}{2\zeta U\left(-\frac{N}{2}, \frac{1}{2}, -\frac{N}{2\zeta}\right)}, \tag{17a}$$

$$M_c^{(\mathrm{CC})} = -\frac{(N_m + N_M)}{\zeta} \times \frac{U\left(1 - N_m, 2 - N_m + N_M, -\frac{N_m+N_M}{\zeta}\right)}{U\left(-N_m, 1 - N_m + N_M, -\frac{N_m+N_M}{\zeta}\right)}. \tag{17b}$$

Remarkably, the expressions above are exact. The prediction should be valid even when the mixture has very few component. Of course, real experimental conditions imply $N \gg 1$. Moreover, we expect from the physical point of view that in such a limit, at fixed number concentration $c = N/V$, N does not play any significant role as shown in the first three panels of Fig. 1.

B. Thermodynamic limit

Since reasonable experimental conditions assure that $N \gg 1$, we focus herein on studying in detail this limit. Furthermore, depending on the arguments, hypergeometric functions are not always easy to compute, which makes even more evident the appropriateness of a more illuminating approach. We consider the thermodynamic limit, which entails the limit $N \to \infty$ but with the concentration $c = N/V$ remaining fixed.

Since performing asymptotic study of the ratios of hypergeometric functions in Eq. (17) is not an easy task, we approach the limit in an alternative way. Specifically, we use Grassmann variables [42] that allow us to derive the partition function in the thermodynamic limit through saddle-point integration [43] after a Hubbard-Stratonovich transformation [44]. Doing so, we are able to compute the melting curve in the thermodynamic limit for both systems of interest, resulting in

$$M_c^{(\mathrm{SC})} = \frac{2}{1 + \sqrt{1 + 4\zeta}}, \tag{18a}$$

$$M_c^{(\mathrm{CC})} = \frac{2}{1 + (1 - 2\phi_m)\zeta + \sqrt{1 + 2\zeta + (1 - 2\phi_m)^2\zeta^2}}, \tag{18b}$$

where we have introduced the fraction of the less concentrated sequence $\phi_m = N_m/N$. The details of the calculations are provided in the Appendix. The convergence of the melting curve to the predicted thermodynamic limit is shown in the three first panels of Fig. 1.

C. Melting temperature

In the context of DNA hybridization and melting, the most analyzed quantity is the melting temperature T_m. The melting temperature is simply defined as the temperature in which half of duplexes are formed. According to our definition of

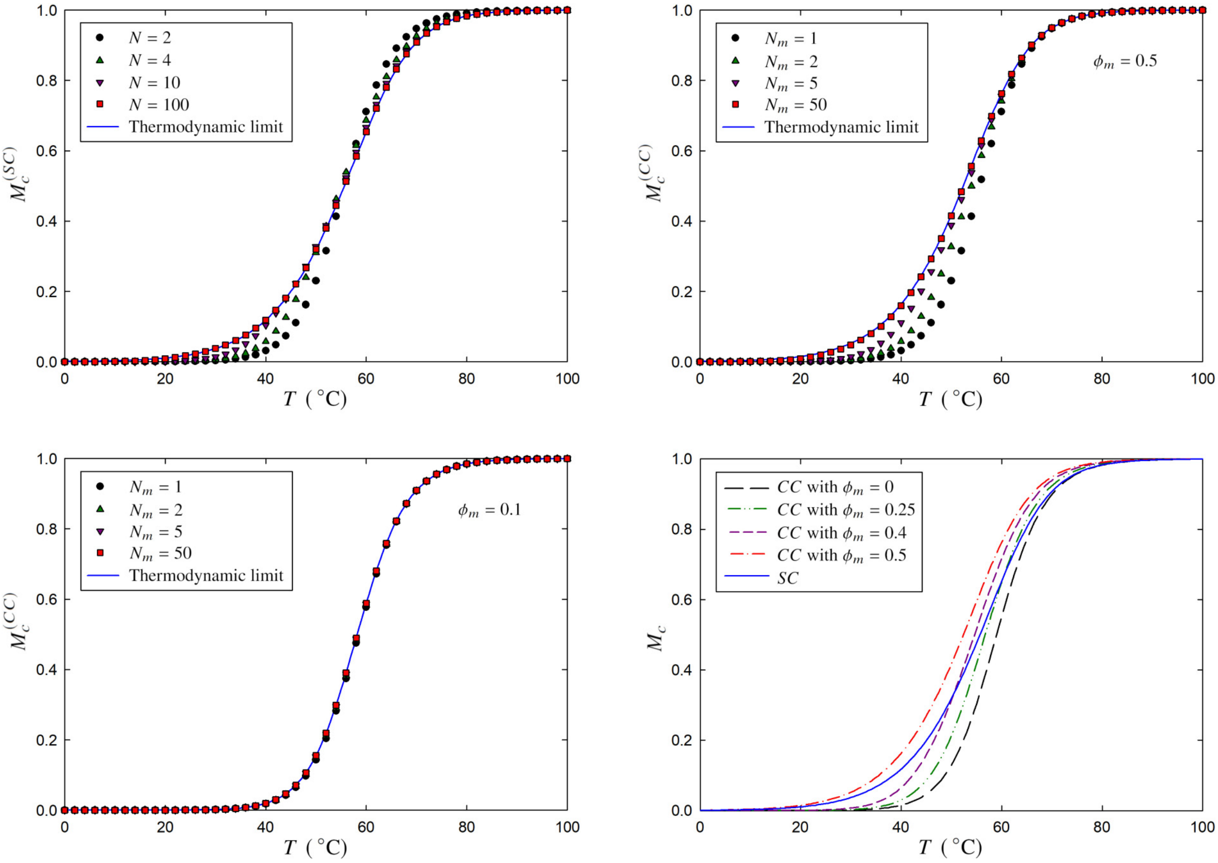

FIG. 1. Melting curve for different DNA mixtures. The convergence of the melting curves for finite systems to the predicted thermodynamic limit when the number of constituents is increased is shown in a SC system (top left), a balanced CC system with $\phi_m = 0.5$ (top right), and an unbalanced CC system with $\phi_m = 0.1$ (bottom left). Bottom right panel presents a comparison of thermodynamic melting curves for CC systems with different values of ϕ_m and a SC system. For the numeric evaluation, we have assumed $\Delta H = -45\ \mathrm{kcal\,mol^{-1}}$, $\Delta S = -120\ \mathrm{cal\,mol^{-1}\,K^{-1}}$, and a concentration of $c = 0.4$ mmol/l.

melting curve, the melting temperature is obtained when we equal $M_c = 0.5$, resulting in

$$T_m^{(\mathrm{SC})} = \frac{\Delta H}{k_B \ln\left(\frac{[c]}{2}\right) + \Delta S}, \tag{19a}$$

$$T_m^{(\mathrm{CC})} = \frac{\Delta H}{k_B \ln\left(\frac{2-3\phi_m}{2}[c]\right) + \Delta S}. \tag{19b}$$

IV. DISCUSSION

Now we highlight the most significant facts that stem from the results derived in the previous section. In Fig. 1, we have illustrated several melting curves for both, SC and CC, cases either in systems with finite size or in the thermodynamic limit. Specifically, we can observe how the convergence to the thermodynamic limit, when increasing the number of constituents, is perfect in the first three panels. Remarkably, as seen in the bottom left panel of Fig. 1, the effect of the finite size almost disappears for very unbalanced CC systems, that is, with low values of ϕ_m. This is in fact intuitive since, in this limit, the less concentrated sequence can be considered as in contact with a bath of strands of the more concentrated sequence. Therefore, the finite size becomes irrelevant since the population of the more concentrated sequence takes relatively high values, legitimating the thermodynamic result, even for low values of N_m. In other words, if we consider a very unbalanced case, that is, $\phi \ll 0.5$, although N_m could be low, N will take high values. Finally, in the bottom right panel, we have plotted several melting curves in the thermodynamic limit of CC systems for different fractions of the less concentrated sequence ϕ_m, as well as the thermodynamic limit of an equivalent SC system. For all the numerical evaluations of the melting curves, we have assumed $\Delta H = -45\ \mathrm{kcal\,mol^{-1}}$, $\Delta S = -120\ \mathrm{cal\,mol^{-1}\,K^{-1}}$ and a total concentration of $c = 0.4$ mmol/l. The thermodynamic parameters considered are consistent with the order of magnitude of complementary sequences around a length of six bases at a reference salt concentration of 1 M [34].

A. Exact results for finite systems

Our exact theory culminating in Eq. (17) provides an analytical prediction for the melting curve in systems with finite size either SC or CC. Considering equivalent free energies and concentrations, the SC system will be more stable than the balanced CC case. In a SC system, all components are available for hybridization, whereas the identity of the oligomer in the CC system is important. Nevertheless, there is a curious unique situation where both of them converge to the same phenomenology. This is when we compare a SC system with

$N = 2$ and a CC system with $N_1 = N_2 = 1$. Under this assumption, the partition function in both cases becomes simply

$$Z = 1 + \zeta, \tag{20}$$

thus obtaining the melting curve

$$M_c = \frac{1}{1+\zeta}. \tag{21}$$

Our theory, as it must be, reflects this consistency. In particular, the melting curves plotted with black circles in the top panels of Fig. 1 are completely identical.

For the best of our knowledge, exact results beyond the thermodynamic limit—i.e., for low number of components—derived in the context of DNA melting are quite scarce [45,46], and usually in the context of molecular simulations, since typical real experiments belongs to the thermodynamic limit. Nevertheless, as stated in the Introduction, the problem of DNA hybridization resembles the (dis)association problem in the field of chemical reactions. Specifically, the SC and CC case analyzed in this work have a direct parallelism with the formation of diatomic molecules either homonuclear or heteronuclear respectively. In spite of a different approach, based on computing the stationary solution of the master equation describing the dynamics of the reaction, the results derived by McQuarry [39] in the context of chemical reactions are fully consistent with ours. The reason behind this consistency is that, therein, they also use our ζ, which only contains information about the interaction pairing and not the possible symmetry of the pairing objects, as the input variable characterizing the interaction.

B. Thermodynamic limit

The convergence of the exact result for increasing values of N to the derived thermodynamic limit shown in the first three panels of Fig. 1 represents a direct validation of our asymptotic calculations. On the one hand, the result for the melting curve in the case of CC systems matches with those provided in the literature. Specifically, most of the time, one finds the results derived in the particular case of a balanced CC system, with $\phi_m = 0.5$. Our result goes beyond the balance case of CC mixtures, explaining the role of the relative concentration through the fraction ϕ_m. This is especially clear in the bottom right panel of Fig. 1, where we have displayed the melting curves for the CC system using different values of ϕ_m as well as we compare with the melting curve for the SC system. Therein, we observe, as stated above, that the SC system is more stable than the balanced CC case, that is, the melting curve for the first remains below the one for the latter. Furthermore, we can see how the melting curve for $\phi_m = 0$ acts as a lower bound for the SC case. Note that when we write $\phi_m = 0$, we are considering the limit of $\phi_m \to 0$ after taking the thermodynamic limit. Indeed, in this extremely unbalanced case there is a bath of available strands to be paired. On the other hand, our result for the melting curve for the SC case differs at first glance from those traditionally found in the literature, for instance in Eq. (15 b) from Ref. [36]. Therein, one finds a factor 8 accompanying ζ in the denominator instead of the factor 4 written in Eq. (18).

Such apparent mismatch is quite curious. On the one hand, our approach (with the 4) has shown to be correct in the asymptotic limit of the exact results, which are consistent with the theory developed for chemical reactions. On the other hand, the traditional melting curve in the literature (with the 8) has been used intensively explaining and fitting data from real experiments. This leads to the question: Why such a difference?

The answer is that this mismatch is not real, but the problem is into the notation. In particular in the definition of ΔS. For us ΔS carries information of the pairing interaction, whereas in the traditional approach, for self-complementary sequences, it additionally includes the entropy reduction coming from the symmetry. We may write $\Delta S_{\text{trad}}^{(\text{SC})} = \Delta S + \Delta S_{\text{sym}}$. In our view, it is more illuminating using equivalent definitions of entropy for pairing strands, either SC or CC. Doing so, we carry the information just from the pure interaction between the two paired objects, without looking into the detail of possible symmetries between these two objects.

Therefore the two approaches are eventually fully compatible, and our derivation helps to point out clearly the origin of the *symmetry correction* systematically applied to self-complementary sequences [27,29,30,33,34,36]. As highlighted in the introduction, this correction ΔS_{sym} is such that $\exp(\Delta S_{\text{sym}}/k_B) \simeq 0.5$. Therefore, if one takes out the correction separately and simplifies the expression, the conventional factor 8 becomes the factor 4 that we have derived and that stems directly from the Boltzmann entropy. Note that, in the conventional approach, this ΔS_{sym} is actually an ad hoc fit parameter that does not need to meet any specific value, potentially affected by other influences. In fact, this has been recently recalculated for DNA [47,48]. Nevertheless, these other influences could be always codified in different values for the initialization entropy for either CC or SC duplex, without any significant change in our approach.

C. Reaction equations approach

The traditional result for the SC case has been usually derived from equilibrium constant arguments. Herein, we give a brief alternative approach which highlight where the mismatch emerges when using the two possible definitions of ΔS. Specifically, we consider the reaction

$$2A \underset{k_b}{\overset{k_f}{\rightleftharpoons}} A_2, \tag{22}$$

which represents the (dis)association of homonuclear diatomic molecules, which also describes the hybridization of self-complementary systems. We write down the macroscopic evolution equation, that is, neglecting fluctuations, for the number of free atoms N_a,

$$\frac{dN_a}{dt} = -k_f \frac{N_a(N_a-1)}{V} + k_b(N - N_a). \tag{23}$$

In the equation above, accompanying the forward rate k_f, we find the number of possible couples in the system normalized by the volume which is $N_a(N_a-1)/2V$ that multiplies the number of atoms that disappear, that is, 2. In turn, along with the backward rate k_b, we have the number of duplexes

$(N - N_a)/2$ that multiplies the number of appearing atoms, which is again 2.

The equilibrium constant, defined as the ratio of the forward rate, yields $K = k_f/k_b = \exp(-\beta\Delta G)$. In this approach, we recall that ΔG has no symmetric consideration. Introducing this relation, the definition of the melting curve $M_c = N_a/N$ and the stationary solution of Eq. (23), then the relation $\zeta = (1 - M_c)/M_c^2$ is obtained, where we have made use of the definition of ζ in Eq. (4) and neglected terms of order N^{-1}. Solving this equation for M_c leads straightforwardly to our result for the SC melting curve in Eq. (18).

On the contrary, in the traditional derivation, the equilibrium constant is directly computed as the fraction between the concentration of duplexes and the squared concentration of the single strands. Therein, when writing $\exp(-\beta\Delta G)$, the ΔS included in ΔG contains the symmetric correction, thus again leading to the same quantitative result obtained through our approach.

An exploitable advantage of the proposed statistical mechanics formulation is that the free energy has full physical meaning, measuring the interaction between the strands. Therefore one could think to actually measures such term from single molecule experiments [49,50], with independence on the CC or SC character of the duplex. This allows a direct comparison between equivalent CC and SC couples. In other words, a finite system containing only two strands with a given free-energy difference behaves in the same way, no matter if we assumed they are CC or SC, whereas this comparison is not so straightforward in the conventional approach because of the offset introduced by the symmetric contribution.

V. CONCLUSION

In this work, we have put forward the fundamentals of statistical physics to the service of the problem of DNA melting. We have provided an equilibrium description of hybridization and melting for nucleic acids, deriving the partition function in any arbitrary mixture made by DNA.

We have then focused on two experimentally relevant systems, made by either SC sequences or different CC of strands. In such cases, the exact melting curves have been derived and we have shown that our results agree with the known expressions in the context of chemical reactions. Remarkably, the proposed statistical mechanics formulation also provides a microscopic justification for the extra entropic contribution that in classic hybridization modeling was required to correctly describe within the same framework the melting of sequences either self-complementary or not. In our view, dealing explicitly with the symmetric correction through the degeneracy of the system, extracting then this contribution from the ΔS that, in our convention, only carries information about the pairing interaction regardless the nature of the pairing objects, allows to compare easier the physics in equivalent duplexes of CC or SC. Of course, both pathways, the conventional one or the one presented here lead to the same quantitative results.

Beyond our exact result, we have derived the melting curve of the systems of interest in the thermodynamic limit, which is the meaningful situation in real experimental conditions. As expected, the convergence to the thermodynamic limit when considering the exact result with increasing number of constituents is excellent.

Although the computation of the melting curve for the SC and CC cases has been made by considering only complementary sequences (and thus just one single way for the attachment between pairs of DNA strands), our model can be generalized to describe pairs with a preferential ways of attachment, even if not complementary. The lack of full complementarity would lead to lower values of ζ, resulting in a system with lower stability and a melting curve shifted to lower temperatures. The exact calculations for such more heterogeneous cases is left for future investigations. Generalization to formation of multimeric complexes of duplexes [45,51,52] is also an appealing field to apply our approach.

ACKNOWLEDGMENTS

S.S. and C.A.P. acknowledge UNIPD STARS grant BioReact 2018. C.A.P. acknowledges the support provided by Grant No. ANR-18-CE30-0013 from the Agence Nationale de la Recherche. T.B. acknowledges support by PRIN2017 project from Ministero Istruzione, Università e Ricerca [ID 2017Z55KCW]. A.M. acknowledges the support from University of Padova through "Excellence Project 2018" of the Cariparo foundation. S.S. and A.M. acknowledges INFN Lincoln grant.

APPENDIX: THERMODYNAMIC LIMIT

We have relegated to this Appendix the mathematical derivation of the melting curve in the thermodynamic limit in order to preserve the clearance in the presentation of the main text. A possible course of action is applying Stirling's approximation for the factorials appearing within the degeneracy of hybridization states $f(n)$. We present instead an approach that is based on the application of auxiliary variables that verify Grassmann algebra, which are commonly used in the context of path integral representation of fermionic fields [42]. Herein, avoiding further technicalities, it suffices to introduce an auxiliary set of independent random variables θ_i which satisfies the following properties:

$$\langle 1\rangle = \langle\theta_i\rangle = 1, \qquad \langle\theta_i^{k\geqslant 2}\rangle = 0,$$
$$\langle f_1(\theta_i) f_2 f(\theta_j)\rangle = \langle f_1(\theta_i)\rangle\langle f_2 f(\theta_j)\rangle \quad (\text{for } j \neq i), \tag{A1}$$

where $\langle\cdot\rangle$ stands for a trace operation.

1. Melting curve for the SC system

We start by writing down the partition function of a system of N self-complementary oligomers, in a smart way, using the properties defined in Eq. (A1) of the set of auxiliary variables θ_i, $i = 1, \ldots, N$,

$$Z = \left\langle \prod_{i=1}^{N}\prod_{j<i}\left(1 + \theta_i\theta_j\frac{\zeta}{N}\right)\right\rangle. \tag{A2}$$

Note that, using the auxiliary variables, we have avoided the explicit writing of the degeneracies of hybridization states $f(n)$ as in Eq. (6) of the main text. Manipulating Eq. (A2) and taking into account the properties of the algebra defined in Eq. (A1), we can reach the expression

$$Z = \left\langle \exp\left[\frac{\zeta}{2N}\left(\sum_{i=1}^{N}\theta_i\right)^2\right]\right\rangle. \tag{A3}$$

We apply now the Hubbard-Stratonovich transformation [44]

$$Z = \left\langle \sqrt{\frac{N}{2\pi\zeta}}\int_{-\infty}^{\infty} dy \exp\left(-N\frac{y^2}{2\zeta} + y\sum_{i=1}^{N}\theta_i\right)\right\rangle. \tag{A4}$$

Then, it is possible to carry out explicitly the trace operation, yielding

$$Z = \sqrt{\frac{N}{2\pi\zeta}}\int_{-\infty}^{\infty} dy \exp\left[N\varphi(y)\right], \tag{A5}$$

with

$$\varphi(y) = -\frac{y^2}{2\zeta} + \ln(1+y). \tag{A6}$$

Importantly, the way in which Eq. (A5) is written allows for a direct application of the method of steepest descent or saddle-point method for approximating the integral [43] which yields, in the asymptotic limit $N \to \infty$,

$$\frac{\ln Z}{N} = \varphi(y^*), \quad \text{with} \quad \left.\frac{\partial\varphi(y)}{\partial y}\right|_{y=y^*} = 0, \tag{A7}$$

where the problem reduces to find the stationary value of the function in the exponent of the integral. Once the maximum is found we can substitute the partition function in Eq. (8), and subsequently in (15), of the main text in order to finally obtain the melting curve

$$M_c^{(\mathrm{SC})} = 1 - 2\zeta\frac{\partial\varphi(y^*)}{\partial\zeta} = \frac{2}{1+\sqrt{1+4\zeta}}, \tag{A8}$$

which is the one presented in Eq. (18) of the main text.

2. Melting curve for the CC system

When we consider a system made by two (different) complementary sequences, the calculations to derive the thermodynamic limit become a little bit more involved. Nevertheless, the line of reasoning and techniques, as seen below, are pretty similar to those used in the previous case. We recall that the total number of oligomers $N = N_m + N_M$ is the sum of the number N_m of of strands of the less concentrated sequence and the number N_M of strands of the more concentrated sequence. Therefore, it is handy to define two sets of auxiliary independent random variables θ_i, $i = 1, \ldots, N_m$ and η_j, $j = 1, \ldots, N_M$. Remarkably, all of them fulfill the properties of the algebra defined in Eq. (A1), the question of labeling with different letters is just a matter of convenience for the notation. We start again by making use of the auxiliary variables to write down the partition function

$$Z = \left\langle \prod_{i=1}^{N_m}\prod_{j=0}^{N_M}\left(1 + \theta_i\eta_j\frac{\zeta}{N}\right)\right\rangle. \tag{A9}$$

Again, we have avoided the explicit writing of the degeneracies of hybridization states $f(n)$. Manipulating Eq. (A9) and taking into account the properties of the algebra defined in Eq. (A1), we reach the expression

$$Z = \left\langle \exp\left[\frac{\zeta}{2N}\left(\sum_{i=1}^{N_m}\theta_i + \sum_{j=1}^{N_M}\eta_i\right)^2 - \frac{\zeta}{2N}\left(\sum_{i=1}^{N_m}\theta_i\right)^2 - \frac{\zeta}{2N}\left(\sum_{j=1}^{N_M}\eta_i\right)^2\right]\right\rangle. \tag{A10}$$

The application of the Hubbard-Stratonovich transformation leads us to the integral form

$$\begin{aligned} Z = \Bigg\langle \left(\frac{N}{2\pi\zeta}\right)^{3/2} &\int_{-\infty}^{\infty} dy_1 \int_{-\infty}^{\infty} dy_2 \int_{-\infty}^{\infty} dy_3 \\ &\times \exp\left[-N\frac{y_1^2}{2\zeta} + y_1\left(\sum_{i=1}^{N_m}\theta_i + \sum_{j=1}^{N_M}\eta_j\right) - N\frac{y_2^2}{2\zeta}\right. \\ &\left. + y_2 i\sum_{i=1}^{N_m}\theta_i - N\frac{y_3^2}{2\zeta} + y_3 i\sum_{j=1}^{N_M}\eta_j\right]\Bigg\rangle. \end{aligned} \tag{A11}$$

The trace can be carry out then and we get

$$\begin{aligned} Z = &\left(\frac{N}{2\pi\zeta}\right)^{3/2} \\ &\times \int_{-\infty}^{\infty} dy_1 \int_{-\infty}^{\infty} dy_2 \int_{-\infty}^{\infty} dy_3 \exp\left[N_m\varphi(y_1, y_2, y_3)\right] \end{aligned} \tag{A12}$$

with

$$\begin{aligned} \varphi(y_1, y_2, y_3) = &-\frac{y_1^2 + y_2^2 + y_3^2}{2\phi_m\zeta} + \ln(1 + y_1 + iy_2) \\ &+ \left(\frac{1}{\phi_m} - 1\right)\ln(1 + y_1 + iy_3). \end{aligned} \tag{A13}$$

We have introduced in the notation the fraction of the less concentrated sequence $\phi_m = N_m/N \in [0, 0.5]$. Note that, when we take the thermodynamic limit, both N_m and N_M go to infinity but the ratios ϕ_m and $\phi_M = 1 - \phi_m$ remain exactly the same. In this occasion, since we have a complex function, it is convenient to deform the contour integration in such a way that the stationary value of φ occurs along the contour. The saddle-point integration has to be performed carefully, taking into account that the search of the stationary value is over the three complex variables. Therefore, we obtain

$$\frac{\ln Z}{N_m} = \varphi(y_1^*, y_2^*, y_3^*). \tag{A14}$$

In order to look for the maximum one enforces the vanishing of the derivatives of $\varphi(y_1, y_2, y_3)$ with respect to the real and imaginary parts of the three variables. There is more than one solution, but only one gives a physical (real and positive) melting curve, which corresponds to the solution with a Hessian matrix negative definite,

$$\begin{aligned} M_c^{(\mathrm{CC})} &= 1 - \zeta\frac{\partial\varphi(y_1^*, y_2^*, y_3^*)}{\partial\zeta} \\ &= \frac{2}{1 + (1-2\phi_m)\zeta + \sqrt{1 + 2\zeta + (1-2\phi_m)^2\zeta^2}}, \end{aligned} \tag{A15}$$

which is the final result provided in Eq. (18) of the main text.

[1] J. Watson and F. Crick, Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid, Nature **171**, 737 (1953).

[2] K. M. Ririe, R. P. Rasmussen, and C. T. Wittwer, Product differentiation by analysis of DNA melting curves during the polymerase chain reaction, Anal. Biochem. **245**, 154 (1997).

[3] G. H. Reed, J. O. Kent, and C. T. Wittwer, High-resolution DNA melting analysis for simple and efficient molecular diagnostics, Pharmacogenomics **8**, 597 (2007).

[4] S. Cuesta-Lopez, D. Angelov, and M. Peyrard, Adding a new dimension to DNA melting curves, Europhys. Lett. **87**, 48009 (2009).

[5] H. Von Keyserling, T. Bergmann, M. Wiesel, and A. M. Kaufmann, The use of melting curves as a novel approach for validation of real-time PCR instruments, BioTechnology **51**, 179 (2011).

[6] J. A. Balog, L. Z. Fehér, and L. G. Puskás, Decoding DNA labels by melting curve analysis using real-time PCR, BioTechnology **63**, 261 (2017).

[7] M. Mandel and J. Marmur, [109] use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA, in *Nucleic Acids, Part B*, Methods in Enzymology, Vol. 12 (Academic Press, New York, 1968), pp. 195–206.

[8] R. Wells, J. Larson, R. Grant, B. Shortle, and C. Cantor, Physicochemical studies on polydeoxyribonucleotides containing defined repeating nucleotide sequences, J. Mol. Biol. **54**, 465 (1970).

[9] R. M. Wartell and A. S. Benight, Thermal denaturation of DNA molecules: A comparison of theory with experiment, Phys. Rep. **126**, 67 (1985).

[10] J.-L. Mergny and L. Lacroix, Analysis of thermal melting curves, Oligonucleotides **13**, 515 (2003).

[11] R. Owczarzy, Melting temperatures of nucleic acids: Discrepancies in analysis, Biophys. Chem. **117**, 207 (2005).

[12] S. A. Rice and A. Wada, On a model of the helix’coil transition in macromolecules. ii, J. Chem. Phys. **29**, 233 (1958).

[13] J. H. Gibbs and E. A. DiMarzio, Statistical mechanics of helix’coil transitions in biological macromolecules, J. Chem. Phys. **30**, 271 (1959).

[14] J. Applequist and V. Damle, Theory of the effects of concentration and chain length on helix-coil equilibria in two stranded nucleic acids, J. Chem. Phys. **39**, 2719 (1963).

[15] D. Poland and H. A. Scheraga, Occurrence of a phase transition in nucleic acid models, J. Chem. Phys. **45**, 1464 (1966).

[16] C. Kittel, Phase transition of a molecular zipper, Am. J. Phys. **37**, 917 (1969).

[17] M. Fixman and J. J. Freire, Theory of DNA melting curves, Biopolymers **16**, 2693 (1977).

[18] M. Y. Azbel, Phase transitions in DNA, Phys. Rev. A **20**, 1671 (1979).

[19] T. Dauxois, M. Peyrard, and A. R. Bishop, Dynamics and thermodynamics of a nonlinear model for DNA denaturation, Phys. Rev. E **47**, 684 (1993).

[20] V. Ivanov, Y. Zeng, and G. Zocchi, Statistical mechanics of base stacking and pairing in DNA melting, Phys. Rev. E **70**, 051907 (2004).

[21] D. Jost and R. Everaers, A unified poland-scheraga model of oligo- and polynucleotide DNA melting: Salt effects and predictive power, Biophys. J. **96**, 1056 (2009).

[22] I. Tinoco, O. C. Uhlenbeck, and M. D. Levine, Estimation of secondary structure in ribonucleic acids, Nature **230**, 362 (1971).

[23] P. N. Borer, B. Dengler, I. Tinoco, and O. C. Uhlenbeck, Stability of ribonucleic acid double-stranded helices, J. Mol. Biol. **86**, 843 (1974).

[24] S. Ghosh, S. Takahashi, T. Endoh, H. Tateishi-Karimata, S. Hazra, and N. Sugimoto, Validation of the nearest-neighbor model for Watson-Crick self-complementary DNA duplexes in molecular crowding condition, Nucleic Acids Res. **47**, 3284 (2019).

[25] L. A. Marky and K. J. Breslauer, Calorimetric determination of base-stacking enthalpies in double-helical DNA molecules, Biopolymers **21**, 2185 (1982).

[26] M. Petersheim and D. H. Turner, Base-stacking and base-pairing contributions to helix stability: Thermodynamics of double-helix formation with ccgg, ccggp, ccggap, accggp, ccggup, and accggup, Biochemistry **22**, 256 (1983).

[27] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky, Predicting DNA duplex stability from the base sequence, Proc. Natl. Acad. Sci. USA **83**, 3746 (1986).

[28] L. A. Marky and K. J. Breslauer, Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves, Biopolymers **26**, 1601 (1987).

[29] M. J. Serra and D. H. Turner, [11] predicting thermodynamic properties of RNA, in *Energetics of Biological Macromolecules*, Methods in Enzymology, Vol. 259 (Academic Press, New York, 1995), pp. 242–261.

[30] N. Sugimoto, S.-i. Nakano, M. Yoneyama, and K.-i. Honda, Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes, Nucleic Acids Res. **24**, 4501 (1996).

[31] D. M. Gray, Derivation of nearest-neighbor properties from data on nucleic acid oligomers. i. Simple sets of independent sequences and the influence of absent nearest neighbors, Biopolymers **42**, 783 (1997).

[32] H. T. Allawi and J. SantaLucia, Thermodynamics and nmr of internal gt mismatches in DNA, Biochemistry **36**, 10581 (1997).

[33] J. SantaLucia, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA **95**, 1460 (1998).

[34] J. SantaLucia and D. Hicks, The thermodynamics of DNA structural motifs, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).

[35] A. S. Benight and R. M. Wartell, Influence of base-pair changes and cooperativity parameters on the melting curves of short DNAS, Biopolymers **22**, 1409 (1983).

[36] R. Owczarzy, P. M. Vallone, F. J. Gallo, T. M. Paner, M. J. Lane, and A. S. Benight, Predicting sequence-dependent melting stability of short duplex DNA oligomers, Biopolymers **44**, 217 (1997).

[37] R. A. Dimitrov and M. Zuker, Prediction of hybridization and melting for double-stranded nucleic acids, Biophys. J. **87**, 215 (2004).

[38] T. Xia, J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner, Thermodynamic parameters for an expanded nearest-neighbor model for formation of rna duplexes with watsoncrick base pairs, Biochemistry **37**, 14719 (1998).

[39] D. A. McQuarrie, Stochastic approach to chemical kinetics, J. Appl. Probab. **4**, 413 (1967).

[40] B. J. McCoy and R. G. Carbonell, Master equation theory for steady state chemical reactions: Dissociation of diatomic molecules in gases, J. Chem. Phys. **66**, 4564 (1977).

[41] N. Lebedev and R. Silverman, *Special Functions and Their Applications*, Dover Books on Mathematics (Dover, London, 1972).

[42] F. A. Berezin, N. Mugibayashi, and A. Jeffrey, *The Method of Second Quantization*, Pure and Applied Physics: A Series of Monographs and Textbooks, Vol. 24 (Academic Press, New York, 1966).

[43] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (Springer, New York, 1999).

[44] J. Hubbard, Calculation of Partition Functions, Phys. Rev. Lett. **3**, 77 (1959).

[45] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, Extracting bulk properties of self-assembling systems from small simulations, J. Phys.: Condens. Matter **22**, 104102 (2010).

[46] D. H. De Jong, L. V. Schäfer, A. H. De Vries, S. J. Marrink, H. J. C. Berendsen, and H. Grubmüller, Determining equilibrium constants for dimerization reactions from molecular dynamics simulations, J. Comput. Chem. **32**, 1919 (2011).

[47] G. Weber, Optimization method for obtaining nearest-neighbour DNA entropies and enthalpies directly from melting temperatures, Bioinformatics **31**, 871 (2014).

[48] I. Ferreira, E. A. Jolley, B. M. Znosko, and G. Weber, Replacing salt correction factors with optimized rna nearest-neighbour enthalpy and entropy parameters, Chem. Phys. **521**, 69 (2019).

[49] M. C. Williams, I. Rouzina, and V. A. Bloomfield, Thermodynamics of DNA interactions from single molecule stretching experiments, Acc. Chem. Res. **35**, 159 (2002).

[50] K. R. Chaurasiya, T. Paramanathan, M. J. McCauley, and M. C. Williams, Biophysical characterization of DNA binding from single molecule force measurements, Phys. Life Rev. **7**, 299 (2010).

[51] T. E. Ouldridge, Inferring bulk self-assembly properties from simulations of small systems with multiple constituent species and small systems in the grand canonical ensemble, J. Chem. Phys. **137**, 144105 (2012).

[52] L. Di Michele, S. J. Bachmann, L. Parolini, and B. M. Mognetti, Communication: Free energy of ligand-receptor systems forming multimeric complexes, J. Chem. Phys. **144**, 161104 (2016).