THE JOURNAL OF CHEMICAL PHYSICS **138**, 055102 (2013)

A sequence-dependent rigid-base model of DNA

O. Gonzalez,[1,a)] D. Petkevičiūtė,[2,a)] and J. H. Maddocks[2,b)]
[1]*Department of Mathematics, University of Texas, Austin, Texas 78712, USA*
[2]*Section de Mathématiques, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

(Received 13 August 2012; accepted 9 January 2013; published online 7 February 2013)

A novel hierarchy of coarse-grain, sequence-dependent, rigid-base models of B-form DNA in solution is introduced. The hierarchy depends on both the assumed range of energetic couplings, and the extent of sequence dependence of the model parameters. A significant feature of the models is that they exhibit the phenomenon of *frustration*: each base cannot simultaneously minimize the energy of all of its interactions. As a consequence, an arbitrary DNA oligomer has an intrinsic or pre-existing stress, with the level of this frustration dependent on the particular sequence of the oligomer. Attention is focussed on the particular model in the hierarchy that has nearest-neighbor interactions and dimer sequence dependence of the model parameters. For a Gaussian version of this model, a complete coarse-grain parameter set is estimated. The parameterized model allows, for an oligomer of arbitrary length and sequence, a simple and explicit construction of an approximation to the configuration-space equilibrium probability density function for the oligomer in solution. The training set leading to the coarse-grain parameter set is itself extracted from a recent and extensive database of a large number of independent, atomic-resolution molecular dynamics (MD) simulations of short DNA oligomers immersed in explicit solvent. The Kullback-Leibler divergence between probability density functions is used to make several quantitative assessments of our nearest-neighbor, dimer-dependent model, which is compared against others in the hierarchy to assess various assumptions pertaining both to the locality of the energetic couplings and to the level of sequence dependence of its parameters. It is also compared directly against all-atom MD simulation to assess its predictive capabilities. The results show that the nearest-neighbor, dimer-dependent model can successfully resolve sequence effects both within and between oligomers. For example, due to the presence of frustration, the model can successfully predict the nonlocal changes in the minimum energy configuration of an oligomer that are consequent upon a local change of sequence at the level of a single point mutation. © *2013 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4789411]

I. INTRODUCTION

The sequence-dependent curvature and flexibility of DNA in solution is of both biological and technological importance. On the biological side, these structural properties of DNA are thought to be critical for its packaging into the cell, recognition by other molecules, and conformational changes during biochemical processes. Indeed, there is a general consensus that the specific basepair sequence of DNA in a genome carries not only a genetic code, but also a structural code that is essential for understanding various processes such as site-specific recognition,[1–4] nucleosome positioning,[5–7] and looping.[8–10] Methylation, which plays a central role in both the epigenetic gene regulation in the normal development of higher organisms and in the development of nearly all types of cancer, is also believed to affect the mechanical properties of the DNA duplex.[11,12] On the technological side, DNA has been used as a material for the controlled fabrication and operation of nano-devices. In this context, the propensity of single strands of DNA to self-assemble into a double-helical structure of complementary basepairs is exploited to build pre-specified three-dimensional structures,[13–15] some of which can perform elementary operations and functions.[16–18] The sequence-dependent mechanical properties of the resulting structure offers a rich design landscape which has yet to be fully exploited.

The detailed study of sequence effects on the intrinsic curvatures and flexibilities of DNA requires a suitable model at the appropriate scale. A primary objective of such a model is to describe sequence-dependent variations in the spatial arrangement of the bases or basepairs of an arbitrary oligomer at the scale of tens to hundreds of basepairs. At present, these scales remain prohibitively expensive for all-atom molecular dynamics (MD) models.[19–23] Although the MD simulation of oligomers on relatively short length and time scales is becoming routine and, for example, the microsecond time scale can now be achieved for short oligomers,[24,25] such simulations are not yet able to reach all the length and time scales of interest; moreover, when feasible, they are done on a case-by-case basis for oligomers of a specific sequence and do not provide a practical, sufficiently explicit method for studying sequence effects over a large number of oligomers. Sequence dependence at the scales of interest also falls below the resolution of the classic worm-like chain and homogeneous elastic models,[26–30] and various related models which have

[a)]Joint first author.
[b)]Author to whom correspondence should be addressed. Electronic mail: john.maddocks@epfl.ch.

received considerable attention in the study of DNA supercoiling, looping, and packaging.[27,31–39] A class of models that are ideally suited for the study of sequence effects at the scales and resolutions of interest are rigid-body models in which individual bases or basepairs are treated as independent, interacting rigid bodies.[32,38,40–44] Such models are coarse-grained and hence simpler to understand than those of the atomistic type, but are more detailed and hence better adapted to capture local sequence-dependent features than models of the homogeneous worm-like chain type.

Although rigid-basepair models have been successful at understanding phenomena on relatively long scales, they are based on a highly idealized representation of DNA and are not sufficiently sophisticated to represent the internal, three-dimensional, sequence-dependent, stereochemical relations between individual bases along an oligomer. For example, phenomena, such as DNA melting and self-assembly,[45–55] cannot be captured with a rigid-basepair model. Rather, models developed in connection with these phenomena range from probabilistic type models with little structural detail, to more physically detailed rigid-base type models with atomistic-like potentials between specified sites, with varying levels of sequence dependence. A main goal of such models has been to capture the thermal denaturation and renaturation of the individual strands of DNA as a function of temperature and ionic conditions. In contrast, here we focus on using a rigid-base model to describe the sequence-dependent curvature and flexibility of double-stranded DNA.[41,43,56–59] We restrict attention to duplex DNA within the B-form family under fixed solvent conditions, and seek to develop a model that can accurately and explicitly predict the sequence-dependent variations in the curvatures and flexibilities of DNA in solution as expressed through an equilibrium configuration-space distribution at the scales of interest, namely, up to a few persistence lengths, or oligomers of a few tens to several hundreds of basepairs in length.

Since early studies, a major issue in the coarse-grain modeling of DNA has been the identification of a fundamental parameter set for interactions that is as simple as possible, yet sufficiently accurate to describe sequence effects at the desired length scales.[59–63] Various evidence suggests that the structural properties of an oligomer, specifically its ground-state or minimum energy shape, can have a noticeable nonlocal dependence on sequence up to the level of tetramers and possibly beyond, suggesting that sequence dependence to at least this length should be included in a parameter set.[20,21,64,65] Due to the exponential growth in the number of sequence combinations, the description of interactions with a sequence dependence up to the tetramer level or beyond would seem to require a rather large set of parameters. For example, the numbers of independent monomer, dimer, trimer, tetramer, and pentamer sequence units are, respectively, 2, 10, 32, 136, and 512. The description of energetic interactions with each of these possible sequence dependencies would, therefore, appear to require correspondingly large numbers of independent sets of interaction parameters. However, as is explained further below, one of the attractive features of our rigid-base model is that a parameter set based on only the two smallest of sequence units, namely, monomers and dimers, can predict ground-states of an oligomer whose local intrinsic shape can depend on the tetramer, and beyond, sequence context.

Once a coarse-grain model with a specific form of parameter dependencies has been decided upon, the other recurring and major modeling issue is to decide upon the data that will be used as a training set to estimate the actual values of the model parameters. The data required to reliably and robustly fit the model parameters must be of sufficiently high structural resolution, and must also cover a sufficiently rich set of sequences. At present, high-resolution experimental data such as that from x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy remains rather sparse, certainly if data on all possible tetramer sequences must be observed. However, the field of MD applied to DNA has improved significantly in recent years, and the recent availability of an extensive database of atomic-resolution DNA simulations[19–21] of a set of oligomers with multiple, independent instances of all 136 tetramer sequence units has opened up the possibility of fully parameterizing a variety of coarse-grain models using a training set extracted from fine-grain MD simulations. That is the strategy that we will pursue.

In this article, we introduce a new hierarchy of models for predicting the relative position, orientation, and energetic coupling between every base in an oligomer of double-helical, B-form DNA with arbitrary sequence, in solution under prescribed, standard, environmental conditions. Motivated by previous work,[43] we focus on models of the rigid-base type in which each individual base is treated as an independent rigid body. For any given oligomer, the models deliver an equilibrium or stationary probability distribution on the associated internal configuration space, where the degrees of freedom or observables are the relative displacement and rotation of each rigid base on each strand of the oligomer. Such a distribution is to be interpreted as a marginal distribution of the fine-grain oligomer-solvent system, where the detailed atomistic level descriptions have been averaged out. Specifically, the DNA backbones are not direct observables in our models, which cannot therefore resolve different backbone states. However, there is sufficient structural resolution to capture the probability of differences in bending, twisting, stretching, and shearing of the basepairs along the contour of the double-helix (so that the standard DNA structural parameters of tilt, roll, twist, shift, slide, and rise are observables), as well as the deformation between the bases within each basepair across the double-helix (so that buckle, propeller, opening, shear, stretch, and stagger are also observables). As a consequence, among other things, the spacing of the major and minor grooves of the double-helix can be predicted.

Consistent with equilibrium statistical mechanics, we assume that our coarse-grain stationary probability distribution is of a canonical form, with a density defined in terms of the exponential of a free energy function. For simplicity, and in the first instance, we also assume that the density is Gaussian, with the associated free energy being a general shifted quadratic function of the coarse-grain variables, namely, the relative displacements and rotations between the rigid bases of the oligomer. While the assumption of a quadratic free energy limits the scope of the model to the small-strain regime within the B-form family of oligomer configurations (and, in particular, complete basepair breaking in any of its forms is precluded), there is no *a priori* limitation on the length scale of the oligomer to which the model can be applied, either large or small. And overall large deformations of long oligomers certainly are captured.

The main novelty of the current work lies in the approach to predict the sequence dependence of the free energy function of a given oligomer of arbitrary sequence. We construct an oligomer free energy as a sum over a nested hierarchy of sequence-dependent local energies that describe physically distinct interactions between various groups of proximal bases. Consistent with a nearest-neighbor assumption, we focus attention on the first two members of the hierarchy that describe the local interactions between the two bases in a monomer, and between the four bases in a dimer. Moreover, we characterize these interactions by a set of parameters that depend only on the local monomer and dimer sequence. The free energy of an arbitrary oligomer of any length is then defined by a construction rule in which the local interaction energies are superimposed. We show that a free energy constructed in this way provides a natural model for the intrinsic curvature and flexibility of an oligomer of arbitrary sequence. Indeed, we show that these properties are determined by the local parameters in a nontrivial way through the construction rule. Specifically, our free energy provides a natural model for the intrinsic *frustration* of an oligomer. The frustration, or pre-existing stress, in an oligomer arises from the fact that each base cannot simultaneously minimize all of its local interactions and must instead find a compromise. As a consequence, our locally parameterized model predicts that the intrinsic or ground-state curvature of an oligomer depends nonlocally on its sequence, as has been observed in various detailed MD simulations.[20,21,64,65] The description of such nonlocal behavior using only local parameters is a feature unique to our rigid-base model. It occurs because of the double-chain connectivity (or topology) of a rigid-base model, and cannot arise in nearest-neighbor, rigid-basepair type models with their single-chain connectivity, as have been considered by a number of authors.[59–63]

We develop and implement a method for estimating a complete parameter set for our nearest-neighbor, dimer-dependent model from atomic-resolution MD data. The overall method comprises two main parts. The first ingredient is the estimation of means and covariances from the MD time series, assuming that these time series are themselves stationary. To obtain parameters consistent with the B-form DNA structural family, we follow the treatment in previous work[43] and employ special procedures for excluding structures with broken H-bonds. Broken bonds provide a signal that a structure is defective, for example, it may have frayed ends. The second ingredient is the numerical optimization of various fitting functionals. Specifically, the model parameter set is fit to the estimated means and covariances of a training set of oligomers using a maximum relative entropy approach in which a Kullback-Leibler divergence between the model and estimated probability densities is minimized. Although both densities are assumed to be of a Gaussian form, numerical minimization is nonetheless required due to the structure of the model. As a concrete example, we explicitly implement the proposed parameter estimation method on an extensive database of MD time series produced by a consortium of groups,[19–21] complemented with additional, and compatible, time series data that we simulated to have training set oligomers with a sufficient diversity of sequences at the leading and trailing ends. Both data sets comprise all-atom simulations, with explicit solvent and ions, of over 50 different oligomers in total, where each oligomer was either 12 or 18 basepairs long, with simulation times of 50–200 ns for each oligomer. From this data, we have obtained an initial, best-fit parameter set for double-stranded, B-form DNA under standard environmental conditions.

We then present several quantitative assessments that illustrate various features and limitations of our nearest-neighbor, dimer-dependent model with its current parameter set. This parameterized model is compared against others in the hierarchy to assess various assumptions pertaining to the locality of the energetic couplings and the level of sequence dependence of its parameters, and compared against direct, all-atom MD simulation to assess its predictive capabilities. For each of several example oligomers, the model predicts a Gaussian probability density on the high-dimensional internal configuration space of the oligomer, that is in good agreement with statistics garnered from direct MD simulation. In addition to direct comparisons of means, covariances, and various marginals, we also adopt the Kullback-Leibler divergence to compare overall differences between probability density functions. The results indicate that the nearest-neighbor, dimer-dependent model with the current parameter set is less satisfactory at modeling the ends of an oligomer than its interior. One possible explanation for this observation is that the current model of the ends of an oligomer is a simple extension of the model in the interior, with no additional assumptions or parameters being introduced to capture any exceptional end effects. We also remark that the assumed Gaussian form of the model precludes it from capturing any type of bi-modal or fraying behavior of the bases of an oligomer. Because some instances of such behavior are evident in the MD data, we *a priori* expect some discrepancies in our comparisons. Nevertheless, within these limitations, our examples show that the model can quantitatively predict the equilibrium statistical properties of many oligomers rather well. Specifically, the model can successfully resolve sequence effects both within and between oligomers, and can successfully predict properties such as the nonlocal effects of single point mutations in the sequence. Compared to more sophisticated models in the hierarchy with larger parameter sets, the nearest-neighbor, dimer-dependent model represents a practical compromise between complexity of the model and accuracy of the model predictions. We also expect the accuracy of the nearest-neighbor model to improve as more MD data becomes available for use as a training set, so that the model parameter set can be further refined.

The presentation is structured as follows. In Sec. II, we establish notation and outline the internal coordinates and quadratic free energy for a general rigid-base model of DNA. In Sec. III, we introduce the concept of hierarchical local energies and define two different nearest-neighbor, sequence-dependent models for the free energy function and describe their properties. In Sec. IV, we describe the MD simulations that were used in our study and a procedure for the estimation, from the observed time series, of the coarse-grain equilibrium probability densities that form our training set. In Sec. V, we describe a method for fitting model predictions of probability densities to the observed training set using a maximum relative entropy approach, and thereby obtain a first best-fit parameter set. In Secs. VI and VII, we present various quantitative assessments that illustrate the effectiveness of the nearest-neighbor, dimer-dependent model with our current parameterization. Finally in Sec. VIII, we summarize our results and conclusions. The supplementary material[66] provides an extensive discussion of the necessary background material that is exploited in the main text, along with further comparisons of predicted and observed quantities for various oligomers.

II. PRELIMINARIES

A. Configuration coordinates

We consider right-handed, double-helical DNA in which bases T, A, C, and G are attached to two, oriented, antiparallel backbone strands and form only the standard Watson-Crick pairs (A, T) and (C, G). Choosing one backbone strand as a reference, a DNA oligomer consisting of n basepairs is identified with a sequence of bases $\mathsf{X}_1\mathsf{X}_2\cdots\mathsf{X}_n$, listed in the 5′ to 3′ direction along the strand, where $\mathsf{X}_a \in \{\mathsf{T}, \mathsf{A}, \mathsf{C}, \mathsf{G}\}$. The basepairs associated with this sequence are denoted by $(\mathsf{X}, \overline{\mathsf{X}})_1, (\mathsf{X}, \overline{\mathsf{X}})_2, \ldots, (\mathsf{X}, \overline{\mathsf{X}})_n$, where $\overline{\mathsf{X}}$ is defined as the Watson-Crick complement of X in the sense that $\overline{\mathsf{A}} = \mathsf{T}$, $\overline{\mathsf{T}} = \mathsf{A}$, $\overline{\mathsf{C}} = \mathsf{G}$, and $\overline{\mathsf{G}} = \mathsf{C}$. The notation $(\mathsf{X}, \overline{\mathsf{X}})_a$ for a basepair indicates that base X is attached to the reference strand, while $\overline{\mathsf{X}}$ is attached to the complementary strand, and there are four possible basepairs $(\mathsf{X}, \overline{\mathsf{X}})_a$ corresponding to the choice $\mathsf{X}_a \in \{\mathsf{T}, \mathsf{A}, \mathsf{C}, \mathsf{G}\}$.

We adopt a coarse-grain model of DNA[43,56,57,67] in which each base is modeled as a rigid object, so that the configuration of an oligomer is equivalent to the configuration of all of its constituent bases. We follow closely the coordinate conventions and notation used in a precursor work;[43] a complete, self-contained description of the coordinates is also provided in the supplementary material.[66] The configuration of an arbitrary base X_a is specified by giving the location of a reference point fixed in the base, and the orientation of a right-handed, orthonormal frame attached to the base. The reference point and frame vectors are defined according to the recent Curves+ implementation[68] of the Tsukuba convention,[67] which provides prescriptions for these quantities in terms of the atomic positions within a base. In the model, the positions of the non-hydrogen atoms in each base of each basepair with respect to the associated reference point and frame are considered to be constant. As a result, once the reference point and frame of each base are specified, so too are the positions of all of the non-hydrogen atoms.

In a rigid-base description, the three-dimensional configuration of a DNA oligomer is determined by the relative rotation and displacement between neighboring bases both across and along the two backbone strands. In part to ensure a simple Watson-Crick symmetry relation between the two possible choices of reference strand, we introduce a basepair origin and frame as the appropriate average of the two associated base reference points and frames, and a junction origin and frame as the appropriate average of two adjacent basepair origins and frames. Thus, the relative rotation and displacement between the bases X_a and $\overline{\mathsf{X}}_a$ across the strands can be described by an intra-basepair coordinate vector $y^a = (\vartheta, \xi)^a \in \mathbb{R}^6$ in the basepair frame, and the relative rotation and displacement between the basepairs $(\mathsf{X}, \overline{\mathsf{X}})_a$ and $(\mathsf{X}, \overline{\mathsf{X}})_{a+1}$ along the strands can be described by an inter-basepair coordinate vector $z^a = (\theta, \zeta)^a \in \mathbb{R}^6$ in the associated junction frame. The relative displacement coordinates $\xi^a, \zeta^a \in \mathbb{R}^3$ are of the standard Cartesian type in the appropriate frame, while the relative rotation coordinates $\vartheta^a, \theta^a \in \mathbb{R}^3$ are of the Cayley type in the appropriate frame, as detailed in the supplementary material.[66]

The definitions of the coordinates $(\vartheta, \xi)^a$ and $(\theta, \zeta)^a$ can be shown to satisfy all the qualitative guidelines set forth in the Cambridge convention[56] for nucleic acid structures, including the symmetry conditions associated with a change of reference strand. Accordingly, we refer to the components of the intra-basepair rotation vector ϑ^a as buckle-propeller-opening, the intra-basepair translation vector ξ^a as shear-stretch-stagger, the inter-basepair rotation vector θ^a as tilt-roll-twist, and the inter-basepair translation vector ζ^a as shift-slide-rise. We remark that because the components of ϑ^a and θ^a are rotational coordinates of the Cayley type, they are not conventional angular coordinates about various axes as employed by many authors; however, they can be put into correspondence with conventional angular coordinates, and in the case of small rotations (measured in radians) are nearly identical.

Notice that the complete configuration of a DNA oligomer is specified by introducing a vector $z^0 = (\theta, \zeta)^0$ of six additional coordinates for the first basepair frame and reference point with respect to an external, lab-fixed frame. Ignoring these six degrees of freedom exactly corresponds to eliminating the overall symmetry of rigid body motion that exists when there is no external potential field.

B. Free energy

Consistent with a rigid-base description of DNA, we adopt a free energy model in which the energy of an arbitrary, n-basepair oligomer is given by a general, shifted quadratic function

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \mathsf{K}(\mathsf{w} - \widehat{\mathsf{w}}) + \widehat{\mathsf{U}}, \tag{1}$$

where $\mathsf{w} = (y^1, z^1, y^2, z^2, \ldots, z^{n-1}, y^n) \in \mathbb{R}^{12n-6}$ is the vector of internal configuration coordinates, $\mathsf{K} \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a symmetric, positive-definite matrix of stiffness parameters, $\widehat{\mathsf{w}} \in \mathbb{R}^{12n-6}$ is a vector of shape parameters that define the ground or minimum energy state, and $\widehat{\mathsf{U}} \geq 0$ is a constant that represents the energy of this state compared to an unstressed state. Thus $\widehat{\mathsf{U}} = 0$ implies that the ground state is unstressed, whereas $\widehat{\mathsf{U}} > 0$ implies that it is stressed. In the latter case, the oligomer is referred to as being pre-stressed or frustrated. We are unaware of a prior discussion of the concept of pre-stress in the context of a coarse-grain model of DNA. Indeed, it cannot arise in a local nearest-neighbor, rigid-basepair model with a linear-chain topology of interactions as have been considered by various authors. However, the possibility of pre-stress or frustration arises in the nearest-neighbor, rigid-base model to be developed here and is a natural consequence of a double-chain topology of interactions.

We assume that the oligomer material parameters $\widehat{\mathsf{U}}$, $\widehat{\mathsf{w}}$, and K are completely determined by the oligomer length n and sequence $\mathsf{X}_1 \cdots \mathsf{X}_n$ along the reference strand. Equivalently, we assume there exist functions $\mathbb{U}$, $\mathbb{W}$, and $\mathbb{K}$ such that

$$\begin{gathered} \widehat{\mathsf{U}} = \mathbb{U}(n, \mathsf{X}_1, \ldots, \mathsf{X}_n), \\ \widehat{\mathsf{w}} = \mathbb{W}(n, \mathsf{X}_1, \ldots, \mathsf{X}_n), \quad \mathsf{K} = \mathbb{K}(n, \mathsf{X}_1, \ldots, \mathsf{X}_n). \end{gathered} \tag{2}$$

The aim of this article is to construct explicit forms of these material parameter functions. Indeed, with such functions in hand, the free energy function in (1) could then be constructed for oligomers of arbitrary length and sequence, which would allow various properties of their shape, stiffness, and frustration to be predicted and studied.

The freedom in the choice of reference strand, taken with the intrinsic objectivity of the free energy, imply that the functions $\mathbb{U}$, $\mathbb{W}$, and $\mathbb{K}$ cannot be completely arbitrary. To see this, let $\mathsf{U}(\mathsf{w})$ and $\mathsf{U}_*(\mathsf{w}_*)$ denote the free energies of an oligomer computed using the two different choices of reference strand: the sequence along one is $\mathsf{X}_1 \cdots \mathsf{X}_n$, and along the other it is $\mathsf{X}_1^* \cdots \mathsf{X}_n^*$. Thus, $\mathsf{U}(\mathsf{w})$ is given by the expression in (1) with the parameters in (2), and $\mathsf{U}_*(\mathsf{w}_*)$ is given by an exactly analogous expression with the parameters

$$\begin{gathered} \widehat{\mathsf{U}}_* = \mathbb{U}(n, \mathsf{X}_1^*, \ldots, \mathsf{X}_n^*), \\ \widehat{\mathsf{w}}_* = \mathbb{W}(n, \mathsf{X}_1^*, \ldots, \mathsf{X}_n^*), \quad \mathsf{K}_* = \mathbb{K}(n, \mathsf{X}_1^*, \ldots, \mathsf{X}_n^*). \end{gathered} \tag{3}$$

From the objectivity condition that $\mathsf{U}(\mathsf{w})$ must equal $\mathsf{U}_*(\mathsf{w}_*)$ for all possible configurations, together with the change-of-strand relations outlined in previous work,[43] we deduce that the material parameter functions must satisfy

$$\begin{gathered} \mathbb{U}(n, \mathsf{X}_1, \ldots, \mathsf{X}_n) = \mathbb{U}(n, \overline{\mathsf{X}}_n, \ldots, \overline{\mathsf{X}}_1), \\ \mathbb{W}(n, \mathsf{X}_1, \ldots, \mathsf{X}_n) = \mathsf{E}_n \mathbb{W}(n, \overline{\mathsf{X}}_n, \ldots, \overline{\mathsf{X}}_1), \\ \mathbb{K}(n, \mathsf{X}_1, \ldots, \mathsf{X}_n) = \mathsf{E}_n \mathbb{K}(n, \overline{\mathsf{X}}_n, \ldots, \overline{\mathsf{X}}_1) \mathsf{E}_n. \end{gathered} \tag{4}$$

Here, $\mathsf{E}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a block, trailing-diagonal matrix formed by $2n-1$ copies of the constant, diagonal matrix $\mathsf{E} = \mathrm{diag}(-1, 1, 1, -1, 1, 1) \in \mathbb{R}^{6\times 6}$, with the property that $\mathsf{E}_n = \mathsf{E}_n^T = \mathsf{E}_n^{-1}$. Specifically, we have

$$\mathsf{E}_n = \begin{pmatrix} & & & \mathsf{E} \\ & & \mathsf{E} & \\ & \iddots & & \\ \mathsf{E} & & & \end{pmatrix}. \tag{5}$$

The relations in (4) are a straightforward consequence of the Watson-Crick symmetry of DNA and can be described as follows. Characterize the 12 types of internal coordinates as being odd or even, where the odd coordinates are buckle, shear, tilt, and shift (one each of intra- and inter-basepair, and one each of translation and rotation) and the remaining 8 are all even. Then under a change of reference strand the odd coordinates of any configuration at any location along an oligomer change sign, whereas the even coordinates remain unaltered. The relations in (4) state that the material parameters delivered by the functions $\mathbb{U}$, $\mathbb{W}$, and $\mathbb{K}$ must be in concordance with this property. Specifically, the frustration energy delivered by $\mathbb{U}$ must be invariant to the choice of reference strand, the shape parameters delivered by $\mathbb{W}$ must behave in precisely the same way as the internal coordinates, and the stiffness parameters delivered by $\mathbb{K}$ must behave in a consistent way; namely, parameters for odd-even couplings change sign under a change of reference strand, whereas parameters for odd-odd and even-even couplings remain unaltered.

C. Configuration density

The equilibrium distribution of the internal coordinates $\mathsf{w} \in \mathbb{R}^{12n-6}$ of a rigid-base model of DNA, in contact with a heat bath at absolute temperature T, is described by the density function[43,44]

$$\rho(\mathsf{w}) = \frac{1}{Z_J} e^{-\mathsf{U}(\mathsf{w})/k_B T}\, J(\mathsf{w}), \quad Z_J = \int e^{-\mathsf{U}(\mathsf{w})/k_B T} J(\mathsf{w})\, d\mathsf{w}. \tag{6}$$

Here, k_B is the Boltzmann constant, Z_J is a normalizing constant, and J is a Jacobian factor which arises due to the non-Cartesian nature of the rotational coordinates, namely,

$$J(\mathsf{w}) = \left[\prod_{a=1}^{n-1} (1 + \tfrac{1}{4}|\theta^a|^2)^{-2}\right] \left[\prod_{a=1}^{n} (1 + \tfrac{1}{4}|\vartheta^a|^2)^{-2}\right]. \tag{7}$$

The statistical mechanical average $\langle \phi \rangle$ of any state function $\phi = \phi(\mathsf{w})$ with respect to the density ρ is given by

$$\langle \phi \rangle := \int \phi(\mathsf{w}) \rho(\mathsf{w})\, d\mathsf{w}, \tag{8}$$

where, in part because we have employed a Cayley parameterization of rotation matrices, all integrations are performed over the domain $\mathbb{R}^{12n-6}$.

Notice that the internal configuration density ρ, and hence the statistical mechanical average of any internal state function ϕ, is invariant under shifts of the oligomer free energy U. Hence, the actual value of the oligomer frustration energy $\widehat{\mathsf{U}}$ appearing in (1) does not explicitly affect the statistical properties of the system. This is consistent with the fact that the free energy of the system is itself only defined up to an arbitrary constant. In (1), the arbitrary constant is chosen so that zero energy corresponds to an unstressed state, which for a given oligomer may or may not correspond to an accessible state.

D. Nondimensionalization

For the purposes of numerics and analysis, it will be convenient to introduce scales and transform the rigid-base model into a dimensionless form. Specifically, we introduce a characteristic scale ℓ for the translational coordinates, a characteristic scale g for the rotational coordinates, and define dimensionless variables $\underline{y}^a$, $\underline{z}^a$, and $\underline{\mathsf{w}}$ by

$$\underline{y}^a = \mathsf{G}^{-1} y^a, \quad \underline{z}^a = \mathsf{G}^{-1} z^a, \quad \underline{\mathsf{w}} = \mathsf{G}_n^{-1}\mathsf{w}, \tag{9}$$

where $\mathsf{G} = \mathrm{diag}(g, g, g, \ell, \ell, \ell) \in \mathbb{R}^{6\times 6}$ is a constant, diagonal matrix and $\mathsf{G}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is the diagonal matrix formed by $2n-1$ copies of the matrix G. Moreover, we use the characteristic scale k_BT and define a dimensionless free energy by $\underline{\mathsf{U}} = \mathsf{U}/k_BT$. Substituting this and the above relations into (1), we obtain

$$\underline{\mathsf{U}}(\underline{\mathsf{w}}) = \frac{1}{2}(\underline{\mathsf{w}} - \widehat{\underline{\mathsf{w}}}) \cdot \underline{\mathsf{K}}(\underline{\mathsf{w}} - \widehat{\underline{\mathsf{w}}}) + \widehat{\underline{\mathsf{U}}}, \tag{10}$$

where $\widehat{\underline{\mathsf{w}}}$, $\underline{\mathsf{K}}$, and $\widehat{\underline{\mathsf{U}}}$ are dimensionless parameters given by

$$\widehat{\underline{\mathsf{w}}} = \mathsf{G}_n^{-1}\widehat{\mathsf{w}}, \quad \underline{\mathsf{K}} = \mathsf{G}_n\mathsf{K}\mathsf{G}_n/(k_BT), \quad \widehat{\underline{\mathsf{U}}} = \widehat{\mathsf{U}}/(k_BT). \tag{11}$$

A dimensionless form of the internal configuration density can also be derived. Specifically, by substituting (9) and (10) into (6) we obtain

$$\underline{\rho}(\underline{\mathsf{w}}) = \frac{1}{\underline{Z}_J} e^{-\underline{\mathsf{U}}(\underline{\mathsf{w}})}\, \underline{J}(\underline{\mathsf{w}}), \quad \underline{Z}_J = \int e^{-\underline{\mathsf{U}}(\underline{\mathsf{w}})}\, \underline{J}(\underline{\mathsf{w}})\, d\underline{\mathsf{w}}, \tag{12}$$

where $\underline{J}$ is the transformed Jacobian factor given by

$$\underline{J}(\underline{\mathsf{w}}) = \left[\prod_{a=1}^{n-1}(1 + \tfrac{1}{4}g^2|\underline{\theta}^a|^2)^{-2}\right]\left[\prod_{a=1}^{n}(1 + \tfrac{1}{4}g^2|\underline{\vartheta}^a|^2)^{-2}\right]. \tag{13}$$

In any given application, the scales ℓ and g would normally be set by the phenomena of interest. The scale ℓ is ideally chosen to describe the magnitude of the variation in the intra- and inter-basepair translational variables, whereas the scale g is ideally chosen to describe the magnitude of the variation in the intra- and inter-basepair rotational variables. In the analysis of molecular dynamics data of DNA, the phenomena of interest are fluctuations of atomic positions on the order of 1 Å. Hence, a reasonable scale for the translational variables is $\ell = 1$ Å because variations in these variables are in direct correspondence to variations in atomic positions. Moreover, a reasonable scale for the rotational variables is $g = 1/5$ (radians) because variations of this size in these rotation variables, about a zero reference value, correspond to a variation of about 1 Å in atomic positions of the atoms making up a base. This follows from the fact that the characteristic size of both a base and the junction between basepairs is approximately 5 Å, so that rotational variations of 1/5 give rise to variations in atomic positions of approximately 1 Å. Throughout the remainder of our developments, we restrict attention to the dimensionless formulation outlined above with these scales and drop the underline notation for convenience.

We remark that the precise values of the factors ℓ and g are inconsequential; rather, it is the order of magnitude of the scaling between rotational and translational variables that is significant when discussing the validity of an approximation. For example, a slightly different scaling was adopted in precursor work.[43] While the difference between these two scalings is not important, we now prefer the rationale for the choice that is given here.

E. Gaussian approximation

Although the free energy U is quadratic, the density ρ is non-Gaussian due to the presence of the Jacobian factor J. However, when the gradient of J is sufficiently small and ρ is sufficiently concentrated, it is reasonable to expect that variations in J can be neglected (see the supplementary material[66]). By the Gaussian approximation of the internal configuration density ρ we mean the density obtained by assuming J to be constant. In this approximation, we have

$$\rho(\mathsf{w}) = \frac{1}{Z}e^{-\mathsf{U}(\mathsf{w})}, \quad Z = \int e^{-\mathsf{U}(\mathsf{w})}\, d\mathsf{w}. \tag{14}$$

By making appropriate choices for the function ϕ in (8), and using standard results for Gaussian integrals,[69] we obtain the moment-parameter relations

$$\langle \mathsf{w}\rangle = \widehat{\mathsf{w}}, \quad \langle \Delta\mathsf{w} \otimes \Delta\mathsf{w}\rangle = \mathsf{K}^{-1}, \quad \langle \mathsf{w}\otimes\mathsf{w}\rangle = \mathsf{K}^{-1} + \widehat{\mathsf{w}}\otimes\widehat{\mathsf{w}}, \tag{15}$$

where $\Delta\mathsf{w} = \mathsf{w} - \widehat{\mathsf{w}}$ and $\otimes$ denotes the usual outer or tensor product of a vector; in components we have $[\Delta\mathsf{w}\otimes\Delta\mathsf{w}]_{pq} = \Delta\mathsf{w}_p\Delta\mathsf{w}_q$.

F. Probability density comparisons

The equilibrium statistical properties of the internal configuration of an oligomer are described by the density $\rho(\mathsf{w})$, which is completely defined by the free energy parameters K and $\widehat{\mathsf{w}}$. Throughout our developments it will be necessary to quantify the difference in two densities $\rho_{\mathrm{m}}(\mathsf{w})$ and $\rho_{\mathrm{o}}(\mathsf{w})$ defined by two different sets of parameters $\{\mathsf{K}_{\mathrm{m}}, \widehat{\mathsf{w}}_{\mathrm{m}}\}$ and $\{\mathsf{K}_{\mathrm{o}}, \widehat{\mathsf{w}}_{\mathrm{o}}\}$. To this end, we appeal to standard results from probability theory[70] and employ the Kullback-Leibler divergence

$$D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}) := \int \rho_{\mathrm{m}}(\mathsf{w}) \ln\left[\frac{\rho_{\mathrm{m}}(\mathsf{w})}{\rho_{\mathrm{o}}(\mathsf{w})}\right] d\mathsf{w}. \tag{16}$$

A more complete discussion of the Kullback-Leibler divergence can be found in the supplementary material,[66] but for our purposes it suffices to observe that in general $D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}) \neq D(\rho_{\mathrm{o}}, \rho_{\mathrm{m}})$, $D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}) \geq 0$ for any ρ_{m} and ρ_{o}, and $D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}) = 0$ if and only if $\rho_{\mathrm{m}} = \rho_{\mathrm{o}}$. And in the special case when ρ_{m} and ρ_{o} are both Gaussian, the integral in (16) can be explicitly evaluated to obtain

$$D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}) = \frac{1}{2}\left[\mathsf{K}_{\mathrm{m}}^{-1} : \mathsf{K}_{\mathrm{o}} - \ln(\det\mathsf{K}_{\mathrm{o}}/\det\mathsf{K}_{\mathrm{m}}) - I : I\right] + \frac{1}{2}(\widehat{\mathsf{w}}_{\mathrm{m}} - \widehat{\mathsf{w}}_{\mathrm{o}})\cdot\mathsf{K}_{\mathrm{o}}(\widehat{\mathsf{w}}_{\mathrm{m}} - \widehat{\mathsf{w}}_{\mathrm{o}}), \tag{17}$$

where a colon denotes the standard Euclidean inner product for square matrices and I denotes the identity matrix of the same dimension as K_{m} and K_{o}.

The divergence $D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}})$ is widely employed in various standard parameter estimation methods in statistics and statistical mechanics.[71–73] When ρ_{o} is interpreted as an observed density and ρ_{m} is interpreted as a model density, then the minimization of $D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}})$ over a space of admissible ρ_{m} yields a best-fit model density ρ_{m}^*. Alternatively, due to the lack of symmetry, minimization of $D(\rho_{\mathrm{o}}, \rho_{\mathrm{m}})$ over ρ_{m} yields a generally different best-fit model density ρ_{m}^{**}. The first approach can be described as model fitting via the maximum relative entropy principle,[74] whereas the second approach, in the Gaussian case, can be shown to correspond to model fitting via the maximum likelihood principle. In the developments that follow, we adopt the maximum relative entropy approach when fitting models to data.

III. MODELS

A. Nearest-neighbor assumption

For the purposes of building a free energy model, we consider different partitions of an oligomer into different types of structural units. Here, we restrict attention to a nearest-neighbor model built on two types of units: 1-mers (or monomers) and 2-mers (or dimers). However, we intentionally establish a notation that generalizes naturally to trimers, tetramers, and so on to facilitate extensions in later work. Specifically, an oligomer of n basepairs and arbitrary sequence $\mathsf{X}_1 \cdots \mathsf{X}_n$ can be partitioned into 1-mer units X_a, $a = 1, \ldots, n$, and 2-mer units $\mathsf{X}_a\mathsf{X}_{a+1}$, $a = 1, \ldots, n-1$. Just as the internal configuration of the oligomer is specified by the coordinate vector $\mathsf{w} \in \mathbb{R}^{12n-6}$, the internal configuration of each 1-mer X_a is specified by the coordinate vector $\mathsf{w}_1^a = y^a \in \mathbb{R}^6$, and the internal configuration of each 2-mer $\mathsf{X}_a\mathsf{X}_{a+1}$ is specified by the coordinate vector $\mathsf{w}_2^a = (y^a, z^a, y^{a+1}) \in \mathbb{R}^{18}$. For convenience, the collection of all 1-mer coordinates will be denoted by $\mathsf{w}_1 = (\mathsf{w}_1^1, \ldots, \mathsf{w}_1^n) \in \mathbb{R}^{6n}$, and the collection of all 2-mer coordinates will be denoted by $\mathsf{w}_2 = (\mathsf{w}_2^1, \ldots, \mathsf{w}_2^{n-1}) \in \mathbb{R}^{18(n-1)}$. We stress that there is considerable redundancy in this notation in that the intra-basepair variables y^a appear in both w_2^a and w_2^{a-1}. This redundancy is both notationally convenient and physically pertinent; the redundancy is the mathematical expression of the physical phenomenon of frustration.

In our developments, it will be necessary to consider various linear maps between the vectors w, w_1, and w_2. The matrix which copies elements of w into the 1-mer vector w_1 is denoted by $\mathsf{P}_1 \in \mathbb{R}^{6n} \times \mathbb{R}^{12n-6}$, so that $\mathsf{w}_1 = \mathsf{P}_1\mathsf{w}$, and the matrix which copies elements of w into the 2-mer vector w_2 is denoted by $\mathsf{P}_2 \in \mathbb{R}^{18(n-1)} \times \mathbb{R}^{12n-6}$, so that $\mathsf{w}_2 = \mathsf{P}_2\mathsf{w}$. The explicit forms of these matrices are

$$\mathsf{P}_1 = \begin{pmatrix} I & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & 0 & 0 & I & & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & I \end{pmatrix}, \quad \mathsf{P}_2 = \begin{pmatrix} I & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & I & 0 & 0 & 0 & & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & 0 & I & 0 & & 0 \\ 0 & 0 & 0 & 0 & I & & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & I \end{pmatrix}, \tag{18}$$

where $0 \in \mathbb{R}^{6\times 6}$ and $I \in \mathbb{R}^{6\times 6}$ denote the zero and identity matrices. It will also be necessary to consider the transpose matrix P_1^T, which maps a 1-mer vector u_1 into an oligomer vector $\mathsf{u} = \mathsf{P}_1^T\mathsf{u}_1$. Here, the entries in u_1 (all of which are intra-basepair coordinates) are mapped to their corresponding location in u, and the remaining entries of u (all of which correspond to inter-basepair coordinates) are zero. Similarly, the transpose matrix P_2^T maps a 2-mer vector u_2 into an oligomer vector $\mathsf{u} = \mathsf{P}_2^T\mathsf{u}_2$. Here, the entries in u_2 are mapped to their corresponding location in u, and overlapping contributions are summed.

We consider a free energy model based on local energies that describe physically distinct interactions within the 1-mer and 2-mer units. Specifically, to any 1-mer X_a we associate an energy of the form

$$\mathsf{U}_1^a(\mathsf{w}_1^a) = \frac{1}{2}(\mathsf{w}_1^a - \widehat{\mathsf{w}}_1^a) \cdot \mathsf{K}_1^a(\mathsf{w}_1^a - \widehat{\mathsf{w}}_1^a), \tag{19}$$

where $\widehat{\mathsf{w}}_1^a \in \mathbb{R}^6$ is a vector of shape parameters that define the minimum energy or ground state of the interaction, and $\mathsf{K}_1^a \in \mathbb{R}^{6\times 6}$ is a symmetric matrix of stiffness parameters that describe the elastic stiffness associated with each internal coordinate and couplings between them. The energy U_1^a is to be interpreted as a model for the intra-basepair interactions between the two bases of $(\mathsf{X}, \overline{\mathsf{X}})_a$. The description of these interactions involves only the intra-basepair coordinates $\mathsf{w}_1^a = y^a$, and the stiffness matrix K_1^a may in general be dense.

Similarly, to any 2-mer $\mathsf{X}_a\mathsf{X}_{a+1}$ we associate an energy

$$\mathsf{U}_2^a(\mathsf{w}_2^a) = \frac{1}{2}(\mathsf{w}_2^a - \widehat{\mathsf{w}}_2^a) \cdot \mathsf{K}_2^a(\mathsf{w}_2^a - \widehat{\mathsf{w}}_2^a), \tag{20}$$

where $\widehat{\mathsf{w}}_2^a \in \mathbb{R}^{18}$ is a vector of shape parameters and $\mathsf{K}_2^a \in \mathbb{R}^{18\times 18}$ is a symmetric matrix of stiffness parameters analogous to before. The energy U_2^a is to be interpreted as a model for all the inter-basepair interactions involving a base of $(\mathsf{X}, \overline{\mathsf{X}})_a$ and a base of $(\mathsf{X}, \overline{\mathsf{X}})_{a+1}$, in other words any nearest-neighbor, base-base interaction across the junction between the basepairs $(\mathsf{X}, \overline{\mathsf{X}})_a$ and $(\mathsf{X}, \overline{\mathsf{X}})_{a+1}$. The description of all these interactions naturally involves the internal coordinates $\mathsf{w}_2^a = (y^a, z^a, y^{a+1})$. The stiffness matrix K_2^a may, in general, be dense and has the natural block form

$$\mathsf{K}_2^a = \begin{pmatrix} \mathsf{K}_{2,11}^a & \mathsf{K}_{2,12}^a & \mathsf{K}_{2,13}^a \\ \mathsf{K}_{2,21}^a & \mathsf{K}_{2,22}^a & \mathsf{K}_{2,23}^a \\ \mathsf{K}_{2,31}^a & \mathsf{K}_{2,32}^a & \mathsf{K}_{2,33}^a \end{pmatrix}, \tag{21}$$

where each entry is an element of $\mathbb{R}^{6\times 6}$. The assumption that the overall matrix is symmetric implies that the diagonal blocks $\mathsf{K}_{2,11}^a$, $\mathsf{K}_{2,22}^a$, and $\mathsf{K}_{2,33}^a$ are each symmetric, and implies that the off-diagonal blocks satisfy $[\mathsf{K}_{2,12}^a]^T = \mathsf{K}_{2,21}^a$, $[\mathsf{K}_{2,13}^a]^T = \mathsf{K}_{2,31}^a$, and $[\mathsf{K}_{2,23}^a]^T = \mathsf{K}_{2,32}^a$.

The local 1-mer and 2-mer energies can be summed in a natural way to obtain an overall oligomer energy. Specifically, we define the oligomer energy as

$$\begin{aligned}\mathsf{U}(\mathsf{w}) &= \sum_{j=1}^{2}\sum_{a=1}^{n-j+1}\mathsf{U}_j^a(\mathsf{w}_j^a)\\ &= \frac{1}{2}\sum_{j=1}^{2}\sum_{a=1}^{n-j+1}(\mathsf{w}_j^a-\widehat{\mathsf{w}}_j^a)\cdot\mathsf{K}_j^a(\mathsf{w}_j^a-\widehat{\mathsf{w}}_j^a). \qquad (22)\end{aligned}$$

Here, the index $j = 1, 2$ is a label for the 1-mer and 2-mer interactions, and for each j, the index $a = 1, \ldots, n - j + 1$ runs over all the j-mers along the oligomer. The oligomer energy can be written in a more convenient matrix form as the sum of two shifted quadratic forms (of different dimensions)

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}\sum_{j=1}^{2}(\mathsf{P}_j\mathsf{w}-\widehat{\mathsf{w}}_j)\cdot\mathsf{K}_j(\mathsf{P}_j\mathsf{w}-\widehat{\mathsf{w}}_j), \qquad (23)$$

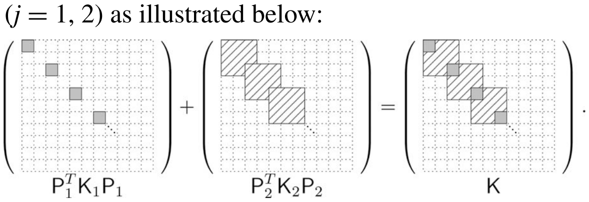
where $\widehat{\mathsf{w}}_j = (\widehat{\mathsf{w}}_j^1, \ldots, \widehat{\mathsf{w}}_j^{n-j+1})$ is a vector containing all the j-mer shape parameters, $\mathsf{K}_j = \mathrm{diag}(\mathsf{K}_j^1, \ldots, \mathsf{K}_j^{n-j+1})$ is a block-diagonal matrix containing all the j-mer stiffness parameters, and P_j is the matrix which copies oligomer coordinates into j-mer coordinates as defined above. By completing squares, we find that this energy can be expressed in the standard form introduced in Eq. (1) of Sec. II B,

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}(\mathsf{w}-\widehat{\mathsf{w}})\cdot\mathsf{K}(\mathsf{w}-\widehat{\mathsf{w}})+\widehat{\mathsf{U}}, \qquad (24)$$

where

$$\begin{aligned}\mathsf{K} &= \sum_{j=1}^{2}\mathsf{P}_j^T\mathsf{K}_j\mathsf{P}_j,\\ \widehat{\mathsf{w}} &= \mathsf{K}^{-1}\left(\sum_{j=1}^{2}\mathsf{P}_j^T\mathsf{K}_j\widehat{\mathsf{w}}_j\right), \qquad (25)\\ \widehat{\mathsf{U}} &= \frac{1}{2}\sum_{j=1}^{2}(\mathsf{P}_j\widehat{\mathsf{w}}-\widehat{\mathsf{w}}_j)\cdot\mathsf{K}_j(\mathsf{P}_j\widehat{\mathsf{w}}-\widehat{\mathsf{w}}_j).\end{aligned}$$

The above relations play a fundamental role in our free energy model. They show how the oligomer-based energy parameters K, $\widehat{\mathsf{w}}$, and $\widehat{\mathsf{U}}$ depend on the j-mer-based energy parameters contained in K_j and $\widehat{\mathsf{w}}_j$. Specifically, from $(25)_1$ and the definitions of P_j and K_j, we deduce that K is a banded, block matrix whose entries depend locally on the entries of K_j $(j = 1, 2)$ as illustrated below:



(S1)

In the illustration, the shaded entries denote the blocks $\mathsf{K}_1^1, \ldots, \mathsf{K}_1^n \in \mathbb{R}^{6\times 6}$ of K_1, the ruled entries denote the blocks $\mathsf{K}_2^1, \ldots, \mathsf{K}_2^{n-1} \in \mathbb{R}^{18\times 18}$ of K_2, grid lines denote elements of $\mathbb{R}^{6\times 6}$, and entries in the double and triple overlaps are summed in the obvious way.

In contrast to the stiffness, the entries of the oligomer shape vector $\widehat{\mathsf{w}}$ do not depend locally on the entries of $\widehat{\mathsf{w}}_j$ ($j = 1, 2$). Indeed, from $(25)_2$ we see that the vector $\widehat{\mathsf{w}}$ is related to the vectors $\widehat{\mathsf{w}}_j$ through the inverse matrix K^{-1}. Specifically, the entries in the product $\mathsf{K}_j\widehat{\mathsf{w}}_j$ depend locally on the entries of $\widehat{\mathsf{w}}_j$; the product is a j-mer vector of weighted j-mer shape parameters. Moreover, by definition of P_j^T, the entries of the product $\mathsf{P}_j^T\mathsf{K}_j\widehat{\mathsf{w}}_j$ also depend locally on those of $\widehat{\mathsf{w}}_j$; the matrix P_j^T maps contributions from each j-mer into its corresponding location in the oligomer, with overlapping contributions being summed. However, the matrix K^{-1} will in general be dense; its diagonal blocks will be dominant with its off-diagonal blocks decaying with distance from the diagonal at a specific rate in accordance with the bandwidth of K. From this we deduce that the oligomer shape parameters in $\widehat{\mathsf{w}}$ will be a convolution of the j-mer shape parameters in $\widehat{\mathsf{w}}_j$. The convolution window will be peaked along the diagonal of K^{-1} and will decay at a rate as described above.

The oligomer energy introduced here provides a natural model for frustration. Indeed, the term $\widehat{\mathsf{U}}$ in $(25)_3$ will, in general, be non-zero. This reflects the fact that, in the minimum energy or ground state $\widehat{\mathsf{w}}$ of the oligomer, each base cannot simultaneously minimize all of its j-mer interactions. Explicitly, each base cannot simultaneously minimize its intra-basepair interaction energy, and its two inter-basepair, cross-junction interaction energies. Instead, each base must find a compromise, which provides the physical explanation for the nonlocal nature of $\widehat{\mathsf{w}}$. We may refer to the forces between bases required to maintain this state of compromise as the frustration forces; their collective energy content over the entire oligomer is precisely the frustration energy $\widehat{\mathsf{U}}$.

B. Sequence-dependence assumptions

A nearest-neighbor free energy for an oligomer of length n is completely determined by the shape and stiffness parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ $(a = 1, \ldots, n)$ and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ $(a = 1, \ldots, n-1)$ introduced above. In general, these parameters may depend on the oligomer length n, the entire oligomer sequence $\mathsf{X}_1\cdots\mathsf{X}_n$, and the location a within the sequence. Of course, we seek the simplest possible model for this dependence compatible with a desired accuracy. Here, we consider two such models for the parameters and outline their properties.

1. Oligomer-based nearest-neighbor model

By an oligomer-based model, we mean one in which the parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ depend on the oligomer length, sequence, and location in the most general way. Specifically, we assume there exist functions $\mathbb{W}_1$, $\mathbb{K}_1$, $\mathbb{W}_2$, and $\mathbb{K}_2$ such that

$$\begin{aligned}\widehat{\mathsf{w}}_1^a &= \mathbb{W}_1(n, \mathsf{X}_1, \ldots, \mathsf{X}_n, a) \in \mathbb{R}^6,\\ \mathsf{K}_1^a &= \mathbb{K}_1(n, \mathsf{X}_1, \ldots, \mathsf{X}_n, a) \in \mathbb{R}^{6\times 6},\\ \widehat{\mathsf{w}}_2^a &= \mathbb{W}_2(n, \mathsf{X}_1, \ldots, \mathsf{X}_n, a) \in \mathbb{R}^{18}, \qquad (26)\\ \mathsf{K}_2^a &= \mathbb{K}_2(n, \mathsf{X}_1, \ldots, \mathsf{X}_n, a) \in \mathbb{R}^{18\times 18}.\end{aligned}$$

In view of the relations in (25), this assumption implies that the oligomer stiffness and shape parameters K and $\widehat{\mathsf{w}}$ could be arbitrary functions of the oligomer length and sequence, subject only to two conditions. First, that K should have the nearest-neighbor sparsity structure illustrated in (S1), and second that the parameters for an oligomer with sequence $\mathsf{X}_1 \cdots \mathsf{X}_n$ are necessarily related by objectivity to those for an oligomer with sequence $\overline{\mathsf{X}}_n \cdots \overline{\mathsf{X}}_1$. Thus in this model there is no finite set of parameters that describe all possible oligomers of all possible lengths. Nevertheless, we will make use of such oligomer-based models as an intermediate step in our consideration of parameter training sets extracted from molecular dynamics simulations.

2. Dimer-based nearest-neighbor model

By a dimer-based model, we mean one in which the parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ depend only on the local dimer sequence $\mathsf{X}_a\mathsf{X}_{a+1}$, but not explicitly on either the oligomer length n nor the location a. Specifically, we assume there exist functions $\mathbb{W}_1$, $\mathbb{K}_1$, $\mathbb{W}_2$, and $\mathbb{K}_2$ such that

$$\begin{aligned} \widehat{\mathsf{w}}_1^a &= \mathbb{W}_1(\mathsf{X}_a) \in \mathbb{R}^6, \quad \mathsf{K}_1^a = \mathbb{K}_1(\mathsf{X}_a) \in \mathbb{R}^{6\times 6}, \\ \widehat{\mathsf{w}}_2^a &= \mathbb{W}_2(\mathsf{X}_a, \mathsf{X}_{a+1}) \in \mathbb{R}^{18}, \quad \mathsf{K}_2^a = \mathbb{K}_2(\mathsf{X}_a, \mathsf{X}_{a+1}) \in \mathbb{R}^{18\times 18}. \end{aligned} \tag{27}$$

Notice that the above relations are assumed to hold at the ends of an arbitrary oligomer as well as in its interior; additional assumptions or parameters could be introduced to capture any exceptional end effects, but we do not explore that line of investigation here. For this dimer-based model, we note that there is a finite set of parameters that describe the energy of all possible oligomers of all possible lengths. Specifically, each of the functions $\mathbb{W}_1(\mathsf{X}_a)$ and $\mathbb{K}_1(\mathsf{X}_a)$ can assume only 4 possible values corresponding to the 4 possible choices of $\mathsf{X}_a \in \{\mathsf{T}, \mathsf{A}, \mathsf{C}, \mathsf{G}\}$, of which only 2 are independent. Similarly, each of the functions $\mathbb{W}_2(\mathsf{X}_a, \mathsf{X}_{a+1})$ and $\mathbb{K}_2(\mathsf{X}_a, \mathsf{X}_{a+1})$ can assume only 16 possible values corresponding to the 16 possible choices of the pair $\mathsf{X}_a, \mathsf{X}_{a+1} \in \{\mathsf{T}, \mathsf{A}, \mathsf{C}, \mathsf{G}\}$, of which only 10 are independent.

The numbers of independent functions is dictated by objectivity. In this respect, it is sufficient to assume that the functions $\mathbb{W}_1$, $\mathbb{K}_1$, $\mathbb{W}_2$, and $\mathbb{K}_2$ are locally objective in the following sense, for any $\mathsf{X}, \mathsf{Y} \in \{\mathsf{T}, \mathsf{A}, \mathsf{C}, \mathsf{G}\}$:

$$\begin{aligned} \mathbb{W}_1(\mathsf{X}) &= \mathsf{E}_1\mathbb{W}_1(\overline{\mathsf{X}}), \quad \mathbb{K}_1(\mathsf{X}) = \mathsf{E}_1\mathbb{K}_1(\overline{\mathsf{X}})\mathsf{E}_1, \\ \mathbb{W}_2(\mathsf{X}, \mathsf{Y}) &= \mathsf{E}_2\mathbb{W}_2(\overline{\mathsf{Y}}, \overline{\mathsf{X}}), \quad \mathbb{K}_2(\mathsf{X}, \mathsf{Y}) = \mathsf{E}_2\mathbb{K}_2(\overline{\mathsf{Y}}, \overline{\mathsf{X}})\mathsf{E}_2. \end{aligned} \tag{28}$$

These conditions imply that the values of the functions are not all independent. Specifically, if we arrange the 4 possible values for X in a table as shown

$$\mathsf{A} \quad \mathsf{G} \quad | \quad \mathsf{T} \quad \mathsf{C},$$

then the value of $\mathbb{W}_1$ for the entries on the right-half of the table are completely determined by those of the left-half, namely, $\mathbb{W}_1(\mathsf{T}) = \mathsf{E}_1\mathbb{W}_1(\mathsf{A})$ and $\mathbb{W}_1(\mathsf{C}) = \mathsf{E}_1\mathbb{W}_1(\mathsf{G})$, and similarly for $\mathbb{K}_1$. Thus, there are only 2 independent values of the monomer parameter functions $\{\mathbb{W}_1, \mathbb{K}_1\}$. Similarly, if we arrange the 16 possible values of XY in a table as shown, where X is vertical and Y is horizontal,

	T	C	A	G
A	AT	AC	AA	AG
G	GT	GC	GA	GG
T	TT	TC	TA	TG
C	CT	CC	CA	CG,

then the value of $\mathbb{W}_2$ for the entries above the table diagonal are completely determined by those below, namely, $\mathbb{W}_2(\mathsf{A}, \mathsf{C}) = \mathsf{E}_2\mathbb{W}_2(\mathsf{G}, \mathsf{T})$ and so on, and similarly for $\mathbb{K}_2$. Moreover, the value of $\mathbb{W}_2$ for each diagonal entry must satisfy the self-symmetry condition $\mathbb{W}_2(\mathsf{A}, \mathsf{T}) = \mathsf{E}_2\mathbb{W}_2(\mathsf{A}, \mathsf{T})$ and so on, and similarly for $\mathbb{K}_2$. From this, we deduce that there are only 10 independent values of the dimer parameter functions $\{\mathbb{W}_2, \mathbb{K}_2\}$ corresponding to a triangular portion of the table with the diagonal included, and that values for the diagonal entries must be invariant under the transformation E_2. Note that the four 2×2 blocks of the table have some physical significance: the two diagonal blocks correspond to purine-pyrimidine and pyrimidine-purine dimer steps, whereas the two off-diagonal blocks correspond to purine-purine and pyrimidine-pyrimidine dimer steps. In summary, a nearest-neighbor, dimer-based model is completely defined by a set of parameters

$$\begin{aligned} &\widehat{\mathsf{w}}_1^\alpha := \mathbb{W}_1(\alpha) \in \mathbb{R}^6, \quad \mathsf{K}_1^\alpha := \mathbb{K}_1(\alpha) \in \mathbb{R}^{6\times 6}, \quad \alpha \in \mathsf{M}, \\ &\widehat{\mathsf{w}}_2^{\alpha\beta} := \mathbb{W}_2(\alpha, \beta) \in \mathbb{R}^{18}, \quad \mathsf{K}_2^{\alpha\beta} := \mathbb{K}_2(\alpha, \beta) \in \mathbb{R}^{18\times 18}, \\ &\qquad\qquad \alpha\beta \in \mathsf{D}, \end{aligned} \tag{29}$$

where M is a set of any 2 independent monomers and D is a set of any 10 independent dimers.

In our later developments, it will be convenient to work with weighted shape parameters instead of the unweighted parameters $\widehat{\mathsf{w}}_1^\alpha$ and $\widehat{\mathsf{w}}_2^{\alpha\beta}$ introduced above. Specifically, we introduce the weighted shape parameters

$$\sigma_1^\alpha := \mathsf{K}_1^\alpha\widehat{\mathsf{w}}_1^\alpha \in \mathbb{R}^6, \quad \sigma_2^{\alpha\beta} := \mathsf{K}_2^{\alpha\beta}\widehat{\mathsf{w}}_2^{\alpha\beta} \in \mathbb{R}^{18}. \tag{30}$$

We remark that the parameters σ_1^α and $\sigma_2^{\alpha\beta}$ are stress-like in character, each being the product of a strain parameter by a stiffness matrix. Their physical interpretation is not immediately evident, but they arise very naturally in the algebra of the problem. Hence, a complete nearest-neighbor, dimer-based parameter set is defined by specifying the values of $\{\sigma_1^\alpha, \mathsf{K}_1^\alpha\}$ for $\alpha \in \mathsf{M}$ and the values of $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ for $\alpha\beta \in \mathsf{D}$. Parameters for all other monomers and dimers can be obtained from the objectivity relations (28), which also imply that the parameters associated with the four palindromic dimer steps must satisfy the appropriate self-symmetry conditions. One such complete parameter set will be presented later.

The shape, stiffness, and frustration parameters $\widehat{\mathsf{w}}$, K, and $\widehat{\mathsf{U}}$ for any given oligomer $\mathsf{X}_1 \cdots \mathsf{X}_n$ can be assembled from the monomer and dimer parameters $\{\sigma_1^\alpha, \mathsf{K}_1^\alpha\}$ and $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ using the relations in (25). Specifically, and omitting the expression for $\widehat{\mathsf{U}}$ for brevity, the expressions for the oligomer parameters $\widehat{\mathsf{w}}$ and K take the forms

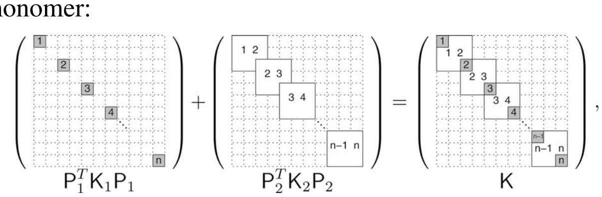
$$\begin{aligned} \mathsf{K} &= \mathsf{P}_1^T\mathsf{K}_1\mathsf{P}_1 + \mathsf{P}_2^T\mathsf{K}_2\mathsf{P}_2 && \in \mathbb{R}^{(12n-6)\times(12n-6)}, \\ \sigma &= \mathsf{P}_1^T\sigma_1 + \mathsf{P}_2^T\sigma_2 && \in \mathbb{R}^{12n-6}, \\ \widehat{\mathsf{w}} &= \mathsf{K}^{-1}\sigma && \in \mathbb{R}^{12n-6}, \end{aligned} \tag{31}$$

where

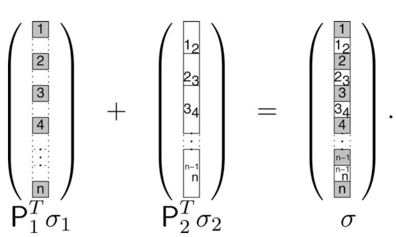
$$\begin{aligned} \mathsf{K}_1 &= \mathrm{diag}(\mathsf{K}_1^1,\ldots,\mathsf{K}_1^n) && \in \mathbb{R}^{6n\times 6n}, \\ \sigma_1 &= (\sigma_1^1,\ldots,\sigma_1^n) && \in \mathbb{R}^{6n}, \\ \mathsf{K}_2 &= \mathrm{diag}(\mathsf{K}_2^1,\ldots,\mathsf{K}_2^{n-1}) && \in \mathbb{R}^{18(n-1)\times 18(n-1)}, \\ \sigma_2 &= (\sigma_2^1,\ldots,\sigma_2^{n-1}) && \in \mathbb{R}^{18(n-1)}, \\ \mathsf{K}_1^a &= \mathsf{K}_1^{\mathsf{X}_a} && \in \mathbb{R}^{6\times 6} && (a=1,\ldots,n), \\ \sigma_1^a &= \sigma_1^{\mathsf{X}_a} && \in \mathbb{R}^{6} && (a=1,\ldots,n), \\ \mathsf{K}_2^a &= \mathsf{K}_2^{\mathsf{X}_a\mathsf{X}_{a+1}} && \in \mathbb{R}^{18\times 18} && (a=1,\ldots,n-1), \\ \sigma_2^a &= \sigma_2^{\mathsf{X}_a\mathsf{X}_{a+1}} && \in \mathbb{R}^{18} && (a=1,\ldots,n-1). \end{aligned} \tag{32}$$

Notice that the oligomer stiffness matrix K and shape vector $\widehat{\mathsf{w}}$ depend directly on the local stiffness and weighted shape parameters $\{\sigma_1^\alpha, \mathsf{K}_1^\alpha\}$ and $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$, but only indirectly on the local unweighted shape parameters $\widehat{\mathsf{w}}_1^\alpha$ and $\widehat{\mathsf{w}}_2^{\alpha\beta}$ via the definition (30). For this reason, we henceforth restrict attention to the weighted parameters. The unweighted parameters are needed explicitly only when either the oligomer frustration parameter $\widehat{\mathsf{U}}$ or the local energy functions U_1^a and U_2^a are needed explicitly.

The dependence of each block of the oligomer stiffness matrix K and weighted shape vector σ upon the oligomer sequence $\mathsf{X}_1\cdots\mathsf{X}_n$ are illustrated below. On the left-hand sides, each single number in a block denotes a dependence on the monomer X_a, while each pair of numbers in a block denotes a dependence on the dimer $\mathsf{X}_a\mathsf{X}_{a+1}$. On the right-hand sides, the double and triple overlapping blocks denote sums as before; notice that the shaded blocks with triple overlaps exhibit an effective dependence on the trimer $\mathsf{X}_{a-1}\mathsf{X}_a\mathsf{X}_{a+1}$ corresponding to the union of two adjacent dimers and a central monomer:



(S2)



(S3)

The overlapping structure of the map from local to oligomer parameters defined in (31) and (32) and illustrated in (S2) and (S3) has some interesting implications. A first implication concerns the observability of the local parameters. In positions within the oligomer arrays where there are no overlaps, we find that the local parameters are directly observable: there is no coupling and the oligomer parameters are equal to the relevant local parameters. In contrast, in positions within the oligomer arrays where there are overlaps, we find that the local parameters are not directly observable: there is coupling and the oligomer parameters are sums of relevant local parameters. Specifically, for the stiffness parameters, the structure of the coupling at interior positions within an oligomer are all identical, of the form $\mathsf{K}_{2,33}^{\alpha\beta} + \mathsf{K}_1^\beta + \mathsf{K}_{2,11}^{\beta\gamma}$, whereas the structure of the coupling at each of the two ends is different, of the form $\mathsf{K}_1^\beta + \mathsf{K}_{2,11}^{\beta\gamma}$ at the leading end, and $\mathsf{K}_{2,33}^{\alpha\beta} + \mathsf{K}_1^\beta$ at the trailing end, for some α, β, and γ. Exactly analogous couplings hold for the weighted shape parameters. In order that these couplings can be resolved in the inverse problem of determining local parameters from oligomer parameters, data from the ends of an oligomer as well as from its interior are required.

A second implication of the structure of the map from local to oligomer parameters concerns the positivity of the local stiffness parameters. The set of local stiffness parameters $\{\mathsf{K}_1^\alpha, \mathsf{K}_2^{\alpha\beta}\}$ is called admissible if it yields a positive-definite oligomer stiffness matrix K for an arbitrary sequence $\mathsf{X}_1\mathsf{X}_2\cdots\mathsf{X}_n$ of arbitrary length $n \geq 2$. In view of the additive, overlapping structure of the map from local to oligomer parameters, a sufficient condition for admissibility is that each of the local stiffness parameter matrices be positive-definite. Alternatively, a weaker set of conditions are also sufficient, namely, that each of the local stiffness parameter matrices be only semi-positive-definite, provided additionally that the reconstructed oligomer matrices for any ten independent sequences of length two (physical dimers) are positive-definite. This latter, weaker set of sufficient conditions is mathematically more convenient than the former in dealing with certain optimization problems associated with the estimation of local parameters. As it happens, the parameter set we extract below has positive-definite local stiffness matrices, but some are nearly only semi-definite due to the presence of some extremely small eigenvalues. Nevertheless, the full set of reconstructed physical dimer stiffness matrices are robustly positive-definite; their smallest eigenvalues are much larger than those of the local matrices. Thus, the weakened sufficient conditions seem to be of some importance in the parameter estimation problem.

IV. THE TRAINING DATA SET

The development of the previous Sec. III resolves the forward problem for a nearest-neighbor, dimer-based model of DNA. In other words, we have described how to reconstruct a shifted quadratic model of the free energy of an oligomer of arbitrary length and sequence starting from a finite parameter set. However, our main interest is the more difficult, inverse problem of how to estimate this finite parameter set from a sufficiently rich training data set. Here, we describe the data set that was used for this purpose. While the data we employed were generated using MD simulation, we remark that our parameter extraction methods could be applied to any analogous, sufficiently sequence-rich and accurate, training data set obtained by other techniques, for example, NMR.

A. Basic assumptions

We consider data for oligomers of known length n_μ and sequence S_μ, $\mu = 1, \ldots, N$. For each oligomer, we assume knowledge of an observed internal configuration density $\rho_{\mu,\mathrm{o}}(\mathsf{w})$ of the dimensionless coordinates $\mathsf{w} \in \mathbb{R}^{12n_\mu-6}$. In a first approximation, we assume that each density is a general Gaussian of the form

$$\rho_{\mu,\mathrm{o}}(\mathsf{w}) = \frac{1}{Z_{\mu,\mathrm{o}}} e^{-(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{o}})\cdot\mathsf{K}_{\mu,\mathrm{o}}(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{o}})/2}, \tag{33}$$

where $\widehat{\mathsf{w}}_{\mu,\mathrm{o}} \in \mathbb{R}^{12n_\mu-6}$ is the observed mean and $\mathsf{K}_{\mu,\mathrm{o}} \in \mathbb{R}^{(12n_\mu-6)\times(12n_\mu-6)}$ is the observed stiffness matrix for the oligomer. We assume $\mathsf{K}_{\mu,\mathrm{o}}$ is symmetric and positive-definite, but make no assumption about its sparsity. Normally, for each oligomer, the density $\rho_{\mu,\mathrm{o}}(\mathsf{w})$ is not known directly, but only indirectly through a sample set or time series. In this case, the observed parameters $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$ and $\mathsf{K}_{\mu,\mathrm{o}}$ must be computed accordingly. Below we describe the computation of these parameters for the specific case of time series data from MD.

B. The ABC data set

Our starting point was a shared pool of MD data on DNA produced by the ABC collaboration within a consortium of groups.[19–21] The ABC collaboration was initiated precisely because of interest in coarse-grain parameter extraction, although the data set is also being exploited in a variety of other ways by other members of the consortium. The ABC data set contains MD simulations of the 39 different 18-mers labeled by $\mu = 1, \ldots, 36$ and $\mu = 54, 55, 56$ in Table I. This set of oligomers was designed to include multiple instances of all 136 possible tetramer sub-sequences away from the ends. Crucially for us, every ABC oligomer was chosen to have 5′-GC and GC-3′ ends. This choice was made to minimize possible convergence issues in the simulations because GC dimer ends were known to be among the most stable against end-fraying. In the development of the theory presented here, and as explained more in the supplementary material,[66] we realized that while we did not need all tetramer sub-sequences to be present in our training set, we did however need a greater variety of end-sequences. We therefore enhanced the original ABC data set with 3 different 18-mers and 14 different 12-mers labeled by $\mu = 37, \ldots, 53$ in Table I. When considering both the reference and complementary strands, these additional oligomers contain all 16 possible 5′-dimer-step ends, and all 16 possible dimer-step-3′ ends. As the additional oligomers were focussed on enhancing the range of end-sequences in the training set, it was significantly faster to simulate 12-mers rather than 18-mers, which partly motivated our choice to use these shorter oligomers.

C. MD simulation protocol

Each DNA oligomer in the data set was simulated using atomic-resolution, explicit-solvent MD. The AMBER suite of programs together with the *parmbsc0* force field[23] was used. Simulations were run in water as modeled by the SPC/E parameters[75] with potassium neutralizing counter ions and a total of 150 mM of KCl salt modeled with the parameters from Dang.[76] For each oligomer, the DNA duplex was built, neutralized, hydrated, and equilibrated using a well-defined protocol described in Ref. 21. The total number of atoms contained in each simulation was approximately 36 000 for each 18-mer, and approximately 17 000 for each 12-mer. Each simulation was run in the *NpT* ensemble with a temperature of 300 K and a pressure of 1 atm, controlled by the Berendsen algorithm.[77] For each oligomer, a time series trajectory was generated, where the length varied between 50 and 200 ns depending on the consortium group running the simulation, and a configuration snapshot was saved every 1 ps.

D. Observed training set data

The program Curves+[68] was used to calculate the coarse-grain intra- and inter-basepair coordinates of each atomic-resolution configuration saved in each MD simulation. In this way, a time series of dimensionless internal coordinate vectors $\mathsf{w}_\mu^{(l)}$, $l = 1, \ldots, L_\mu$, was obtained for each oligomer S_μ in Table I, where L_μ denotes the total number of snapshots for the oligomer. The simplest way to determine the oligomer parameters $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$ and $\mathsf{K}_{\mu,\mathrm{o}}$ is to assume that the time series $\mathsf{w}_\mu^{(l)}$ is ergodic with respect to the density $\rho_{\mu,\mathrm{o}}(\mathsf{w})$. In this case, the statistical mechanical averages over configuration space appearing in the moment relations (15) can be replaced with averages over the time series, and the desired parameters can

TABLE I. Sequences S_μ contained in the MD data set.

μ	S_μ	S_μ	μ
1	GCTATATATATATATAGC	GCTAGATAGATAGATAGC	29
2	GCATTAATTAATTAATGC	GCGCGGGCGGGCGGGCGC	30
3	GCGCATGCATGCATGCGC	GCGTGGGTGGGTGGGTGC	31
4	GCCTAGCTAGCTAGCTGC	GCACTAACTAACTAACGC	32
5	GCCGCGCGCGCGCGCGGC	GCGCTGGCTGGCTGGCGC	33
6	GCGCCGGCCGGCCGGCGC	GCTATGTATGTATGTAGC	34
7	GCTACGTACGTACGTAGC	GCTGTGTGTGTGTGTGGC	35
8	GCGATCGATCGATCGAGC	GCGTTGGTTGGTTGGTGC	36
9	GCAAAAAAAAAAAAAAGC	AAACAATAAGAA	37
10	GCCGAGCGAGCGAGCGGC	AAAGAACAATAA	38
11	GCGAAGGAAGGAAGGAGC	AAATAACAAGAA	39
12	GCGTAGGTAGGTAGGTGC	GGGAGGTGGCGG	40
13	GCTGAGTGAGTGAGTGGC	GGGCGGAGGTGG	41
14	GCAGCAAGCAAGCAAGGC	GGGCGGTGGAGG	42
15	GCAAGAAAGAAAGAAAGC	GGGTGGAGGCGG	43
16	GCGAGGGAGGGAGGGAGC	GGGTGGCGGAGG	44
17	GCGGGGGGGGGGGGGGGC	AAATAAAAATAAGAACAA	45
18	GCAGTAAGTAAGTAAGGC	AAATAACAATAAGAACAA	46
19	GCGATGGATGGATGGAGC	GGGAGGGGGAGGCGGTGG	47
20	GCTCTGTCTGTCTGTCGC	GACATGGTACAG	48
21	GCACAAACAAACAAACGC	ACGATCCTAGCA	49
22	GCAGAGAGAGAGAGAGGC	ATGCTAATCGTA	50
23	GCGCAGGCAGGCAGGCGC	AGCTGAAGTCGA	51
24	GCTCAGTCAGTCAGTCGC	CGAACTTCAAGC	52
25	GCATCAATCAATCAATGC	GTCTACCATCTG	53
26	GCGTCGGTCGGTCGGTGC	GCATAAATAAATAAATGC	54
27	GCTGCGTGCGTGCGTGGC	GCATGAATGAATGAATGC	55
28	GCACGAACGAACGAACGC	GCGACGGACGGACGGAGC	56

then be estimated by the expressions

$$\widehat{\mathsf{w}}_{\mu,\mathrm{o}} := \frac{1}{L_\mu}\sum_{l=1}^{L_\mu}\mathsf{w}_\mu^{(l)}, \quad \mathsf{K}_{\mu,\mathrm{o}}^{-1} := \frac{1}{L_\mu}\sum_{l=1}^{L_\mu}\Delta\mathsf{w}_\mu^{(l)}\otimes\Delta\mathsf{w}_\mu^{(l)}, \tag{34}$$

where $\Delta\mathsf{w}_\mu := \mathsf{w}_\mu - \widehat{\mathsf{w}}_{\mu,\mathrm{o}}$.

While the averaging relations in (34) are entirely standard, they must nevertheless be treated with caution for at least three reasons. First, the time series may not be long enough for the ergodicity assumption to be a good approximation; a test of this is to consider palindromic oligomers and assess whether the estimates of $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$ and $\mathsf{K}_{\mu,\mathrm{o}}$ satisfy the requisite palindromic symmetries enjoyed by the exact values of these parameters. Second, our basic assumptions that the density $\rho_{\mu,\mathrm{o}}(\mathsf{w})$ and the associated time series $\mathsf{w}_\mu^{(l)}$ are Gaussian may not be a good approximation; a test of this is to construct the marginal distribution (or histogram) of each component of the internal coordinate vector from the time series and assess whether each marginal is indeed Gaussian. The results of both of these tests are discussed in Sec. VI for two representative oligomers S_μ, with more cases presented in the supplementary material,[66] and with the results for all of the oligomers of Table I being available at the site http://lcvmwww.epfl.ch/cgDNA. Third, the estimate for the inverse stiffness matrix, or covariance, given in (34) is known to be sensitive to outliers in the time series. There are a variety of ways of attempting a more robust estimate, but we here opt for simplicity and use only an explicit, physical criterion to exclude outliers as described next.

Our objective is to parameterize a coarse-grain, sequence-dependent, Gaussian model of B-form DNA. However, in a large ensemble of long MD simulations, there are inevitably events where the DNA departs from a configuration that could be described as B-form. Indeed, for each sequence, we observed that basepairs transiently broke and re-formed during the MD simulations. In particular, for the oligomers with non-GC ends, there were significant periods where the ends were frayed, although in most cases the double-helix did eventually reform. As we cannot hope to capture such phenomena with a Gaussian model, we filtered our time series; specifically, we eliminated from the averages in (34) all snapshots with one or more broken hydrogen bond anywhere along the oligomer. Following previous work,[43,58] we considered a hydrogen bond to be broken if the distance between donor and acceptor was greater than 4 Å. Our data suggest (see the supplementary material[66]) that the distribution of the donor-acceptor distance is close to Gaussian with a mean of approximately 3 Å and a standard deviation of approximately 0.1–0.2 Å, which is in agreement with *ab initio* calculations and high-resolution crystal structures.[78] Hence by using a threshold of 4 Å, or approximately five standard deviations above the mean, we exclude structures that are significantly outside the scope of our quadratic model. We remark that although they are deliberately excluded in the current treatment, frayed, melted, and other non-B-DNA structures are of course of considerable interest.[45–55,79–81]

Using (34) together with our filtering criterion, we obtained an observed shape (or mean) vector $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$, an observed stiffness matrix $\mathsf{K}_{\mu,\mathrm{o}}$, and an observed internal configuration density $\rho_{\mu,\mathrm{o}}(\mathsf{w})$ for each oligomer S_μ, $\mu = 1, \ldots, 53$ in Table I, which corresponds to 39 distinct 18-mers and 14 distinct 12-mers. After filtering, a total of over 2.1×10^6 snapshots or approximately 2.1 microseconds of MD simulation time remained in this data set. Specifically, for each of the above oligomers there remained an average of approximately 40 000 accepted configurations after filtering. We also initially considered the oligomers S_μ, $\mu = 54, 55, 56$, but found that there remained comparatively few snapshots after filtering. Hence, these latter oligomers were removed from the training data set and not considered further in our analysis.

E. Kullback-Leibler scale

To quantify various modeling errors it is convenient to set a scale for the Kullback-Leibler divergence introduced in (16). Moreover, since the divergence is a measure of the distance between two probability densities on the same configuration space, it is desirable to consider two separate scales: one for 18-mers and one for 12-mers. For 18-mers, we define a scale D_o as the average of $D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}})$ over all distinct pairs of 18-mer sequences S_{μ_1} and S_{μ_2} in the data set. As detailed in the supplementary material,[66] by direct computation using (17), we find

$$D_\mathrm{o} = \underset{\substack{n_\mu=18 \\ \mu_1\neq\mu_2}}{\mathrm{avg}}\; D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}) \doteq 85. \tag{35}$$

A similar calculation could be done to determine a scale for 12-mers. However, for simplicity, we just scale the above result by a factor of 2/3 to account for the smaller dimension of the configuration space and use the same symbol D_o. Hence for 12-mers, we use the scale $D_\mathrm{o} \doteq 57$. The appropriate value of D_o provides a characteristic scale for the divergence between the internal configuration densities of a pair of oligomers of given length in our data set that is due to the variation in the two sequences.

V. PARAMETER ESTIMATION

Here, we use our training data set to estimate, or fit, the parameters in our nearest-neighbor models. While our main interest is to fit the parameters for the nearest-neighbor model with dimer-based sequence dependence, as an intermediate step we first fit a nearest-neighbor model with oligomer-based sequence dependence to each oligomer in the data set. This procedure is adopted to better quantify the modeling errors in the two distinct modeling assumptions of nearest-neighbor interactions, and of dimer sequence dependence of the parameter set. The two-stage approach also enhances our numerical treatment of the fitting problem. In what follows, we use the subscripts M and m to denote quantities associated with the two nearest-neighbor models with, respectively, general oligomer-based sequence dependence, and dimer-based sequence dependence.

A. Oligomer-based fitting

In an oligomer-based model, each oligomer S_μ is described by a model probability density

$$\rho_{\mu,\mathrm{M}}(\mathsf{w}) = \frac{1}{Z_{\mu,\mathrm{M}}} e^{-(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{M}})\cdot\mathsf{K}_{\mu,\mathrm{M}}(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{M}})/2}, \tag{36}$$

where $\widehat{\mathsf{w}}_{\mu,\mathrm{M}}$ is the shape vector and $\mathsf{K}_{\mu,\mathrm{M}}$ is the symmetric, positive-definite stiffness matrix for the oligomer. As described in Sec. III B 1, the parameters $\widehat{\mathsf{w}}_{\mu,\mathrm{M}}$ and $\mathsf{K}_{\mu,\mathrm{M}}$ may depend upon the oligomer length and sequence in the most general way, subject only to the restriction that $\mathsf{K}_{\mu,\mathrm{M}}$ possess a specified sparsity pattern corresponding to the assumed type and range of interactions. We will be primarily concerned with the rigid-base nearest-neighbor model, with the stiffness matrix sparsity pattern illustrated in (S1), but we will also briefly discuss two other models with different assumed interactions: a smaller stencil which corresponds to a rigid-basepair nearest-neighbor model, and a larger stencil which corresponds to a rigid-base next-nearest-neighbor model.

For a specified sparsity pattern, a best-fit oligomer-based model for the oligomer S_μ in the training set is a density $\rho^*_{\mu,\mathrm{M}}$, with parameters $\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$ and $\mathsf{K}^*_{\mu,\mathrm{M}}$, which satisfies

$$\rho^*_{\mu,\mathrm{M}} = \underset{\rho_{\mu,\mathrm{M}}}{\mathrm{argmin}}\; D(\rho_{\mu,\mathrm{M}}, \rho_{\mu,\mathrm{o}}). \tag{37}$$

That is, among all model densities of the form (36) with a specified sparsity pattern, a best-fit density has a minimum divergence to the observed density in the training set. Using (37) and the explicit expression in (17) for the divergence between Gaussian densities, we find that the parameters associated with a best-fit density must satisfy the necessary conditions

$$\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}} = \widehat{\mathsf{w}}_{\mu,\mathrm{o}},$$
$$\mathsf{K}^*_{\mu,\mathrm{M}} = \underset{\mathsf{K}_{\mu,\mathrm{M}}}{\mathrm{argmin}}\; \frac{1}{2}\left[\mathsf{K}^{-1}_{\mu,\mathrm{M}}{:}\mathsf{K}_{\mu,\mathrm{o}} - \ln(\det \mathsf{K}_{\mu,\mathrm{o}}/\det \mathsf{K}_{\mu,\mathrm{M}}) - I{:}I\right], \tag{38}$$

where the minimum is taken over the set of symmetric matrices of the specified sparsity. We remark that the properties of the Kullback-Leibler divergence can be used to show that the above matrix optimization problem can also be regarded as that of finding a stiffness matrix $\mathsf{K}_{\mu,\mathrm{M}}$ of specified sparsity such that the associated covariance matrix $\mathsf{K}^{-1}_{\mu,\mathrm{M}}$ has a minimum distance, in an appropriate sense, to the observed covariance matrix $\mathsf{K}^{-1}_{\mu,\mathrm{o}}$ from the training set (see the supplementary material[66]). Since the functional in the minimization problem (38) is continuous and bounded below among positive-definite matrices and becomes unbounded above as a definite matrix approaches a semi-definite one, we expect that a minimum exists within the set of symmetric, positive-definite matrices of the specified sparsity. Indeed, using two different numerical procedures, a custom written numerical gradient flow and an implementation using the optimization code *Hanso*,[82,83] we have been able to find a minimum for

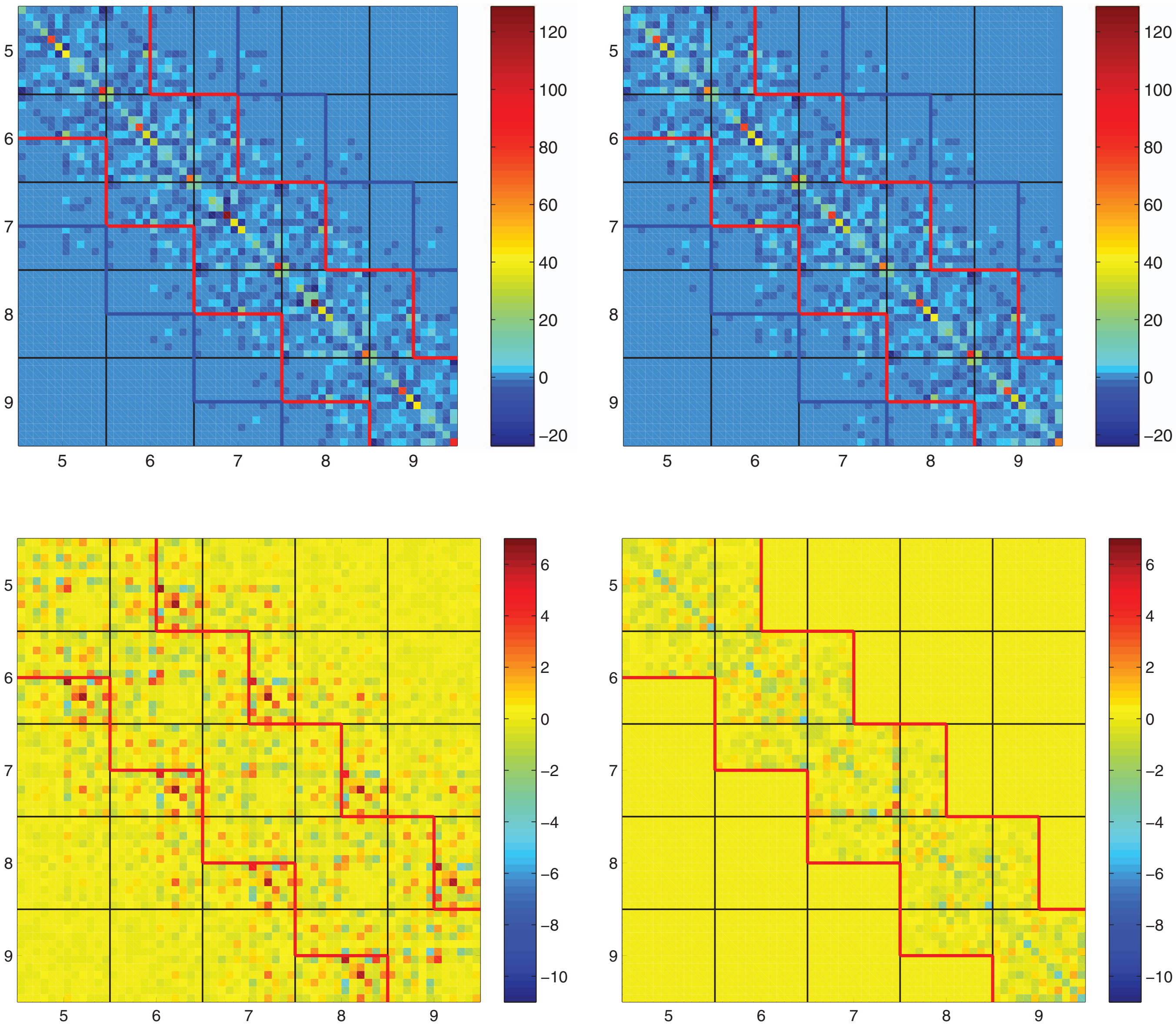


FIG. 1. Stiffness sub-matrices in dimensionless units for the training set oligomers S_1 and S_3. Black grid lines denote 12×12 blocks corresponding to intra- and inter-basepair coordinates (y^a, z^a) at positions $a = 5, \ldots, 9$ along the oligomer. Red and blue lines denote the sparsity stencils for a nearest-neighbor and next-nearest-neighbor, rigid-base free energy model, respectively. (Top left) sub-matrix of $\mathsf{K}_{3,\mathrm{o}}$. (Top right) sub-matrix of $\mathsf{K}_{1,\mathrm{o}}$. (Bottom left) sub-matrix of the difference $\mathsf{K}_{3,\mathrm{o}} - \mathsf{K}^*_{3,\mathrm{M}}$ for best-fit oligomer-based model with nearest-neighbor stencil. (Bottom right) sub-matrix of the difference $\mathsf{K}^*_{3,\mathrm{M}} - \mathsf{K}^*_{3,\mathrm{m}}$ for best-fit oligomer-based and dimer-based models with nearest-neighbor stencil; see Sec. V B. All entries outside the red stencil on the bottom right vanish by definition, while on the bottom left they are identical with the plot immediately above, but now the color scale is much different.

each oligomer in the training set for each of the three prescribed sparsity patterns. While such minima may only be local, our computations suggest that any possible multiple minima are rather isolated: for each oligomer and sparsity pattern, small changes to the initial condition did not change the location of the minima, and the outputs of the two different codes were extremely close in all cases.

Before presenting the results of our computations, we explain in more detail the three sparsity patterns or stencils that we consider. Figure 1 presents sub-matrices of the observed stiffness $\mathsf{K}_{\mu,\mathrm{o}}$ for the two training set oligomers S_μ with $\mu = 1, 3$. The qualitative features visible in this figure are similar for all sub-matrices over all oligomers in the training set (see http://lcvmwww.epfl.ch/cgDNA). Specifically, all of the largest entries lie within a 6×6 block-diagonal stencil (not explicitly marked). The inter-basepair portion of such a stencil is associated with a rigid-basepair free energy model that is an entirely local function of the inter-basepair coordinates at each junction: each set of junction variables are decoupled from all other configuration variables, either inter- or intra-basepair. Such a model is rather standard in much of the literature on the coarse-graining of DNA. Nevertheless, as shown in the figure, it is evident that considerably more of the signal in the observed stiffness matrices $\mathsf{K}_{\mu,\mathrm{o}}$ lie within the overlapping 18×18 nearest-neighbor, rigid-base stencil marked in red, and yet more within the overlapping 30×30 next-nearest-neighbor, rigid-base stencil marked in blue. We remark that the next-nearest-neighbor stencil still excludes

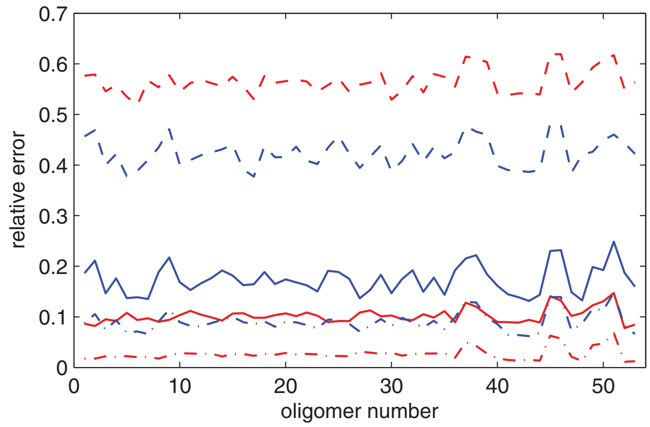


FIG. 2. Relative error in densities and stiffnesses versus oligomer number between the best-fit oligomer-based model with three different sparsity patterns and the observed quantities. (Red curves) relative error in densities $D(\rho^*_{\mu,\mathrm{M}}, \rho_{\mu,\mathrm{o}})/D_\mathrm{o}$. (Blue curves) relative error in stiffness matrices $||\mathsf{K}^*_{\mu,\mathrm{M}} - \mathsf{K}_{\mu,\mathrm{o}}||/||\mathsf{K}_{\mu,\mathrm{o}}||$. (Dashed curves) 6 × 6 block-diagonal stencil. (Solid curves) overlapping 18 × 18 stencil. (Dashed-dotted curves) overlapping 30 × 30 stencil. $||\cdot||$ denotes the Frobenius norm.

entries of the same magnitude to those that it captures beyond the nearest-neighbor stencil. This observation suggests that it would be interesting to consider an even longer range interaction model, but we do not pursue that line of investigation here.

Figure 2 presents an assessment, for each of the three sparsity stencils, of the relative error between the best-fit oligomer-based model density $\rho^*_{\mu,\mathrm{M}}$ and the observed density $\rho_{\mu,\mathrm{o}}$ for each of the training set oligomers S_μ, $\mu = 1, \ldots, 53$. Specifically, for each stencil and oligomer, the Kullback-Leibler divergence between $\rho^*_{\mu,\mathrm{M}}$ and $\rho_{\mu,\mathrm{o}}$ is shown, scaled into a relative divergence using the appropriate value of D_o as introduced in Sec. IV E. Because it is only differences in stiffness matrices that contribute to the divergences in this case, we also plot the relative error, in the Frobenius norm, between the best-fit stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{M}}$ and the observed stiffness matrix $\mathsf{K}_{\mu,\mathrm{o}}$ for each stencil and oligomer. For the 6 × 6 block-diagonal stencil which contains the nearest-neighbor, rigid-basepair model, the relative error in the densities is around 55% for each oligomer, whereas the relative error in the stiffness matrices is lower, around 45%. For the overlapping 18 × 18 stencil associated with a nearest-neighbor, rigid-base model, the relative error in the densities is around 10% for each oligomer, whereas the relative error in the stiffness matrices is now larger, around 20%. For reference, a portion of the difference matrix $\mathsf{K}_{\mu,\mathrm{o}} - \mathsf{K}^*_{\mu,\mathrm{M}}$ for the 18 × 18 stencil is shown in Figure 1 (bottom left panel) for the oligomer S_μ with $\mu = 3$. For the overlapping 30 × 30 stencil associated with a next-nearest-neighbor rigid-base model, the relative error in the densities is less than 5% for nearly all oligomers, whereas the relative error in the stiffness matrices is around 10%.

The above results suggest, as would be expected, that oligomer-based free energy models with longer range interactions provide better fits of the training set data than models with shorter range interactions. In particular, the 6 × 6 block-diagonal stencil provides a rather poor fit of the data, whereas the overlapping 18 × 18 and 30 × 30 stencils provide increasingly better fits. Based on these results, we can now confirm quantitatively for all of the oligomers S_μ, $\mu = 1, \ldots, 53$ in our training set an observation first made qualitatively in previous work[43] for one oligomer, namely, that a nearest-neighbor, rigid-basepair model with oligomer sequence dependence, as is reflected in the 6 × 6 block-diagonal stencil, is in rather poor agreement with MD simulations at the scale of tens of basepairs, while a nearest-neighbor, rigid-base model with oligomer sequence dependence, as is reflected in the overlapping 18 × 18 stencil, is in reasonably good agreement with MD simulations. The robustness of this conclusion has been further confirmed in a contemporaneous analysis[42] of the Dickerson dodecamer, which describes multiple independent instances of multiple microsecond simulations under a range of MD protocols, and uses 3DNA[58] rather than Curves+[68] coordinates for the rigid bases.

B. Dimer-based fitting

The primary objective of this article is to move beyond coarse-grain oligomer-based fitting of MD simulations, to the construction of a finite coarse-grain parameter set that will allow the prediction of equilibrium distributions for oligomers of arbitrary sequence, with no further MD simulation required. To that end, and in the first instance, we make a compromise between the quality of the possible fit to oligomer-based observations as limited by the assumed stencil size, and the complexity of the associated model parameter set, and will hereafter consider only the overlapping 18 × 18 stencil associated with a nearest-neighbor rigid-base model, with model parameters that depend only on the dimer sequence context.

In the nearest-neighbor, dimer-based model, each oligomer S_μ is described by a model probability density

$$\rho^{\mathcal{P}}_{\mu,\mathrm{m}}(\mathsf{w}) = \frac{1}{Z_{\mu,\mathrm{m}}} e^{-(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{m}})\cdot\mathsf{K}_{\mu,\mathrm{m}}(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{m}})/2}, \tag{39}$$

where $\widehat{\mathsf{w}}_{\mu,\mathrm{m}}$ is the shape vector and $\mathsf{K}_{\mu,\mathrm{m}}$ is the symmetric, positive-definite stiffness matrix for the oligomer. As described in Sec. III B 2, the vector $\widehat{\mathsf{w}}_{\mu,\mathrm{m}}$ and the matrix $\mathsf{K}_{\mu,\mathrm{m}}$ are constructed from a finite, dimer-based parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$. By a best-fit parameter set for our collection of training set oligomers S_μ we mean a set $\mathcal{P}^*$ satisfying

$$\mathcal{P}^* = \underset{\mathcal{P}}{\operatorname{argmin}} \sum_{\mu=1}^{53} D(\rho^{\mathcal{P}}_{\mu,\mathrm{m}}, \rho^*_{\mu,\mathrm{M}}). \tag{40}$$

That is, among all parameter sets, a best-fit parameter set should minimize the sum of the divergences between the dimer-based constructed densities and the best-fit oligomer-based densities with the nearest-neighbor sparsity pattern. Presumably, if the collection of training set oligomers is sufficiently rich, such a best-fit parameter set should not only provide a good description of the training set oligomers, but also any other oligomer of arbitrary length and sequence.

Notice that the fitting problem for the dimer-based model is posed in terms of the best-fit oligomer-based densities $\rho^*_{\mu,\mathrm{M}}$, with the prescribed nearest-neighbor sparsity pattern, rather than directly in terms of the observed densities $\rho_{\mu,\mathrm{o}}$ from the training set. We believe that this choice leads to a better separation of modeling errors, and a better numerical treatment of the fitting problem. First, we can quantify modeling errors due to the dimer-based assumption on the level of sequence dependence, independent of the nearest-neighbor assumption on the level of energetic coupling. Second, we can generate better initial guesses for the numerical treatment of the minimization problem in (40) because $\mathsf{K}_{\mu,\mathrm{m}}$ and $\mathsf{K}^*_{\mu,\mathrm{M}}$ have the same nearest-neighbor sparsity structure, and direct comparisons between these two matrices can be made. Specifically, a numerical gradient flow procedure was used to treat the minimization problem, and the initial guess for a best-fit parameter set plays an important role in the overall success of the procedure.

To generate an initial guess for a best-fit parameter set, we developed an approximate solution of (40) using a least-squares approach. For motivation, notice that the sum of the divergences in (40) achieves its lowest possible value of zero when the constructed stiffness matrix $\mathsf{K}_{\mu,\mathrm{m}}$ and shape vector $\mathsf{w}_{\mu,\mathrm{m}}$ are equal to the given stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{M}}$ and shape vector $\mathsf{w}^*_{\mu,\mathrm{M}}$ for each oligomer S_μ, $\mu = 1, \ldots, 53$ in the training set. Equality between these matrices and vectors in general cannot be achieved due in part to the relatively small size of the parameter set and the large size of the training set, and in part to fundamental differences between the dimer- and oligomer-based models. Working with weighted shape vectors for convenience, it is thus reasonable to seek a parameter set that satisfies, in a least-squares sense, the over-determined system of linear equations

$$\left.\begin{aligned} \mathsf{K}_{\mu,\mathrm{m}} &= \mathsf{K}^*_{\mu,\mathrm{M}} \\ \sigma_{\mu,\mathrm{m}} &= \sigma^*_{\mu,\mathrm{M}} \end{aligned}\right\}, \quad \mu = 1, \ldots, 53. \tag{41}$$

The matrices $\mathsf{K}_{\mu,\mathrm{m}}$ and vectors $\sigma_{\mu,\mathrm{m}}$ on the left-hand side of the above equation are explicit functions of the unknown parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ as defined in (31) and (32). The matrices $\mathsf{K}^*_{\mu,\mathrm{M}}$ and vectors $\sigma^*_{\mu,\mathrm{M}}$ on the right-hand side of the above equation are known data that have been determined, or can be obtained from the oligomer-based fit.

A procedure was developed to construct a least-squares solution of (41) (see the supplementary material[66]) and thereby obtain an admissible parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ to be used as an initial guess in our numerical minimization of the nonlinear, Kullback-Leibler objective functional in (40). In our numerical minimization of this functional, we employed a numerical gradient flow procedure in which the semi-positive-definiteness conditions on the stiffness parameters and the objectivity conditions on all parameters as discussed in Sec. III B 2 were explicitly enforced, including the self-symmetry conditions associated with palindromic dimers. Using this procedure, we were able to numerically minimize the functional in (40) and thereby obtain an admissible, best-fit parameter set $\mathcal{P}^*$. As can be expected due to the high dimensionality and nonlinear nature of the problem, different choices of initial guess lead to different numerically computed minima. Indeed, our computations suggest that this minimization problem is rather delicate and worthy of further study. Throughout the remainder of our developments, we describe properties of one specific choice of a best-fit parameter set $\mathcal{P}^*$, which seems typical of the optimal approximations found thus far.

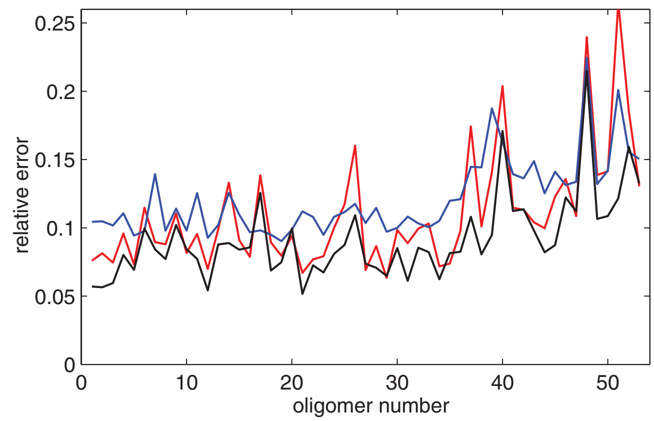


FIG. 3. Relative error in densities, stiffnesses, and shapes versus oligomer number between the best-fit dimer-based model and the best-fit oligomer-based model with nearest-neighbor sparsity. (Red curves) relative error in densities $D(\rho^*_{\mu,\mathrm{m}}, \rho^*_{\mu,\mathrm{M}})/D_\mathrm{o}$. (Blue curves) relative error in stiffness matrices $\|\mathsf{K}^*_{\mu,\mathrm{m}} - \mathsf{K}^*_{\mu,\mathrm{M}}\|/\|\mathsf{K}^*_{\mu,\mathrm{M}}\|$. (Black curves) relative error in shape vectors $|\mathsf{w}^*_{\mu,\mathrm{m}} - \mathsf{w}^*_{\mu,\mathrm{M}}|/|\mathsf{w}^*_{\mu,\mathrm{M}}|$. $\|\cdot\|$ and $|\cdot|$ denote the Frobenius and Euclidean norms, respectively.

Figure 3 presents an assessment of the modeling errors incurred in approximating an oligomer-based model density $\rho^*_{\mu,\mathrm{M}}$ by a dimer-based model density $\rho^*_{\mu,\mathrm{m}}$ constructed with our best-fit parameter set $\mathcal{P}^*$ for each of the training set oligomers S_μ, $\mu = 1, \ldots, 53$. Here, the modeling error for each oligomer is due to differences between the stiffness matrices $\mathsf{K}^*_{\mu,\mathrm{M}}$ and $\mathsf{K}^*_{\mu,\mathrm{m}}$, and the shape vectors $\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$ and $\widehat{\mathsf{w}}^*_{\mu,\mathrm{m}}$, and both of these differences arise due to the finiteness of the parameter set $\mathcal{P}^*$. In physical terms, this modeling error reflects the difference between a quadratic, nearest-neighbor free energy model in which the parameters of the model are allowed to depend on the sequence composition of the entire oligomer, and one in which the parameters depend only on the composition of the local dimer. As can be seen, the relative errors between the internal configuration densities, stiffness matrices, and shape vectors are all in the range $5\% - 15\%$ for the oligomers with $\mu = 1, \ldots, 36$, with slightly higher errors for the oligomers with $\mu = 37, \ldots, 53$. One possible explanation for these higher errors is that this latter set of oligomers all have non-GC dimer ends, which are much less represented in the training data set. Nevertheless, our conclusion is that the dimer-based model with our best-fit parameter set $\mathcal{P}^*$ is able to well resolve sequence variations across all the training set oligomers. Example model constructions, for oligomers within and outside of the training set, are discussed in Secs. VI and VII below.

C. The $\mathcal{P}^*$ parameter set

The full numerical data for our best-fit parameter set $\mathcal{P}^* = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ is provided in the supplementary material,[66] including electronic files for download and a visual presentation, and MATLAB scripts for the construction of the shape vector and stiffness matrix for any given oligomer of arbitrary length and sequence. A related and more comprehensive set of MATLAB scripts for use with our model and parameter set is also available at the site http://lcvmwww.epfl.ch/cgDNA.

The parameter set for our model comprises the 1-mer stiffness parameter matrices $\mathsf{K}_1^\alpha \in \mathbb{R}^{6\times 6}$ and weighted shape parameter vectors $\sigma_1^\alpha \in \mathbb{R}^6$, along with the 2-mer stiffness parameter matrices $\mathsf{K}_2^{\alpha\beta} \in \mathbb{R}^{18\times 18}$ and weighted shape parameter vectors $\sigma_2^{\alpha\beta} \in \mathbb{R}^{18}$. Strong variations between the two independent sets of 1-mer parameters $\{\sigma_1^\alpha, \mathsf{K}_1^\alpha\}$ suggests that differences in the intra-basepair interactions within the two independent basepairs (monomers) are being captured. Similarly, variations in the ten independent sets of 2-mer parameters $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ suggests that sequence-dependent differences in the inter-basepair, e.g., stacking, interactions within the ten independent dimer steps are also being captured. In addition to notable differences between all ten independent 2-mer parameter sets, there are also striking similarities within the three independent families of parameters for purine-pyrimidine, pyrimidine-purine, and purine-purine dimer steps.

We remark that inspection of the eigenvalues of the 1-mer and 2-mer stiffness parameter matrices reveals a number of rather small eigenvalues that could reasonably be approximated as zero. However, the small eigenvalues for the stiffness parameter matrices do not translate into correspondingly small eigenvalues for any oligomer stiffness matrix of length two or more assembled from the parameter stiffness matrices. Although some of the individual 1-mer and 2-mer interaction energies have rather soft modes, they always stabilize each other when superposed to yield an oligomer energy that is relatively stiff.

Differences in the precise definitions of the intra- and inter-basepair coordinates employed by various authors make a detailed comparison of sequence-dependent model parameter values somewhat complicated. Nevertheless, for purposes of comparison with other results in the literature, it is of interest to consider the expected, or average, configuration for a sequence-averaged homogeneous oligomer. One of various ways to do this is to make a reconstruction using a sequence-averaged best-fit parameter set $\mathcal{P}^{*,\mathrm{avg}} = \{\sigma_1^{\mathrm{avg}}, \mathsf{K}_1^{\mathrm{avg}}, \sigma_2^{\mathrm{avg}}, \mathsf{K}_2^{\mathrm{avg}}\}$ obtained via Euclidean averaging of the 1-mer and 2-mer parameters over all 4 possible monomers and 16 possible dimers, respectively. The set $\mathcal{P}^{*,\mathrm{avg}}$ can be interpreted as providing a homogeneous, nearest-neighbor model of DNA in which the occurrence of each of the four possible basepairs is assumed to be equally likely at each position in an oligomer. Using this homogeneous parameter set, a model shape vector $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$ and stiffness matrix $\mathsf{K}_{\mathrm{h,m}}^*$ can then be reconstructed for a homogeneous oligomer of arbitrary length. Despite the fact that the oligomer is homogeneous, and the input parameters are constant, there are, as detailed in the supplementary material,[66] significant end effects in the model reconstruction of the ground state $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$. For some shape parameters, for example propeller, visible end effects penetrate to a depth of 4 basepairs from either end.

For sufficiently long oligomers, the entries of the shape $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$ and stiffness $\mathsf{K}_{\mathrm{h,m}}^*$ parameters approach constant values away from the ends. The constant, interior values of $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$ are shown in Table II, which also contains comparable data reported in two other sources, namely, sequence-averaged values computed directly from the original atomistic MD ABC database,[21] and from crystal structures of DNA-protein complexes.[59] To make meaningful comparisons, we also provide our constant, interior homogeneous shape parameters in both dimensional Curves+ coordinates[68] and dimensional 3DNA coordinates.[58] We remark that the computation of a 3DNA version of our results is not straightforward, and involves various choices as discussed in the supplementary material.[66] We note that the entries in the second, third, and fourth columns of Table II, which are all based ultimately on MD and are expressed in either Curves+ or 3DNA coordinates, are in reasonable agreement. The fifth column, which contains inter-basepair data from crystal structures expressed in 3DNA coordinates, has noticeable differences compared to the previous three columns: the value of roll is smaller, twist is larger, and slide is of the opposite sign.

TABLE II. Comparison of sequence-averaged, homogeneous coarse-grain ground-state shapes away from ends. (First column) homogeneous shape in dimensionless units obtained from a reconstruction using the sequence-averaged best-fit parameter set $\mathcal{P}^{*,\mathrm{avg}}$ of the dimer-based model of this article. (Second column) same as first, but in Curves+[68] dimensional units of degrees and Angstroms. (Third column) sequence-averaged shape obtained directly from averaging MD simulation data; Curves+ coordinates taken from Table 1 of Ref. 21. (Fourth column) same as first, but expressed in 3DNA[58] coordinates. (Fifth column) sequence-averaged inter-basepair shape obtained from DNA crystal structure data; 3DNA coordinates taken from Table 1 of Ref. 59.

	1 Dimensionless	2 ° or Å	3 ° or Å	4 ° or Å	5 ° or Å
Buckle	0	0	1.2	0.0	…
Propeller	− 1.09	− 12.4	− 11.0	− 12.5	…
Opening	0.11	1.2	2.1	− 0.8	…
Shear	0	0	0.02	0.00	…
Stretch	0.02	0.02	0.03	-0.03	…
Stagger	0.17	0.17	0.09	0.15	…
Tilt	0	0	− 0.3	0.0	0
Roll	0.26	2.9	3.6	2.9	1.4
Twist	2.96	33.0	32.6	33.0	35.4
Shift	0	0	− 0.05	0.00	0
Slide	− 0.56	− 0.56	− 0.44	− 0.62	0.35
Rise	3.31	3.31	3.32	3.32	3.32

In the sequence-independent, homogeneous case and away from the ends, we note that there are coarse-grain quantities, defined independently of any specific coordinates, that can be compared without ambiguity. Specifically, away from the ends, the reference points of the bases along each strand in a homogeneous oligomer should lie on a circular helix, the geometric properties of which are independent of any choice of coordinates. Using our sequence-averaged parameters, we find that there are 10.9 basepairs in a complete revolution of this helix. Remarkably, this value of the helical repeat is within 5% of the experimentally reported value of 10.4 ± 0.1.[84] Other geometric properties of this helix can also be computed. For example, for the pitch or vertical rise we obtain a value of 35.4 Å per revolution, or equivalently 3.25 Å per basepair, and for the radius (traced out by the reference points) we obtain a value of 1.49 Å. From this and the comparisons in Table II, we conclude that our best-fit parameter set is consistent, in the sense of sequence-averaged values, with accepted properties of B-form DNA.

VI. EXAMPLE OLIGOMERS FROM THE TRAINING SET

To illustrate the accuracy of the sequence dependence of our dimer-based model, we used it to reconstruct the shape vector $\widehat{\mathsf{w}}^*_{\mu,\mathrm{m}}$, stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{m}}$, and density $\rho^*_{\mu,\mathrm{m}}$ for each of the training set oligomers S_μ listed in Table I. For brevity, we only present results for the two oligomers S_1 and S_8 in the main text, with the additional cases S_3 and S_{42} in the supplementary material,[66] and results for all the oligomers online at http://lcvmwww.epfl.ch/cgDNA. Specifically, for each oligomer we used our model and best-fit parameter set to construct $\widehat{\mathsf{w}}^*_{\mu,\mathrm{m}}$, $\mathsf{K}^*_{\mu,\mathrm{m}}$, and $\rho^*_{\mu,\mathrm{m}}$, and compare these to the analogous quantities $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$, $\mathsf{K}_{\mu,\mathrm{o}}$, and $\rho_{\mu,\mathrm{o}}$ observed in MD simulation. Additionally, we also make comparison to the analogous quantities $\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$, $\mathsf{K}^*_{\mu,\mathrm{M}}$, and $\rho^*_{\mu,\mathrm{M}}$ of the oligomer-based model.

Figures 4 and 5 show entries of the shape vector and stiffness matrix as a function of position along the oligomers S_1 and S_8, respectively. The top four panels in each figure show the entries of the observed shape parameter vector $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$ (which are the MD time series averages) in solid lines, and the constructed dimer-based model shape parameter vector $\widehat{\mathsf{w}}^*_{\mu,\mathrm{m}}$ in dashed lines, each versus sequence position. In each figure, the base sequence of the reference strand of the oligomer is indicated on the abscissa, with intra-basepair parameter values indicated at each basepair, and inter-basepair parameter values indicated at each junction; for clarity these discrete values are interpolated by piecewise linear curves. The bottom four panels in each figure are analogous and show the diagonal entries of the observed stiffness matrix $\mathsf{K}_{\mu,\mathrm{o}}$ in solid lines, and of the constructed dimer-based model stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{m}}$ in dashed lines. For further comparison, the diagonal entries of the nearest-neighbor oligomer-based model stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{M}}$ are also indicated in dashed-dotted lines. Although, as illustrated in Figure 1, the stiffness matrices have many nonzero entries, we choose for brevity to plot only the diagonal entries.

The data in Figures 4 and 5 illustrate the rather high quality of the dimer-based model constructions. Frequently, the differences between the observed and the constructed quantities are indistinguishable, and with very few exceptions, the pointwise differences in the quantities are less than the variation with sequence. Visually, the errors in the intra-basepair shape and stiffness parameters appear larger, but the scales in the plots of the intra- and inter-basepair parameters are of necessity different, although the units are identical. For both intra- and inter-basepair shape parameters, rather few errors are larger than 0.1 Å in translational variables and 2° in rotational variables. All constructed parameters shown for oligomers S_1 and S_8 are visibly consistent with the periodicity of their interior sequences. By design, the constructed parameters exactly satisfy the requisite symmetry conditions for the palindromic oligomer S_1. The observed parameters computed directly from the MD time series data (shown in the solid lines) for the most part also closely satisfy the requisite symmetries, but errors can arise from a lack of convergence of the MD simulation of the relevant oligomer. For example, the breaking of evenness in the plot of the observed shape parameter stagger, and of the observed stiffness parameter twist-twist in Figure 4 violates the palindromic symmetry of oligomer S_1, and must reflect a lack of convergence of the MD time series. Results for other sequences (see the supplementary material[66]) indicate that there is a tendency for the constructed quantities to exhibit larger errors at the ends than the interior, which may reflect the fact that all possible dimer ends are not equally represented in our training data set.

Figures 6–8 show various one-dimensional marginal distributions (or histograms) for each type of intra- and inter-basepair coordinate at each location along the two oligomers S_1 and S_8. These marginal distributions provide a way to assess the quality of the Gaussian assumption in our modeling approach and further illustrate sequence and end effects. For each type of coordinate, at each location along each oligomer S_μ, we compare four different marginal distributions: the actual distribution obtained directly from the MD data (solid lines), and the three fitted distributions associated with the training set density $\rho_{\mu,\mathrm{o}}$ (dotted lines), the oligomer-based model density $\rho^*_{\mu,\mathrm{M}}$ (dashed-dotted lines), and the constructed dimer-based model density $\rho^*_{\mu,\mathrm{m}}$ (dashed lines). Whereas the marginal distribution obtained from the MD data could in principle be far from Gaussian, the marginal distribution associated with each model fit must necessarily be Gaussian because the model density is itself a (high-dimensional) Gaussian. For the intra-basepair coordinates considered in Figure 6, the monomer at each position on the reference strand is marked in each panel. For the inter-basepair coordinates considered in Figures 7 and 8, the dimer at each junction on the reference strand is marked in each panel. The middle rows of panels in the figures reflect the repeating sub-sequence in the interiors of the oligomers, whereas the middle columns of panels reflect the different recurring monomers or dimers within each sub-sequence; similarities in the marginals due to periodicity and differences due to sequence dependence are quite apparent.

Figure 6 shows the marginal distributions for the intra-basepair coordinates along oligomer S_8. Notice that the four distributions for each coordinate at each position are practically indistinguishable. This shows that each of the three Gaussian densities $\rho_{\mu,\mathrm{o}}$, $\rho^*_{\mu,\mathrm{M}}$, and $\rho^*_{\mu,\mathrm{m}}$ can adequately represent the actual distribution of each intra-basepair coordinate along the oligomer. Similar results and conclusions hold for the distributions of intra-basepair coordinates along oligomer S_1; the four distributions at each position on this oligomer are in even better agreement than they are for oligomer S_8 and are accordingly not shown. Figures 7 and 8 provide analogous plots for the inter-basepair coordinates. Now it can be seen that there are cases where the actual marginal distribution obtained from the MD data is noticeably non-Gaussian, even away from the ends. The marginals of slide for the various TA dimers in Figure 7, and the marginals of twist for the various CG dimers in Figure 8 are among the most non-Gaussian cases (see the supplementary material[66] for further examples). The bi-modal properties of the CG dimer have been

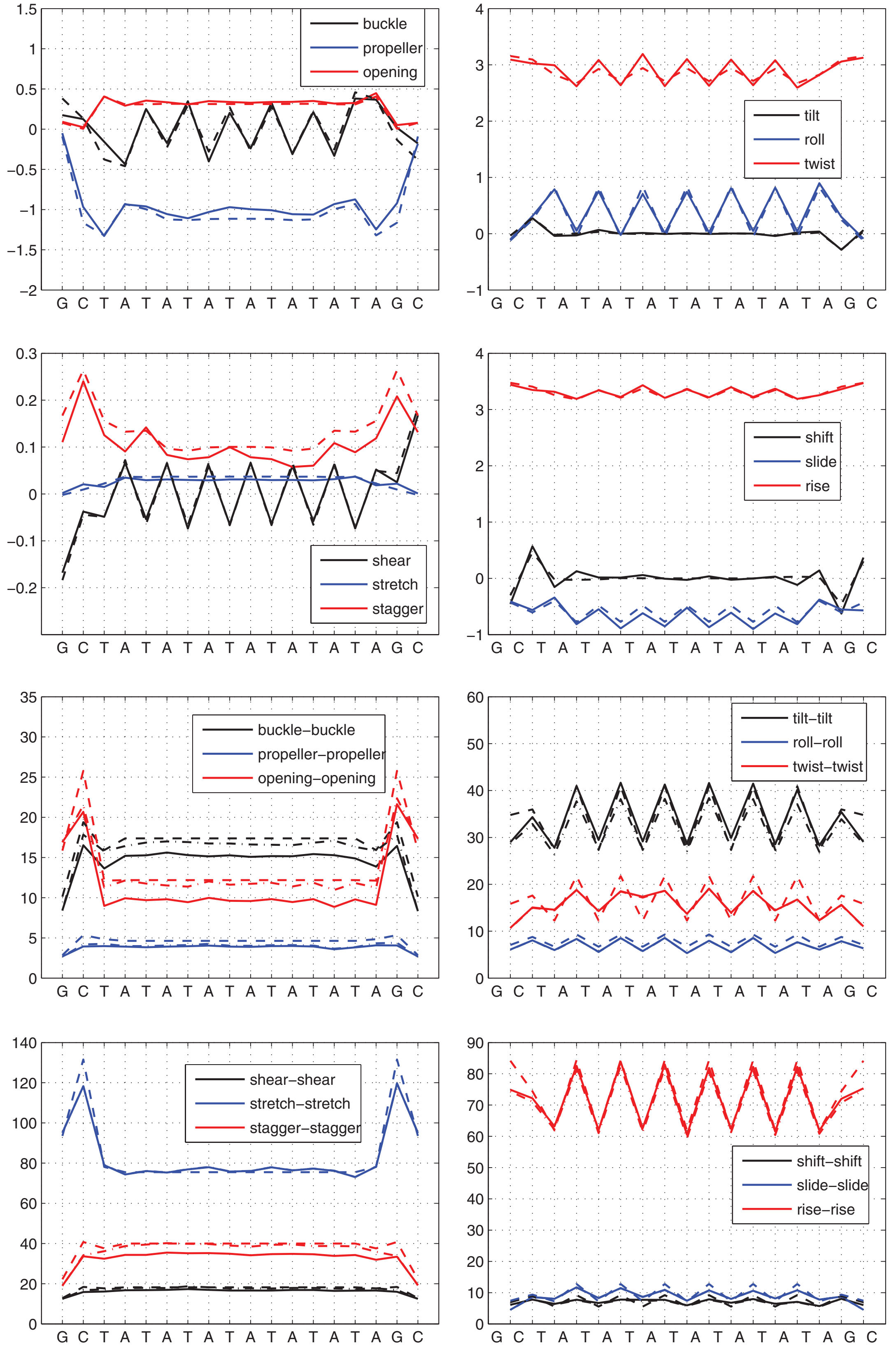


FIG. 4. Entries of shape vectors and stiffness matrices in dimensionless units for the palindromic, interior period two, 18-mer S_1 from the training set. (Top four panels) entries of observed vector $\widehat{\mathsf{w}}_{1,o}$ (solid) and constructed dimer-based model vector $\widehat{\mathsf{w}}^*_{1,m}$ (dashed). (Bottom four panels) diagonal entries of observed matrix $\mathsf{K}_{1,o}$ (solid), constructed dimer-based model matrix $\mathsf{K}^*_{1,m}$ (dashed) and nearest-neighbor oligomer-based model matrix $\mathsf{K}^*_{1,M}$ (dashed-dotted).

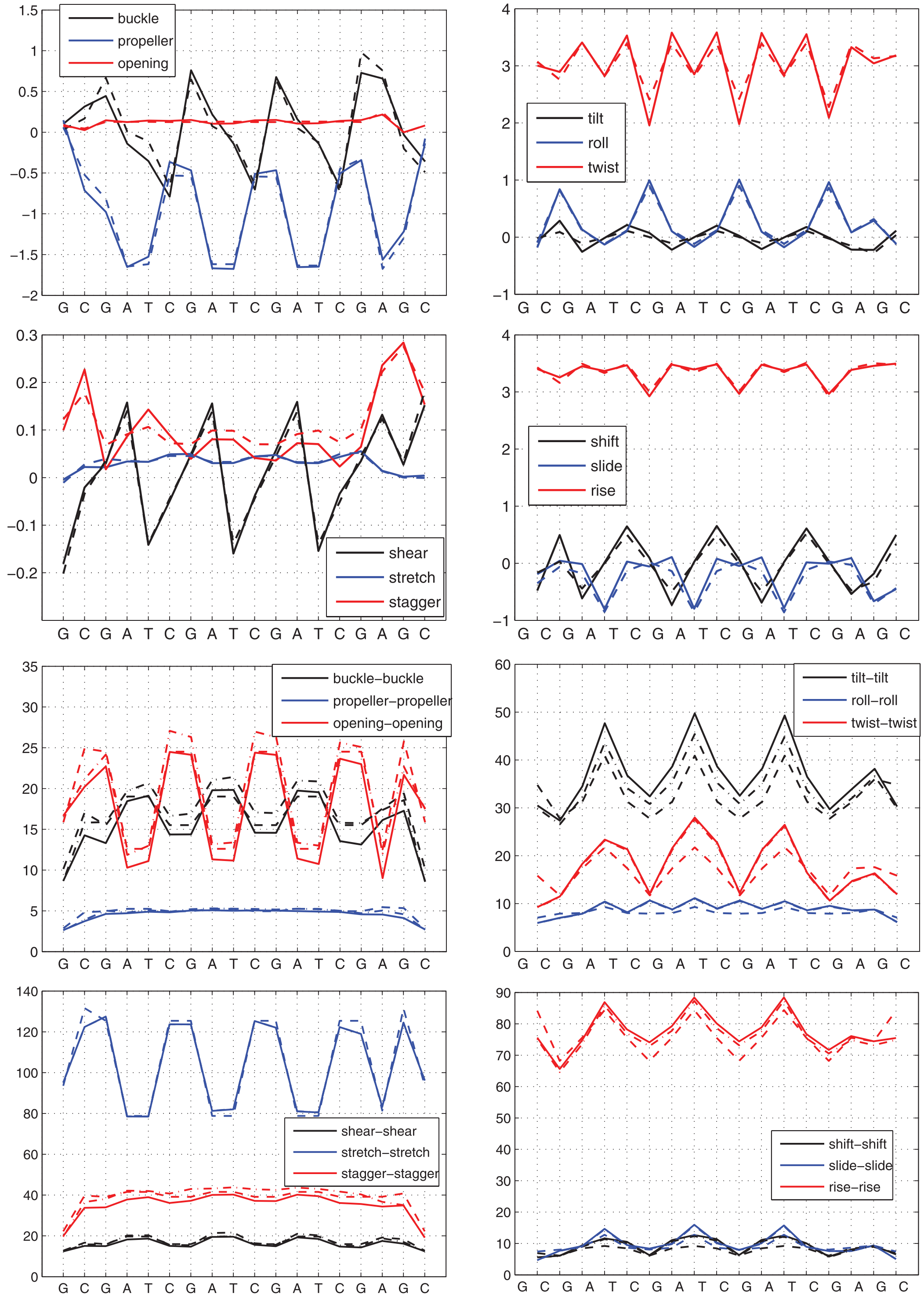


FIG. 5. Entries of shape vectors and stiffness matrices in dimensionless units for the non-palindromic, interior period four, 18-mer S_8 from the training set. (Top four panels) entries of observed vector $\widehat{\mathsf{w}}_{8,\mathrm{o}}$ (solid) and constructed dimer-based model vector $\widehat{\mathsf{w}}^{*}_{8,\mathrm{m}}$ (dashed). (Bottom four panels) diagonal entries of observed matrix $\mathsf{K}_{8,\mathrm{o}}$ (solid), constructed dimer-based model matrix $\mathsf{K}^{*}_{8,\mathrm{m}}$ (dashed), and nearest-neighbor oligomer-based model matrix $\mathsf{K}^{*}_{8,\mathrm{M}}$ (dashed-dotted).

noted previously.[21] While such bi-modal and otherwise non-Gaussian behavior is beyond the scope of the Gaussian approach considered here, the results show that the dimer-based model with the best-fit parameter set can capture the dominant features of sequence variation in a satisfactory way. Specifically, when comparing the dimer-based model construction to the MD data, the errors in both the mean and half-width of the constructed marginal of any coordinate can be seen to be almost always qualitatively smaller than the variation in these quantities due to sequence.

VII. AN EXAMPLE OLIGOMER NOT FROM THE TRAINING SET

To further illustrate the accuracy and discriminatory resolution of the sequence dependence of our dimer-based model,

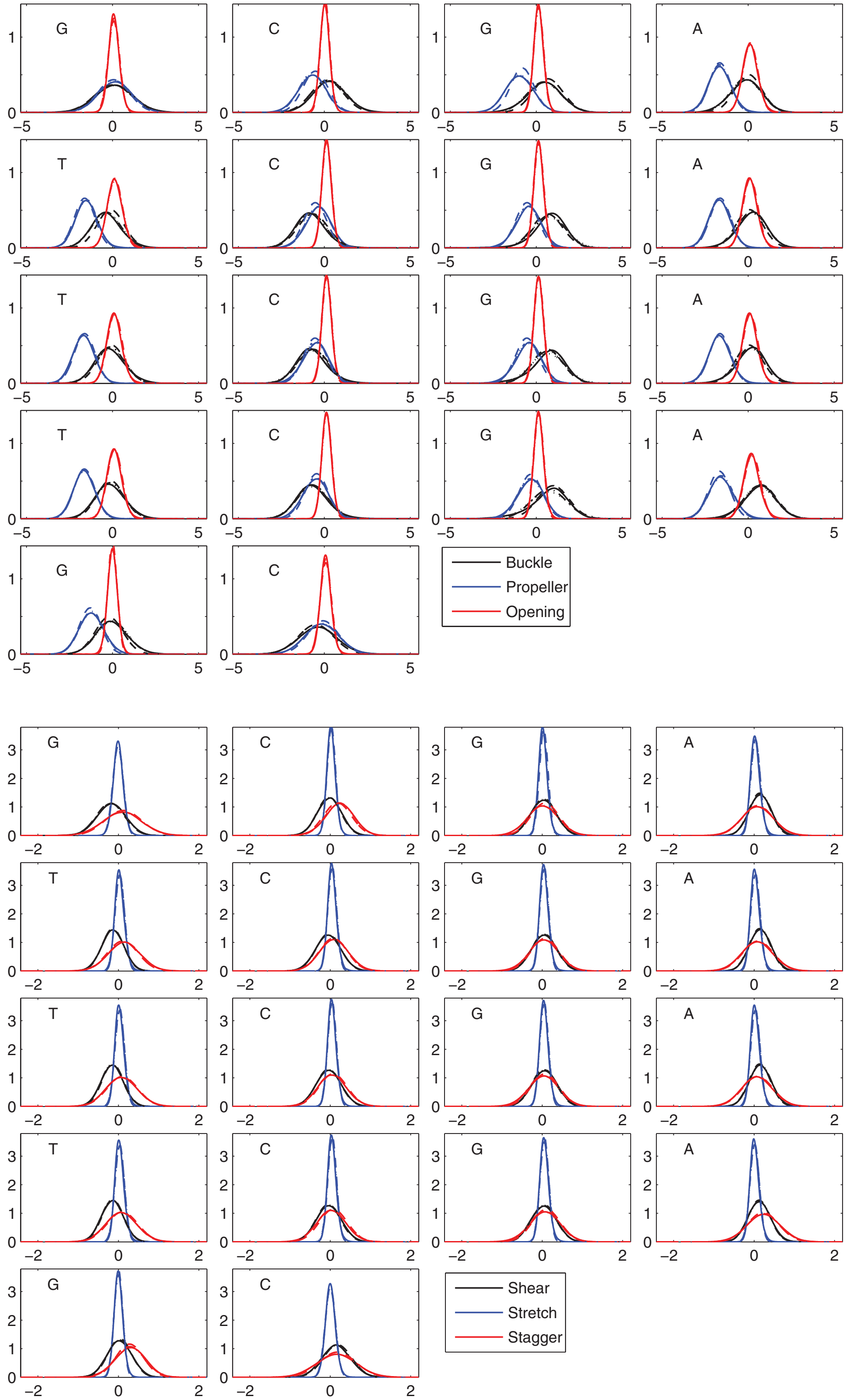


FIG. 6. Normalized marginal distributions for intra-basepair coordinates at each position along oligomer S_8. Positions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each position shows the monomer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{8,o}$, dotted; density $\rho^*_{8,M}$, dashed-dotted; density $\rho^*_{8,m}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units. The marginals from the different sources are virtually indistinguishable.

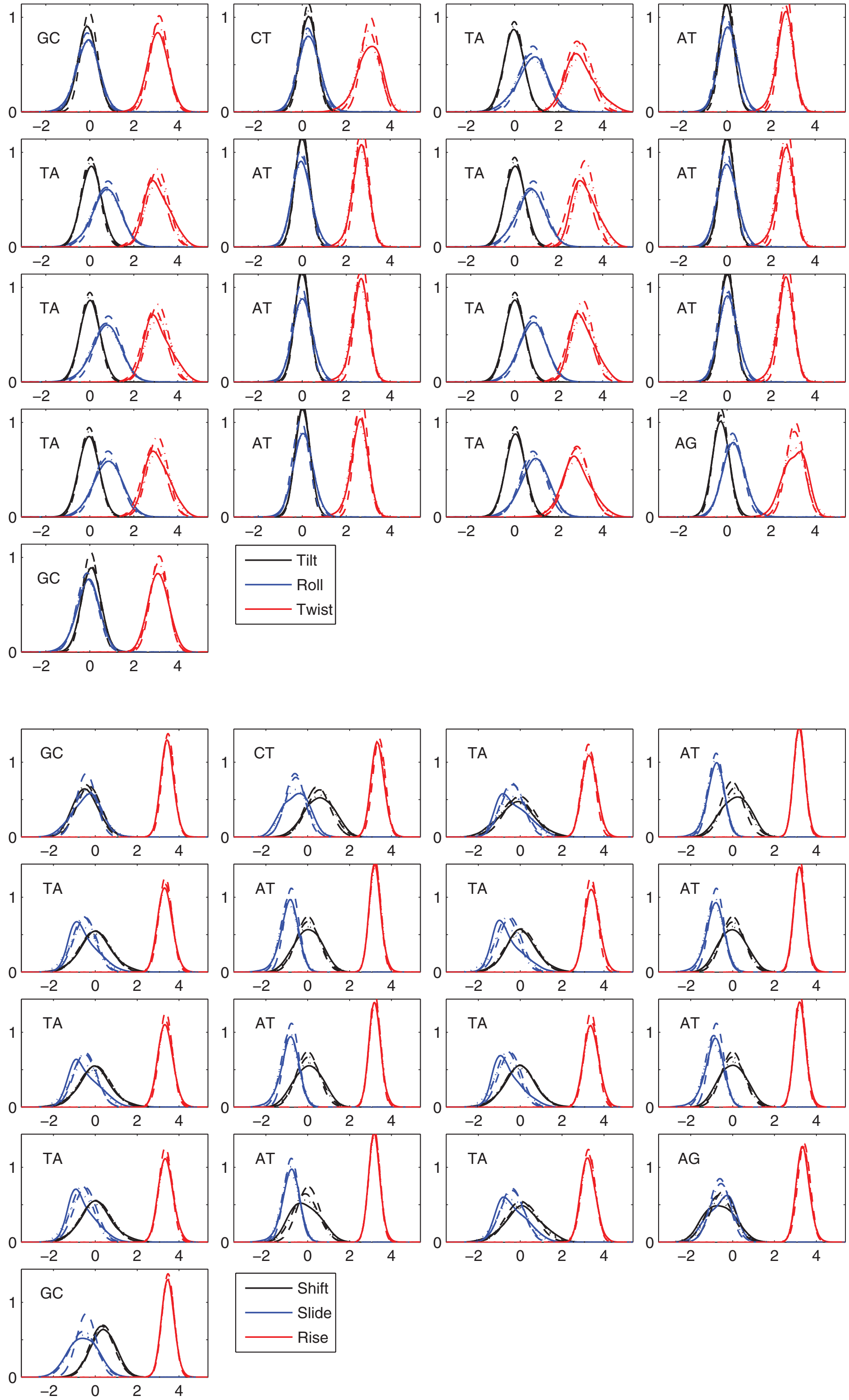


FIG. 7. Normalized marginal distributions for inter-basepair coordinates at each junction along oligomer S_1. Junctions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each junction shows the dimer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{1,\mathrm{o}}$, dotted; density $\rho^*_{1,\mathrm{M}}$, dashed-dotted; density $\rho^*_{1,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units.

we used it to predict the shape vector $\widehat{\mathsf{w}}^*_{1',\mathrm{m}}$ and stiffness matrix $\mathsf{K}^*_{1',\mathrm{m}}$ for an 18-basepair oligomer $\mathsf{S}_{1'}$ not in the original training set. Specifically, oligomer $\mathsf{S}_{1'}$ is a single point mutation of the training set oligomer S_1; these two oligomers differ by a single base at position 6: $\mathsf{S}_{1'}$ has a T in this position, whereas S_1 has an A. We constructed the quantities $\widehat{\mathsf{w}}^*_{1',\mathrm{m}}$ and $\mathsf{K}^*_{1',\mathrm{m}}$ directly from our existing best-fit parameter set $\mathcal{P}^*$ discussed in Sec. V C. To assess the accuracy of the prediction, we then carried out a full MD simulation of oligomer $\mathsf{S}_{1'}$ and obtained an observed shape vector $\widehat{\mathsf{w}}_{1',\mathrm{o}}$ and stiffness

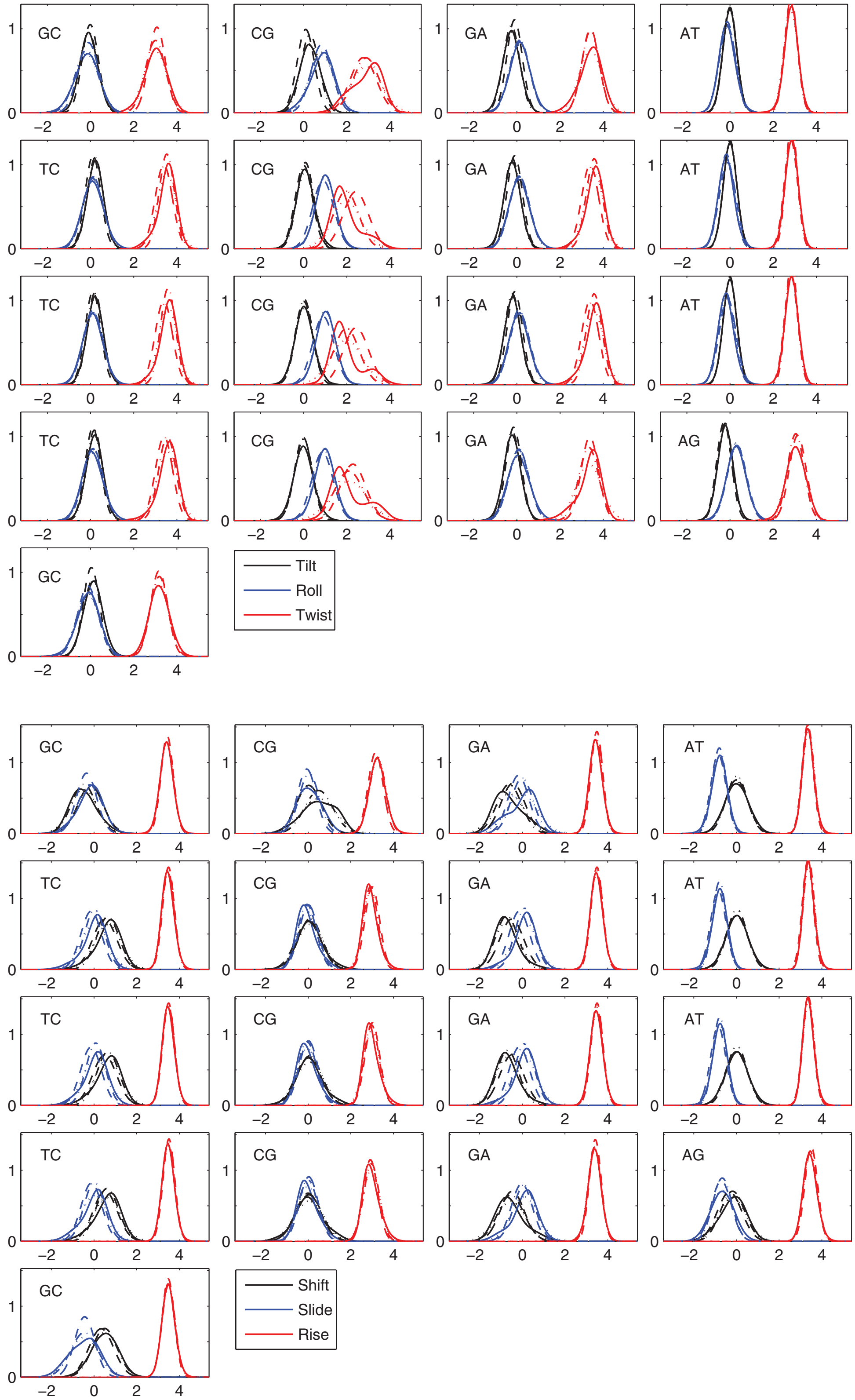


FIG. 8. Normalized marginal distributions for inter-basepair coordinates at each junction along oligomer S_8. Junctions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each junction shows the dimer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{8,\mathrm{o}}$, dotted; density $\rho^*_{8,\mathrm{M}}$, dashed-dotted; density $\rho^*_{8,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units.

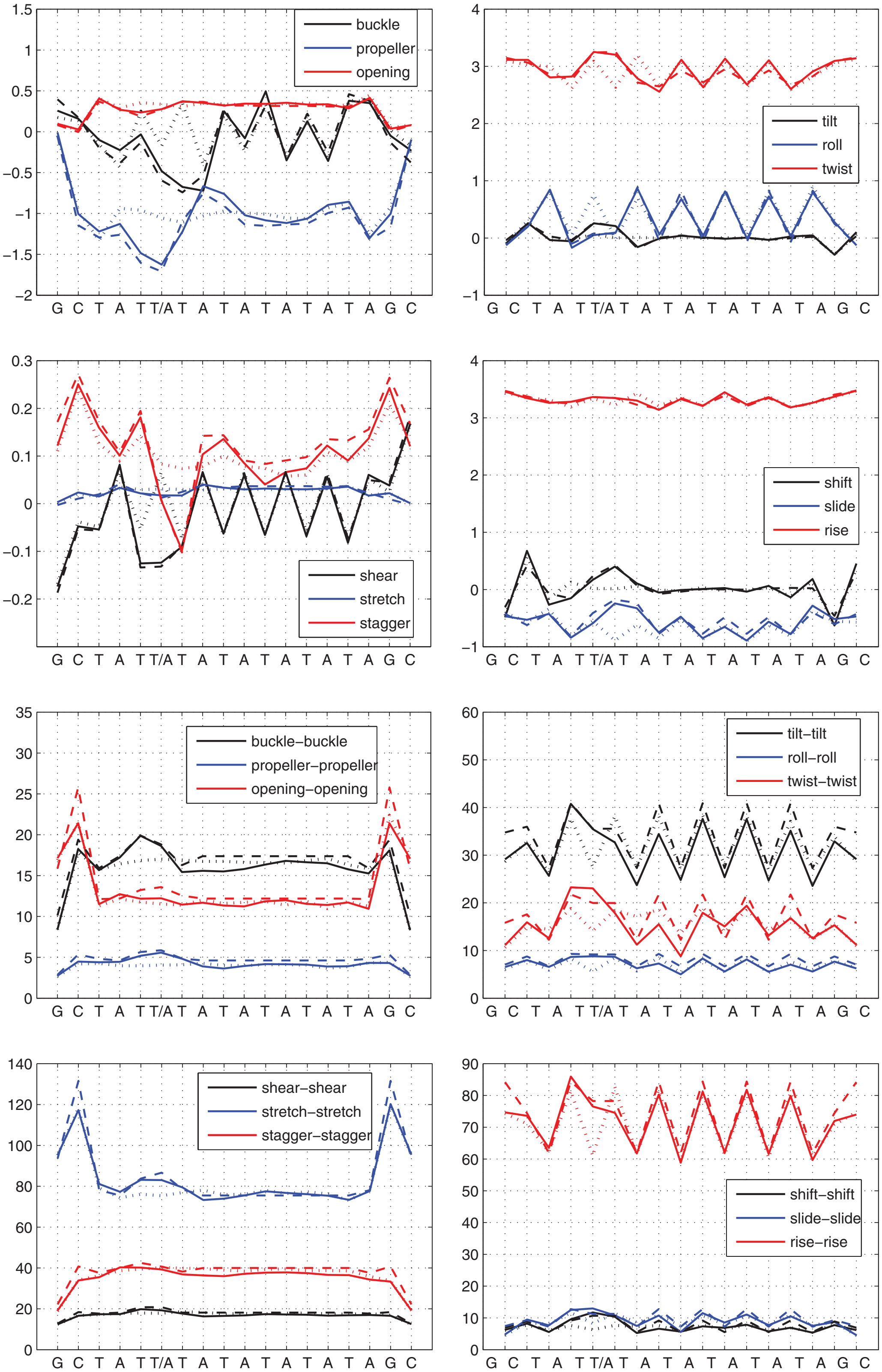


FIG. 9. Entries of shape vectors and stiffness matrices in dimensionless units for oligomer $S_{1'}$, which is a single point mutation of S_1. The base sequence of the reference strand is indicated on the horizontal axis, where the symbol T/A denotes the mutation site. (Top four panels) entries of observed vector $\widehat{w}_{1',o}$ (solid) and constructed dimer-based model vector $\widehat{w}^*_{1',m}$ (dashed) for $S_{1'}$, and observed vector $\widehat{w}_{1,o}$ (dotted) for S_1. (Bottom four panels) diagonal entries of nearest-neighbor oligomer-based model matrix $K^*_{1',M}$ (solid) and constructed dimer-based model matrix $K^*_{1',m}$ (dashed) for $S_{1'}$, and observed matrix $K_{1,o}$ (dotted) for S_1.

matrix $\mathsf{K}_{1',o}$ for comparison. We stress that the MD data for oligomer $\mathsf{S}_{1'}$ were not part of the training set from which $\mathcal{P}^*$ was derived.

Figure 9 shows the results of our prediction and comparison. The top four panels show the entries of the predicted shape parameter vector $\widehat{\mathsf{w}}^*_{1',m}$ in dashed lines and the observed vector $\widehat{\mathsf{w}}_{1',o}$ in solid lines as a function of position along oligomer $\mathsf{S}_{1'}$. For reference, the entries of the observed shape parameter vector $\widehat{\mathsf{w}}_{1,o}$ from the original, unmutated oligomer S_1 is taken from Figure 4 and is here plotted with dotted lines. In each panel, the base sequence of the reference strand of the oligomer is indicated on the abscissa just as before, with the mutation site indicated at position 6. The bottom four panels are analogous and show the diagonal entries of the nearest-neighbor oligomer-based model stiffness matrix $\mathsf{K}^*_{1',M}$ in solid lines and the constructed dimer-based model stiffness matrix $\mathsf{K}^*_{1',m}$ in dashed lines for oligomer $\mathsf{S}_{1'}$, and for reference the diagonal entries of the observed stiffness matrix $\mathsf{K}_{1,o}$ of the unmutated oligomer S_1 are shown in dotted lines.

The data in Figure 9 show that a change of one base in an oligomer can have pronounced effects on the oligomer shape and stiffness parameters. The changes in the oligomer shape parameters can be significantly nonlocal as illustrated by the predicted and observed data in the top four panels. Indeed, the predicted and observed effects of the mutation are in rather good agreement. The nonlocal effects are most pronounced for buckle, propeller and stagger, where the major effects are spread over approximately five bases, and are less pronounced, but still noticeable, for shear, twist, shift and slide, where the effects are spread over approximately three bases. For other quantities, for example, roll, the effect of the mutation appears rather local, and for others, for example, opening, stretch, and rise, any effect is not clearly visible on the scale of the plots. The changes in the oligomer stiffness parameters are illustrated in the bottom four panels. The dimer-based model, by design, predicts that the stiffness parameters can only change locally near the mutation site, and here too the prediction is in rather good agreement with the observed results.

Table III provides a further quantitative comparison between the predicted and observed effects of the single point mutation. The table compares the constructed dimer-based model and observed values of the intra-basepair parameters buckle and propeller for the central monomer X_8 of the pentamer sub-sequence $\mathsf{X}_6\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9\mathsf{X}_{10}$ of oligomers S_1 and $\mathsf{S}_{1'}$. Specifically, this pentamer has sequence ATATA in S_1 and TTATA in $\mathsf{S}_{1'}$. The mutation site is at position 6, and we compare the values of buckle and propeller for the central monomer A at position 8. The nonlocal effects of the mutation can be interpreted as context effects for the parameters of the central monomer. The table shows that the observed values of buckle and propeller for the central monomer depend significantly on the surrounding pentamer context, and that the dimer-based model can predict this dependence. The table also compares the constructed dimer-based model and observed values of the inter-basepair parameter slide for the junction in the central dimer $\mathsf{X}_7\mathsf{X}_8$ of the tetramer sub-sequence $\mathsf{X}_6\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9$ of oligomers S_1 and $\mathsf{S}_{1'}$. This tetramer has sequence ATAT in S_1 and TTAT in $\mathsf{S}_{1'}$, with the central dimer in both cases being TA. The table shows that the observed value of Slide for the central dimer step depends significantly on the surrounding tetramer context, and that our dimer-based model can again predict this nonlocal dependence.

TABLE III. Comparison between predicted and observed nonlocal sequence dependence of buckle, propeller and slide in dimensionless units from two different sequence contexts arising in oligomers S_1 and $\mathsf{S}_{1'}$. (Top) values of buckle and propeller in the central monomer $\mathsf{X}_8 = \mathsf{A}$ of the pentamer sub-sequence $\mathsf{X}_6\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9\mathsf{X}_{10}$. (Bottom) values of slide in the central junction $\mathsf{X}_7\mathsf{X}_8 = \mathsf{TA}$ of the tetramer sub-sequence $\mathsf{X}_6\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9$.

S_1,$\mathsf{S}_{1'}$ pentamer $\mathsf{X}_6\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9\mathsf{X}_{10}$	ATATA	TTATA
Buckle at X_8 (observed)	− 0.27	− 0.73
Buckle at X_8 (predicted)	− 0.25	− 0.53
Propeller at X_8 (observed)	− 1.03	− 0.72
Propeller at X_8 (predicted)	− 1.09	− 0.75
S_1,$\mathsf{S}_{1'}$ tetramer $\mathsf{X}_6\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9$	ATAT	TTAT
Slide at $\mathsf{X}_7\mathsf{X}_8$ (observed)	−0.56	−0.29
Slide at $\mathsf{X}_7\mathsf{X}_8$ (predicted)	−0.48	−0.24

VIII. SUMMARY AND CONCLUSIONS

This presentation has introduced a novel hierarchy of coarse-grain, sequence-dependent, rigid-base models of B-form DNA in solution, each of which provides an immediate approximation of the configuration-space equilibrium distribution for oligomers of arbitrary length and sequence. The hierarchy depends on both the assumed range of energetic couplings, and the assumed extent of sequence dependence of the model parameters. Attention was focussed on the particular model in the hierarchy that has nearest-neighbor interactions and dimer sequence dependence of the model parameters. For a Gaussian version of this model, a complete coarse-grain parameter set was estimated starting from a recent and extensive database of atomic-resolution MD simulations. The Kullback-Leibler divergence between probability density functions was used to make several quantitative assessments of the accuracy of the nearest-neighbor, dimer-dependent model with the specific parameter set. This particular model was compared both against others in the hierarchy to assess the validity of various assumptions pertaining to the locality of the energetic couplings and the level of sequence dependence of its parameters, and also against an all-atom MD simulation not in the training data set to assess its predictive capabilities. The results suggest that, compared to more sophisticated models with larger parameter sets, the nearest-neighbor, dimer-dependent model represents a practical compromise between simplicity and accuracy. Moreover, within various natural limitations pertaining to bi-modality and end effects, the model can quantitatively predict the equilibrium statistical properties of various oligomers rather well. Specifically, the model can successfully resolve sequence effects both within and between oligomers. In particular, the model can successfully predict the nonlocal structural consequences of a single point mutation.

Perhaps the most significant conceptual feature of all of our models is that they exhibit the phenomenon of frustration. To our knowledge, this phenomenon has not been considered or exploited in previous efforts on modeling the sequence-dependent curvature and flexibility of DNA; it provides a simple mechanism for describing and understanding the nonlocal dependence on sequence of the minimum energy shape of an oligomer, as has been observed in recent MD studies. This frustration, or pre-existing stress, adds a new dimension to the structural code of DNA and may potentially have functional implications in recognition, binding and other kinds of interactions that involve full or partial denaturation of the double-helix. Sequences with a high, localized concentration of frustration energy could well be denaturation hot-spots.

Our nearest-neighbor, dimer-dependent model with its initial parameter set, should, we hope, already provide a useful tool for understanding the sequence-dependent mechanics of DNA in various contexts. Due to the relatively high structural resolution of a rigid-base model, the ability to predict the ground-state or minimum energy conformation of an oligomer could prove useful in the refinement of future crystallographic and spectroscopic structures of DNA, especially in the analysis of larger structures for which only partial information may be available. Moreover, the ability to predict the elastic coupling matrix for an entire oligomer could prove useful in studying the interdependence of displacements and rotations of the bases of an oligomer in response to external loads. A detailed analysis of DNA wrapped around a nucleosome using our model is one obvious target application.[5–7] The sign and magnitude of the elastic coupling constants in our parameter set provide useful information about the polarity and strength of correlations and could reveal key mechanisms of the initial, sequence-dependent pathways in various protein binding and related interactions.

Another evident application of our model is to study cyclization and other looping rates of DNA oligomers as a function of their sequence.[8–10] In the context of our model, this classic experimental technique can be described as computing the marginal probability for the relative displacement and orientation of the first and last basepair frame in the oligomer. Because of the nonlinear geometry involved in the reconstruction of the global shape of an oligomer, this is not an explicit computation even within a Gaussian model, but it is a well-studied problem at least when the starting point is a rigid-basepair model. Although such looping problems can be addressed within the context of the discrete models themselves,[38] they can also be considered in the context of the analogous limiting continuum models, which can be computationally advantageous on the pertinent longer length scales of several tens to several hundreds of basepairs.[29,31,32,85] For discrete rigid-basepair models, the continuum limit is an inhomogeneous elastic rod or worm-like chain.[26,29,40] In contrast, the continuum limit of our nearest-neighbor, dimer-dependent, rigid-base model leads naturally to an inhomogeneous, continuum bi-rod model,[86] which incorporates structural features beyond the conventional elastic rod and worm-like chain models, and whose mathematical analysis may offer enhanced understanding of DNA supercoiling, looping, and related phenomena, including the prediction of hot-spots for denaturation.

In another direction for future work, it is evident that the nearest-neighbor model could itself be refined in various ways. Most simply it is certainly to be expected that the accuracy of the model parameterization will improve as more MD simulation data becomes available. The question of ergodicity of the MD simulations underlying our training data set remains open. The database used here had between 50 and 200 ns simulations of each oligomer, but longer simulations would certainly be desirable, and are already becoming available. A 10 μs simulation of a single ABC oligomer (carried out on specialized hardware[87]) suggests that a database of 3 μs simulations should be sufficient to sample the main fluctuations of B-family DNA at room temperature.[88] It is also certainly true that as the MD potentials underlying the fine-grain simulation data evolve, the quality of the associated coarse-grain parameters should improve correspondingly. One benefit of using MD simulations to provide the coarse-grain training data set is that if it is desired to study the effects of different solvent and ion conditions on the coarse-grain model, then it is just necessary to run the appropriate set of MD simulations, thereby modify the training data set, and reapply the parameter extraction procedure described here. Similarly, if a coarse-grain model of methylated bases is desired, this can be simply done provided that an appropriate set of MD simulations is available, as has already been carried out for a rigid-basepair model.[89] Now of course there will be a larger parameter set to allow for methylated and unmethylated bases.

Staying within the context of Gaussian models, the methodology introduced here could be used to estimate parameters for more sophisticated models in our hierarchy with beyond nearest-neighbor interactions, in much the same way as is done here explicitly for the nearest-neighbor, dimer-dependent case. The errors plotted in Figure 2 suggest that a next-nearest-neighbor type of model may provide a significant improvement in accuracy and hence be worthy of future study. The estimation of parameters for such a model would naturally require MD training data with a sufficiently rich set of sequences, including a sufficiently rich set of end sequences.

Finally, and more challengingly, it is evident that it would be of interest to introduce functional forms in the free energy that are more general than shifted quadratics, and to fit parameters in the resulting non-Gaussian distributions in order to attempt to capture effects such as the bi-modality,[21,90] which is evident in marginal distributions for some junction parameters at some dimer steps, end-fraying, and the sequence-dependent probability of DNA melting. Such a generalization would appear to be challenging, but it would be of considerable interest to combine the strengths of our Gaussian rigid-base models with their sequence-dependent frustration and detailed configuration variables, with existing models[51,54] of melting, which are non-Gaussian but with a comparatively simplified approximation to configuration space with fewer degrees of freedom.

ACKNOWLEDGMENTS

The authors thank all colleagues in the ABC consortium for sharing their trajectories, and Jérémy Curuksu for collaborating in the initial analysis of the ABC data set. D.P. and J.H.M. were supported in part by the Swiss National Science Foundation under Award No. 200021-126666.

[1]N. Becker, L. Wolff, and R. Everaers, Nucleic Acids Res. **34**, 5638 (2006).
[2]G. Holman, M. Zewail-Foote, A. Smith, K. Johnson, and B. Iverson, Nat. Chem. **3**, 875 (2011).
[3]P. Poulain, A. Saladin, B. Hartmann, and C. Prevost, J. Comput. Chem. **29**, 2582 (2008).
[4]R. Rohs, X. Jin, S. West, R. Joshi, B. Honig, and R. Mann, Annu. Rev. Biochem. **79**, 233 (2010).
[5]N. Kaplan, I. Moore, Y. Fondufe-Mittendorf, A. Gossett, D. Tillo, Y. Field, E. LeProust, T. Hughes, J. Lieb, J. Widom, and E. Segal, Nature (London) **458**, 362 (2009).
[6]A. Morozov, K. Fortney, D. Gaykalova, V. Studitsky, J. Widom, and E. Siggia, Nucleic Acids Res. **37**, 4707 (2009).
[7]T. Schlick, J. Hayes, and S. Grigoryev, J. Biol. Chem. **287**, 5183 (2012).
[8]T. Cloutier and J. Widom, Proc. Natl. Acad. Sci. U.S.A. **102**, 3645 (2005).
[9]R. Schleif, Annu. Rev. Biochem. **61**, 199 (1992).
[10]J. Vilar and S. Leibler, J. Mol. Biol. **331**, 981 (2003).
[11]F. Michor, J. Liphardt, M. Ferrari, and J. Widom, Nat. Rev. Cancer **11**, 657 (2011).
[12]P. Severin, X. Zou, H. Gaub, and K. Schulten, Nucleic Acids Res. **39**, 8740 (2011).
[13]A. Carbone and N. Seeman, in *Aspects of Molecular Computing*, Lecture Notes in Computer Science Vol. 2950, edited by N. Jonoska, G. Paun, and G. Rozenberg (Springer, 2004), pp. 121–137.
[14]R. Goodman, I. Schaap, C. Tardin, C. Erben, R. Berry, C. Schmidt, and A. Turberfield, Science **310**, 1661 (2005).
[15]D. Han, S. Pal, J. Nangreave, Z. Deng, Y. Liu, and H. Yan, Science **332**, 342 (2011).
[16]R. Barish, R. Schulman, P. Rothemund, and E. Winfree, Proc. Natl. Acad. Sci. U.S.A. **106**, 6054 (2009).
[17]E. Franco, E. Friedrichs, J. Kim, R. Jungmann, R. Murray, E. Winfree, and F. Simmel, Proc. Natl. Acad. Sci. U.S.A. **108**, E784 (2011).
[18]R. Schulman, B. Yurke, and E. Winfree, Proc. Natl. Acad. Sci. U.S.A. **109**, 6405 (2012).
[19]D. Beveridge, G. Barreiro, K. Byun, D. Case, T. Cheatham III, S. Dixit, E. Giudice, F. Lankas, R. Lavery, J. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. Thayer, P. Varnai, and M. Young, Biophys. J. **87**, 3799 (2004).
[20]S. Dixit, D. Beveridge, D. Case, T. Cheatham III, E. Giudice, F. Lankas, R. Lavery, J. Maddocks, R. Osman, H. Sklenar, K. Thayer, and P. Varnai, Biophys. J. **89**, 3721 (2005).
[21]R. Lavery, K. Zakrzewska, D. Beveridge, T. Bishop, D. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer, Nucleic Acids Res. **38**, 299 (2010).
[22]M. Orozco, A. Noy, and A. Perez, Curr. Opin. Struct. Biol. **18**, 185 (2008).
[23]A. Perez, I. Marchan, D. Svozil, J. Sponer, T. Cheatham III, C. Laughton, and M. Orozco, Biophys. J. **92**, 3817 (2007).
[24]T. Drsata, A. Perez, M. Orozco, A. Morozov, J. Sponer, and F. Lankas, J. Chem. Theory Comput. **9**(1), 707 (2013).
[25]A. Perez, F. Luque, and M. Orozco, J. Am. Chem. Soc. **129**, 14739 (2007).
[26]N. Becker and R. Everaers, Phys. Rev. E **76**, 021923 (2007).
[27]C. Benham and S. Mielke, Annu. Rev. Biomed. Eng. **7**, 21 (2005).
[28]G. Chirikjian, J. Phys.: Condens. Matter **22**, 323103 (2010).
[29]J. Maddocks, in *A Celebration of Mathematical Modeling: The Joseph B. Keller Anniversary Volume*, edited by D. Givoli, M. Grote, and G. Papanicolaou (Kluwer Science, 2004), pp. 113–136.
[30]T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide* (Springer, New York, 2002).
[31]L. Cotta-Ramusino and J. Maddocks, Phys. Rev. E **82**, 051924 (2010).
[32]L. Czapla, D. Swigon, and W. Olson, J. Chem. Theory Comput. **2**, 685 (2006).
[33]S. Harvey, A. Petrov, B. Devkota, and M. Boz, Phys. Chem. Chem. Phys. **11**, 10553 (2009).
[34]S. Mielke, N. Gronbech-Jensen, and C. Benham, Phys. Rev. E **77**, 031924 (2008).
[35]J. Peters and L. Maher III, Q. Rev. Biophys. **43**, 23 (2010).
[36]T. Schlick and O. Perisic, Phys. Chem. Chem. Phys. **11**, 10729 (2009).
[37]A. Vologodskii and V. Rybenkov, Phys. Chem. Chem. Phys. **11**, 10543 (2009).
[38]Y. Zhang and D. Crothers, Biophys. J. **84**, 136 (2003).
[39]Y. Zhou and G. Chirikjian, Macromolecules **39**, 1950 (2006).
[40]B. Coleman, W. Olson, and D. Swigon, J. Chem. Phys. **118**, 7127 (2003).
[41]O. Gonzalez and J. Maddocks, Theor. Chem. Acc. **106**, 76 (2001).
[42]F. Lankas, in *Innovations in Biomolecular Modeling and Simulations*, edited by T. Schlick (Royal Society of Chemistry, 2012), Vol. 2, pp. 3–32.
[43]F. Lankas, O. Gonzalez, L. Heffler, G. Stoll, M. Moakher, and J. Maddocks, Phys. Chem. Chem. Phys. **11**, 10565 (2009).
[44]J. Walter, O. Gonzalez, and J. Maddocks, SIAM Multi. Model. Simul. **8**, 1018 (2010).
[45]R. DeMille, T. Cheatham III, and V. Molinero, J. Phys. Chem. B **115**, 132 (2011).
[46]K. Doi, T. Haga, H. Shintaku, and S. Kawano, Philos. Trans. R. Soc. A **368**, 2615 (2010).
[47]G. Freeman, D. Hinckley, and J. de Pablo, J. Chem. Phys. **135**, 165104 (2011).
[48]M. Machado, P. Dans, and S. Pantano, Phys. Chem. Chem. Phys. **13**, 18134 (2011).
[49]A. Morriss-Andrews, J. Rottler, and S. Plotkin, J. Chem. Phys. **132**, 035105 (2010).
[50]S. Niewieczerzal and M. Cieplak, J. Phys.: Condens. Matter **21**, 474221 (2009).
[51]T. Ouldridge, A. Louis, and J. Doye, J. Chem. Phys. **134**, 085101 (2011).
[52]A. Savelyev and G. Papoian, Proc. Natl. Acad. Sci. U.S.A. **107**, 20340 (2010).
[53]H. Tepper and G. Voth, J. Chem. Phys. **122**, 124906 (2005).
[54]A. Wildes, N. Theodorakopoulos, J. Valle-Orero, S. Cuesta-Lopez, J.-L. Garden, and M. Peyrard, Phys. Rev. Lett. **106**, 048101 (2011).
[55]J. Zou, W. Liang, and S. Zhang, Int. J. Numer. Methods Eng. **83**, 968 (2010).
[56]R. Dickerson, M. Bansal, C. Calladine, S. Diekmann, W. Hunter, O. Kennard, R. Lavery, H. Nelson, W. Olson, W. Saenger, Z. Shakked, H. Sklenar, D. Soumpasis, C.-S. Tung, E. von Kitzing, A. Wang, and V. Zhurkin, J. Mol. Biol. **205**, 787 (1989).
[57]M. El Hassan and C. Calladine, J. Mol. Biol. **251**, 648 (1995).
[58]X.-J. Lu and W. Olson, Nucleic Acids Res. **31**, 5108 (2003).
[59]W. Olson, A. Gorin, X.-J. Lu, L. Hock, and V. Zhurkin, Proc. Natl. Acad. Sci. U.S.A. **95**, 11163 (1998).
[60]A. Bolshoy, P. McNamara, R. Harrington, and E. Trifonov, Proc. Natl. Acad. Sci. U.S.A. **88**, 2312 (1991).
[61]I. Brukner, R. Sanchez, D. Suck, and S. Pongor, J. Biomol. Struct. Dyn. **13**, 309 (1995).
[62]Y. Liu and D. Beveridge, J. Biomol. Struct. Dyn. **18**, 505 (2001).
[63]P. Santis, A. Palleschi, M. Savino, and A. Scipioni, Biochemistry **29**, 9269 (1990).
[64]M. Packer, M. Dauncey, and C. Hunter, J. Mol. Biol. **295**, 71 (2000).
[65]M. Packer, M. Dauncey, and C. Hunter, J. Mol. Biol. **295**, 85 (2000).
[66]See supplementary material at http://dx.doi.org/10.1063/1.4789411 for supplemental appendices, complete parameter set, and accompanying program files.
[67]W. Olson, M. Bansal, S. Burley, R. Dickerson, M. Gerstein, S. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger, and H. Berman, J. Mol. Biol. **313**, 229 (2001).
[68]R. Lavery, M. Moakher, J. Maddocks, D. Petkeviciute, and K. Zakrzewska, Nucleic Acids Res. **37**, 5917 (2009).
[69]R. Hogg and A. Craig, *Introduction to Mathematical Statistics*, 3rd ed. (Macmillan, New York, 1970).
[70]S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).
[71]E. Jaynes, Phys. Rev. **106**, 620 (1957).
[72]E. Jaynes, Phys. Rev. **108**, 171 (1957).
[73]E. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[74]A. Majda and X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows* (Cambridge University Press, Cambridge, 2006).
[75]H. Berendsen, J. Grigera, and T. Straatsma, J. Phys. Chem. **91**, 6269–6271 (1987).
[76]L. Dang, J. Am. Chem. Soc. **117**, 6954 (1995).
[77]H. Berendsen, J. Postma, W. van Gunsteren, A. DiNola, and J. Haak, J. Chem. Phys. **81**, 3684–3690 (1984).

[78] A. Srinivasan, R. Sauers, M. Fenley, A. Boschitsch, A. Matsumoto, A. Colasanti, and W. Olson, Biophys. Rev. **1**, 13 (2009).
[79] R. Blake and S. Delcourt, Biopolymers **29**, 393 (1990).
[80] D. Crothers and H. Spatz, Biopolymers **10**, 1949 (1971).
[81] R. Wartell, J. Klysik, W. Hillen, and R. Wells, Proc. Natl. Acad. Sci. U.S.A. **79**, 2549 (1982).
[82] J. Burke, A. Lewis, and M. Overton, SIAM J. Optim. **15**, 751 (2005).
[83] A. Lewis and M. Overton, Math. Program. (2012).
[84] J. Wang, Proc. Natl. Acad. Sci. U.S.A. **76**, 200 (1979).
[85] P. Furrer, R. Manning, and J. Maddocks, Biophys. J. **79**, 116 (2000).
[86] M. Moakher and J. Maddocks, Arch. Ration. Mech. Anal. **177**, 53 (2005).
[87] D. Shaw, M. Deneroff, R. Dror, J. Kuskin, R. Larson, J. Salmon, C. Young, B. Batson, K. Bowers, J. Chao, M. Eastwood, J. Gagliardo, J. Grossman, C. Ho, D. Ierardi, I. Kolossváry, J. Klepeis, T. Layman, C. McLeavey, M. Moraes, R. Mueller, E. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. Wang, Commun. ACM **51**, 91 (2008).
[88] R. Lavery and T. Cheatham III, private communication (December 2012).
[89] A. Perez, C. Castellazzi, F. Battistini, K. Collinet, O. Flores, O. Deniz, M. Ruiz, D. Torrents, R. Eritja, M. Soler-Lopez, and M. Orozco, Biophys. J. **102**, 2140 (2012).
[90] P. Dans, A. Perez, I. Faustino, R. Lavery, and M. Orozco, Nucleic Acids Res. **40**, 10668 (2012).