Multiscale Model. Simul.
Vol. 15, No. 3, pp. 1073–1107

ABSOLUTE VERSUS RELATIVE ENTROPY PARAMETER ESTIMATION IN A COARSE-GRAIN MODEL OF DNA*

O. GONZALEZ†, M. PASI‡, D. PETKEVIČIŪTĖ§, J. GLOWACKI¶, AND J. H. MADDOCKS¶

Abstract. Maximum entropy procedures for estimating coarse-grain parameters from molecular dynamics (MD) simulation data are considered within the specific context of the sequence-dependent *cgDNA* rigid-base model of DNA. We describe a quite general approach that exploits a maximum *absolute* entropy principle to fit an observed matrix of covariances subject to the constraint of only allowing a prescribed sparsity pattern of nearest-neighbor interactions in the free energy. We also allow indefinite local stiffness-matrix parameter blocks that nevertheless always generate a positive-definite model stiffness matrix. Beginning from a database of atomic-resolution MD simulations of a library of short DNA oligomers in explicit solvent, these procedures deliver a complete parameter set for the *cgDNA* model. Due to the intrinsic linear structure of DNA and the convergence characteristics of the MD time series data, the maximum absolute entropy parameter set yields significantly improved predictions of persistence lengths, when compared to a previous parameter set that was fit to the same MD data, but using a maximum *relative* entropy fitting principle and local stiffness-matrix parameter blocks that were constrained to be semidefinite.

Key words. entropy principles, parameter estimation, molecular modeling, statistical mechanics

AMS subject classifications. 62H12, 82B05, 82D60, 15A83, 15A09

DOI. 10.1137/16M1086091

1. Introduction.

An important problem in molecular biology is to understand how the mechanical properties of DNA depend on the sequence of bases along its two backbones. Properties that influence bending, twisting, shearing, and stretching in different directions along and across the two strands are believed to be essential in various biological processes such as DNA looping [34], nucleosome positioning [35], and other DNA-protein interactions, and gene regulation [26], all of which depend on the probability of DNA to adopt various three-dimensional configurations [4]. Consequently models at differing length scales are needed to quantify how the mechanical properties of DNA depend upon its sequence. The intermediate scales of a few tens to a few hundreds of base pairs are of particular biological interest. The study of sequence-dependent effects at such scales requires the development of specialized coarse-grain models, because, with contemporary computational resources, all-atom molecular dynamics (MD) simulations at these lengths remain intensive, particularly given the large number of possible sequences, while sequence-dependent behavior is

*Received by the editors July 22, 2016; accepted for publication (in revised form) April 19, 2017; published electronically July 18, 2017.

http://www.siam.org/journals/mms/15-3/M108609.html

Funding: The second, third, fourth, and fifth authors were supported in part by the Swiss National Science Foundation under award 200020 143613/1 to the fifth author.

†Department of Mathematics, University of Texas at Austin, Austin, TX 78712 (og@math.utexas.edu).

‡Section de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland. Current address: Centre for Biomolecular Sciences, University of Nottingham, University Park, UK (marco.pasi@nottingham.ac.uk).

§Section de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland. Current address: Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Lithuania (daiva.petkeviciute@ktu.lt).

¶Section de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland (jaroslaw.glowacki@epfl.ch, john.maddocks@epfl.ch).

below the resolution of the standard, homogeneous, wormlike chain or uniform, elastic rod coarse-grain models.

The *cgDNA* model, recently introduced in [13, 31], was developed to predict a sequence-dependent free energy and associated equilibrium probability distribution for an oligomer[1] of B-form DNA of arbitrary sequence in solvent under prescribed environmental conditions at these intermediate length scales. The model is of the rigid-base type in which each base on each strand of the DNA is considered as a rigid entity interacting with its nearest neighbors. The coarse-grain configuration of a given DNA oligomer is described by an independent set of standard internal helicoidal coordinates corresponding to the relative, three-dimensional rotation and displacement between neighboring bases both along and across the two backbone strands. The model is then completed by a parameter set, which depends only on the local dinucleotide[2] sequence along the oligomer, combined with a rule for constructing a shifted, quadratic approximation to the free energy for any oligomer from its sequence. The stiffness, or coefficient, matrix in this quadratic form is banded, reflecting the assumption of there being only nearest-neighbor interactions, and has a local dependence on sequence. In contrast the shift, or ground-state configuration, of the quadratic energy has a nonlocal sequence dependence due to the fact that it arises from a completion-of-squares operation involving the inverse of the banded stiffness matrix. This nonlocality of sequence-dependence is a unique feature of the *cgDNA* model within coarse-grain descriptions of DNA. It encapsulates the frustration or pre-existing stress in an oligomer. As a consequence the *cgDNA* model has been shown to successfully resolve observed sequence effects, both within and between DNA oligomers, down to the resolution of nonlocal changes in the ground-state configuration due to a single point mutation in the sequence.

Once a *cgDNA* parameter set has been estimated, the construction of the *cgDNA* free energy function for an oligomer of arbitrary sequence is an essentially trivial computation, and the configurational statistics of that oligomer are then described by an associated Gaussian equilibrium probability density on the space of internal coordinates, for which efficient sampling techniques are available. Associated software and examples are described in [27, 32]. In particular, the computational effort associated with applying the *cgDNA* model to a new sequence is far less than that required in an additional MD simulation for each sequence of interest, so that studies of much larger ranges in both length scales and sequence variation become feasible.

Nevertheless the scientific utility of the *cgDNA* model depends upon the accuracy of its parameter set. In this presentation we consider mathematical procedures for estimating these sequence-dependent parameter sets. While we describe our results within the specific context of the *cgDNA* model, the maximum entropy approach to enforcing a prescribed sparsity pattern in the stiffness (or precision) matrix in a Gaussian is potentially also of wider interest [3, 6, 10, 36, 38]. For example, in the specific application fields of numerical weather prediction and data assimilation, both sparse covariance and sparse inverse covariance (or precision) matrix estimates are adopted using other techniques such as *tapering* [2, 5, 11, 33, 37].

As our starting point, we assume that a library of atomic-resolution MD simula-

[1] The standard molecular biology term *oligomer* is used here and throughout as a synonym for a short fragment or molecule of DNA.

[2] A *nucleotide* is the basic structural unit of DNA, comprising a base, of one of four types T, A, C, or G, together with its linked phosphate and sugar groups. The dinucleotide sequence context then means the sequence step along one strand, such as TA, GG, etc.

tions of an ensemble of training oligomers in explicit solvent[3] is available, and then we seek to estimate coarse-grain parameters from this database. Although other types of data, for example NMR data, could potentially be used, we focus on establishing methods tailored to the particular case of data coming from a finer-grain MD simulation due to both its easy availability and the continually improving state-of-the-art in the field; see, e.g.. [1, 8, 24, 29]. In particular, by taking MD time series simulation data for a library of sequences as our starting point, we have access to changes in solvent condition, such as temperature and salt concentration and species, just by running the appropriate MD simulation of the library; albeit we must rely on the accuracy of the underlying MD simulation protocols and potentials. A (contemporary) complete MD data set of time series for a single sequence library is of the order of 0.5 TB in size, from which an associated *cgDNA* parameter set is to be extracted, which comprises a total of 1592 independent scalars (that, as described below, can be naturally grouped into certain small independent vectors and symmetric matrices). Consequently the current mathematical considerations could reasonably be described as the development of efficient machine learning techniques for extracting model parameters from our particular structured MD simulation big data sets.

Our parameter estimation procedure comprises three steps. The first is the estimation of a coarse-grain configurational mean vector and covariance matrix from a fine-grain MD time series for each oligomer in a training set, assuming that these time series are stationary; for more details, see [13, 21]. The second is the fitting of the observed mean and covariance of each training set oligomer by a descriptive Gaussian model that is required to have the banded stiffness matrix that expresses the nearest-neighbor interaction assumption. And the third and final step is the estimation of the locally sequence-dependent parameters (which allow the prediction of *cgDNA* model free energies for oligomers of arbitrary sequence) by fitting to the banded Gaussian description of each oligomer in the training library. In this paper we focus on the second and third steps of the procedure and examine the impact of various mathematical choices in the associated fits.

Regarding the second step of our procedure, we compare two fitting strategies based on maximizing either an absolute or a relative entropy for the probability density function of the oligomer. Due to the linear structure of DNA, and the related convergence characteristics of MD time series data, we here argue that the fit based on maximum absolute entropy is more natural than the maximum relative entropy fit that we adopted previously. Specifically, the approach based on absolute entropy employs data from only a band about the diagonal of the estimated covariance matrix, whereas the approach based on relative entropy employs data from the entire estimated covariance matrix, and we present numerical evidence to suggest that the data that is close to the diagonal has a smaller error with respect to its assumed equilibrium or stationary value than the data that is far away. Moreover, the maximum absolute entropy fit can be constructed using a simple, local inversion algorithm [12], whereas the relative entropy fit requires numerical optimization techniques. We note that in the absolute entropy case, the maximization problem reduces to a matrix completion problem that has been previously studied [3, 6, 10, 14, 19, 22, 36, 38]. Furthermore, although we only consider means and covariances in this work, higher-order moments are also of interest, and it is understood that these could be accommodated in the maximum absolute entropy approach in a natural way [16, 17, 18].

[3] The term *explicit solvent* refers to an explicit, fully atomistic representation of the water molecules and ions making up the solvent within an MD simulation.

In the third step of the procedure, we seek to estimate the parameters in the *cgDNA* model by fitting them against the banded Gaussian description of each training oligomer. To this end, we characterize a best-fit parameter set as one that maximizes a natural objective function on a space of parameter sets, and we examine various assumptions on the choice of parameter space. As we will see, an important requirement of the *cgDNA* model is that it produce a positive-definite stiffness matrix, and hence a well-defined probability density function, for an oligomer of arbitrary sequence above some minimal length. This requirement can be guaranteed under various different restrictions on the parameter set, specifically on the parameter stiffness matrices associated with the local base-pair and junction energies, which are described in detail later. In previous work [13], these parameter matrices were assumed to be positive-semidefinite, which complicates both the characterization and numerical treatment of the parameter fitting problem. Although apparently natural, this positive-semidefinite restriction is not required by any physical consideration associated with the model. And we show here that positivity of the *cgDNA* model stiffness matrix for any sequence can be guaranteed with a class of parameter matrices that are themselves indefinite. This generalization simplifies the parameter space and allows for a faster and more efficient numerical treatment of the fitting problem. Moreover, the admissible set of parameter matrices includes, but is strictly bigger than, positive-semidefinite, so that the fit is by definition improved.

As an illustration of our results, we compare two different best-fit parameter sets for the *cgDNA* model. One set, referred to as *cgDNAparamset*1, was described in the prior work [13], while the other, *cgDNAparamset*2, is new. Whereas the set *cgDNAparamset*1 was based on a maximum relative entropy description of each training oligomer, and was computed using a simple gradient flow method with positive-semidefinite restrictions on the parameter stiffness matrices, the set *cgDNAparamset*2 is based on a maximum absolute entropy description, and was computed using a Newton–Broyden method without any restrictions on the definiteness of the parameter stiffness matrices. The set *cgDNAparamset*2 is also shown to be a locally unique optimizer for our fitting procedure in the sense that it satisfies the requisite first-order necessary, and second-order sufficient, conditions on the gradient and Hessian of the objective function. *cgDNAparamset*2 contains indefinite parameter stiffness matrices but satisfies a set of sufficient conditions for any *cgDNA* stiffness matrix to be positive-definite.

Finally, we show that the predictive capabilities of *cgDNA* with *cgDNAparamset*2 are noticeably improved as compared to *cgDNAparamset*1. Specifically, while the two parameter sets each predict the sequence-dependent variations in ground-state shape within and between oligomers rather well, we find that *cgDNAparamset*2 is significantly better than *cgDNAparamset*1 in predicting the stiffness properties of oligomers in the sense of persistence length. We attribute this improvement to the use of more natural choices in our parameter estimation procedure.

2. Maximum entropy estimation of coarse-grain equilibrium distributions.

Before turning to a description of the specific application to modeling DNA, we first describe the general class of problems that we treat and the general mathematical approaches that we adopt.

2.1. Coarse-grain equilibrium distributions.

Our assumed starting point for modeling is that we are considering a system that can be described with coarse-grain coordinates $\mathsf{w} \in \mathbb{R}^m$, and which is *microscopic* in the informal sense that the problem is one of statistical mechanics, where the physical observables are expectations of an underlying equilibrium or stationary probability density function $\rho(\mathsf{w})$ defined with respect to the standard measure $d\mathsf{w}$. Moreover, we assume that this density can be expressed in the exponential, or Boltzmann, form

$$\rho(\mathsf{w}) = \frac{1}{Z} e^{-U(\mathsf{w})}, \tag{1}$$

where $U(\mathsf{w})$ is the free energy of the problem (expressed in units of $k_B T$), and Z is the normalizing constant (or partition function).

The expected, or average, value of any function ϕ over the equilibrium distribution ρ is defined in the usual way as

$$\langle \phi \rangle_\rho := \int \phi(\mathsf{w}) \rho(\mathsf{w}) \, d\mathsf{w}, \tag{2}$$

where all integrals here and throughout are over the space $\mathbb{R}^m$ unless mentioned otherwise. The mean $\mu_\rho \in \mathbb{R}^m$ and (centred) covariance $\mathsf{C}_\rho \in \mathbb{R}^{m \times m}$ will play key roles in our development, and are defined as usual to be

$$\mu_\rho := \langle \mathsf{w} \rangle_\rho, \qquad \mathsf{C}_\rho := \langle \Delta\mathsf{w} \otimes \Delta\mathsf{w} \rangle_\rho, \tag{3}$$

where $\Delta\mathsf{w} = \mathsf{w} - \mu_\rho$ and $\otimes$ denotes the outer or tensor product of a vector, so that in components we have $[\mathsf{x} \otimes \mathsf{x}]_{pq} = \mathsf{x}_p \mathsf{x}_q$ for any vector x. Whereas the vector μ_ρ can be arbitrary, the matrix C_ρ is always symmetric positive-semidefinite, and we will consider cases where it can be assumed to also be positive-definite.

Our first goal is to predict optimal approximations to the model equilibrium probability density function $\rho(\mathsf{w})$, or equivalently the energy $U(\mathsf{w})$, from a finite set of observed data. Our assumed starting point for model parameter estimation is that we have an estimated mean μ and covariance $\mathsf{C} = \mathsf{C}^T > 0$ for the problem. These input data could in principle be obtained either from experiment, or, as in the case of our DNA modeling application, from a time series generated by a finer-grain simulation such as atomistic MD. We further assume that there is a banded index set $\mathcal{N}$, corresponding to the entries of a symmetric $m \times m$ matrix within specified overlapping diagonal blocks (and containing all diagonal entries), for which all large entries of C^{-1} have indices in $\mathcal{N}$. Of course the precise characterization of which entries of C^{-1} are large and which are small is ultimately a modeling decision. The index set complementary to $\mathcal{N}$ will be denoted by $\mathcal{N}^c$.

Because we assume that we have estimated values of only the mean and covariance of the distribution ρ, we will be lead naturally to Gaussian models of the form

$$\rho(\mathsf{w}) = \frac{1}{Z} e^{-\frac{1}{2}(\mathsf{w}-\mu)\cdot\mathsf{K}(\mathsf{w}-\mu)}, \tag{4}$$

in which the free energy $U(\mathsf{w})$ is a shifted quadratic form, and by well-known results for Gaussian integrals, e.g., [15], the normalizing constant Z is known explicitly, the shift μ is the mean, and the inverse of the stiffness matrix K is the covariance, i.e., $\mathsf{K}^{-1} = \mathsf{C}$. If it is instead assumed that estimates of higher-order moments, or other non-quadratic expectations, are available, we would instead be lead naturally to nonquadratic models of the free energy [16, 17, 18].

2.2. Absolute and relative entropy. We will use the standard notions of absolute and relative (Shannon) entropies in formulating our estimation strategies and adopt the sign conventions employed in [25]. By the absolute entropy of a density $\rho(\mathsf{w})$ we mean

$$D_{\mathrm{abs}}(\rho) = -\int \rho(\mathsf{w}) \ln\left[\rho(\mathsf{w})\right] d\mathsf{w}. \tag{5}$$

For the case of a bounded domain of unit measure, the absolute entropy satisfies $D_{\mathrm{abs}}(\rho) \le 0$ and $D_{\mathrm{abs}}(\rho) = 0$ if and only if ρ is the uniform density. (These observations follow from the notion of relative entropy introduced below, with one density being unity.) On unbounded domains, the existence and characterization of maximizing densities is more delicate. Intuitively, $D_{\mathrm{abs}}(\rho)$ can be interpreted as a measure of uniformity; densities that maximize $D_{\mathrm{abs}}(\rho)$ within a prescribed class can be interpreted as the most uniform (least biased) in that class. In the special case when ρ is a Gaussian, with a stiffness (or inverse covariance) matrix K, the integral in (5) can be explicitly evaluated to obtain

$$D_{\mathrm{abs}}(\rho) = \frac{1}{2}\Big[\ln\left(\det(2\pi I)/\det\mathsf{K}\right) + I : I\Big], \tag{6}$$

where a colon denotes the standard Frobenius inner product for square matrices, and I denotes the identity matrix of the same dimension m as K so that $I : I = m$.

By the relative entropy of a density $\rho_2(\mathsf{w})$ with respect to a density $\rho_1(\mathsf{w})$ we mean

$$D_{\mathrm{rel}}(\rho_2, \rho_1) = -\int \rho_2(\mathsf{w}) \ln\left[\frac{\rho_2(\mathsf{w})}{\rho_1(\mathsf{w})}\right] d\mathsf{w}. \tag{7}$$

This quantity is a nonsymmetric measure of the difference between ρ_2 and ρ_1; it satisfies $D_{\mathrm{rel}}(\rho_2, \rho_1) \le 0$ for any ρ_2 and ρ_1, and $D_{\mathrm{rel}}(\rho_2, \rho_1) = 0$ if and only if $\rho_2 = \rho_1$. More specifically, the functional $-D_{\mathrm{rel}}(\rho_2, \rho_1)$ is referred to as the Kullback–Leibler divergence [20]; it defines a premetric on the set of probability densities but is not a metric since it is nonsymmetric and does not satisfy the triangle inequality. Following [25], we prefer to work with $D_{\mathrm{rel}}(\rho_2, \rho_1)$ rather than $-D_{\mathrm{rel}}(\rho_2, \rho_1)$ purely for notational consistency with (5). As in the absolute entropy case, the relative entropy can also be interpreted as a measure of uniformity; densities ρ_2 within a prescribed class that maximize $D_{\mathrm{rel}}(\rho_2, \rho_1)$ for given ρ_1 can be interpreted as the most uniform with respect to ρ_1. In the special case when ρ_2 and ρ_1 are both Gaussian, with stiffnesses K_2 and K_1, and means μ_2 and μ_1, the integral in (7) can be explicitly evaluated as before to obtain

$$\begin{aligned} D_{\mathrm{rel}}(\rho_2, \rho_1) = \frac{1}{2}\Big[\ln\left(\det\mathsf{K}_1/\det\mathsf{K}_2\right) - \mathsf{K}_2^{-1} : \mathsf{K}_1 + I : I\Big] \\ - \frac{1}{2}(\mu_2 - \mu_1)\cdot\mathsf{K}_1(\mu_2 - \mu_1). \end{aligned} \tag{8}$$

2.3. Estimation via maximum absolute entropy. Both absolute and relative entropies have been widely employed in various parameter estimation methods in statistics [3, 6, 10, 36, 38] and statistical mechanics [16, 17, 18]. For instance, the maximization of the absolute entropy $D_{\mathrm{abs}}(\rho_{\mathrm{m}})$ over a class of model densities ρ_{m} yields a best-fit density $\rho_\diamond$ within the class; such a density is a best-fit in the sense that it maximizes entropy, and can be understood as being a most uniform (or least biased) density in the class. This approach is referred to as model fitting via the maximum entropy principle, to which we will sometimes add the adjective *absolute* to emphasize the distinction from the relative entropy case. Both the functional form and parameters of a best-fit density can be obtained from the maximum entropy principle approach, and analytic solutions are known in some cases.

The simplest case of interest is when the admissible class is defined as all normalized densities ρ_{m} whose mean is a prescribed vector μ, and whose covariances are completely prescribed as a symmetric, positive-definite matrix C, so that a best-fit density satisfies

$$\rho_\diamond := \underset{\rho_{\mathrm{m}} \in R}{\operatorname{argmax}} \; D_{\mathrm{abs}}(\rho_{\mathrm{m}}), \tag{9}$$

where R is a set of smooth, normalized model density functions given by

$$R = \{\rho_{\mathrm{m}} \mid \mu_{\rho_{\mathrm{m}}} = \mu, \quad \mathsf{C}_{\rho_{\mathrm{m}}} = \mathsf{C}\}. \tag{10}$$

In this simple case, it is well known (see, e.g., [25]) that a best-fit density $\rho_\diamond$ must necessarily take the Gaussian form (4) for some vector $\mu_\diamond$ and symmetric, positive-definite matrix $\mathsf{K}_\diamond$. Moreover, the parameters $\mu_\diamond$ and $\mathsf{K}_\diamond$ are explicitly related to the constraint data μ and C of the maximization; specifically,

$$\mu_\diamond = \mu, \qquad \mathsf{K}_\diamond = \mathsf{C}^{-1}. \tag{11}$$

If it is instead assumed a priori that the density to be approximated is Gaussian of the form (4), then the same best-fit parameters (11) could instead be obtained from the maximum likelihood principle applied to a finite ensemble of data with sample mean μ and covariance C.

We remark that whether or not the best-fit model density $\rho_\diamond(\mathsf{w})$ is a good approximation to the assumed underlying density $\rho(\mathsf{w})$ is another matter entirely; it is related to properties of the higher-order moments of $\rho(\mathsf{w})$, and, if the data μ and C were estimated from an associated time series, to properties of that time series, such as stationarity and ergodicity.

2.4. Banded models via maximum relative entropy estimation. In contrast to absolute entropy, models can also be fit based on the notion of relative entropy. For instance, whenever an observed density ρ_{o} is available as a prior, the maximization of the relative entropy $D_{\mathrm{rel}}(\rho_{\mathrm{m}}, \rho_{\mathrm{o}})$ over a class of model densities ρ_{m} yields a best-fit density, which has been termed model fitting via the maximum relative entropy principle; see, e.g., [25]. Analytic solutions for best-fit densities defined using the maximum relative entropy principle do not appear to be known for the classes of model densities to be considered here; indeed, questions of existence and uniqueness can be delicate, and we proceed only formally. However, we note that, at least in the context of our applications, the development of robust numerical routines for computing Gaussian, maximum relative entropy best-fit densities is straightforward [13].

In general, the stiffness matrix $\mathsf{K}_\diamond = \mathsf{C}^{-1}$ arising in (11) will be dense, but motivated by the particular data C arising in our DNA application, we now introduce the further assumption that there is a banded index set $\mathcal{N}$ for which all entries of $\mathsf{K}_\diamond$ with indices in the complementary index set $\mathcal{N}^c$ are small. In order to simplify by reducing the number of nonvanishing parameters in the model, it is then natural to seek a best-fit Gaussian density whose associated stiffness matrix is constrained to be banded, with a sparsity pattern corresponding to the index set $\mathcal{N}$. In our previous work [13], we obtained such best-fit densities using a Gaussian version of the maximum relative entropy principle; specifically, we set

$$\rho_{\mathrm{rel}} := \underset{\rho_{\mathrm{m}} \in R}{\operatorname{argmax}} \; D_{\mathrm{rel}}(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}), \tag{12}$$

where R is a set of normalized Gaussian model density functions with symmetric, positive-definite, banded stiffness matrices defined as

$$R = \left\{ \rho_{\mathrm{m}} \mid \rho_{\mathrm{m}} = \frac{1}{Z_{\mathrm{m}}} e^{-\frac{1}{2}(\mathsf{w}-\mu_{\mathrm{m}})\cdot\mathsf{K}_{\mathrm{m}}(\mathsf{w}-\mu_{\mathrm{m}})}, \quad \mathsf{K}_{\mathrm{m}}^T = \mathsf{K}_{\mathrm{m}} > 0, \quad [\mathsf{K}_{\mathrm{m}}]_{\mathcal{N}^c} = 0 \right\}, \tag{13}$$

where we took the observed or prior density ρ_{o} to be the Gaussian $\rho_{\mathrm{o}} = \rho_{\diamond}$ with positive-definite, dense stiffness matrix $\mathsf{K}_{\mathrm{o}} = \mathsf{K}_{\diamond}$, arising from the maximum absolute entropy fit (11). Since the (i, j) entry in the stiffness matrix being nonzero indicates that the i^{th} and j^{th} entries in the configuration coordinate vector are coupled in the free energy, we note that an assumption of only nearest-neighbor interactions is equivalent to an appropriate choice of the index set $\mathcal{N}$.

The maximization problem (12) can be simplified. Specifically, in view of the explicit, decoupled expression in (8) for the relative entropy in the Gaussian case, we can immediately conclude that the mean vector for any maximizing density must be $\mu_{\mathrm{rel}} = \mu_{\mathrm{o}}$, which follows from the fact that K_{o} is positive-definite. As a result, the stiffness matrix for any maximizing density must satisfy

$$\mathsf{K}_{\mathrm{rel}} = \underset{\substack{\mathsf{K}_{\mathrm{m}}^T = \mathsf{K}_{\mathrm{m}} > 0 \\ [\mathsf{K}_{\mathrm{m}}]_{\mathcal{N}^c} = 0}}{\operatorname{argmax}} \; \frac{1}{2}\Big[\ln\left(\det \mathsf{K}_{\mathrm{o}} / \det \mathsf{K}_{\mathrm{m}}\right) - \mathsf{K}_{\mathrm{m}}^{-1} : \mathsf{K}_{\mathrm{o}} + I : I \Big]. \tag{14}$$

As described in [13], and at least for our specific DNA application data, the characterization (14) of $\mathsf{K}_{\mathrm{rel}}$ was found to be amenable to different iterative numerical optimization algorithms, each of which converged to the same optimizer for various different initial guesses. Hence for each set of input data (μ, C), we obtained a best-fit Gaussian density $\rho_{\mathrm{rel}}(\mathsf{w})$ with mean vector $\mu_{\mathrm{rel}} = \mu$ and banded stiffness matrix $\mathsf{K}_{\mathrm{rel}}$. And we note that all entries of the observed covariance C enter into the characterization (14).

2.5. Banded models via maximum absolute entropy estimation. We now observe that a best-fit Gaussian density with a banded stiffness matrix can be characterized in a different way, namely via an appropriate maximum *absolute* entropy principle. Specifically, for input data (μ, C), we may consider a best-fit density defined by

$$\rho_{\mathrm{abs}} := \underset{\rho_{\mathrm{m}} \in R}{\operatorname{argmax}} \; D_{\mathrm{abs}}(\rho_{\mathrm{m}}), \tag{15}$$

where R is a set of smooth, normalized model density functions, whose functional form and parameters are arbitrary, but whose mean and covariance are constrained by the definition

$$R = \{\rho_{\mathrm{m}} \mid \mu_{\rho_{\mathrm{m}}} = \mu, \quad [\mathsf{C}_{\rho_{\mathrm{m}}}]_{\mathcal{N}} = [\mathsf{C}]_{\mathcal{N}}\}. \tag{16}$$

Thus the best-fit density is the one that maximizes the absolute entropy functional over all model densities that are consistent with the estimated mean vector μ, and the subset $[\mathsf{C}]_{\mathcal{N}}$ of the estimated covariance matrix associated with the index set $\mathcal{N}$. When the index set $\mathcal{N}$ corresponds to the entire matrix, then (15) and (16) reduce to (9) and (10). However, whenever the index set $\mathcal{N}$ does not correspond to the entire matrix, then the covariances outside the index set play no role in the fit, which is in contrast to the relative entropy fit (14), which involves all entries of the estimated covariance matrix C for all stencils $\mathcal{N}$.

The maximization problem in (15) and (16) is well studied [3, 6, 10, 36, 38]. By a slight generalization of a classic result [6], provided that the subset $[\mathsf{C}]_{\mathcal{N}}$ of the estimated covariance includes all of the diagonal entries (variances) and is known to be drawn from a symmetric, positive-definite matrix, both of which conditions are satisfied in our case, a best-fit density $\rho_{\text{abs}}(\mathsf{w})$ exists, is unique, and is a Gaussian with mean vector $\mu_{\text{abs}} = \mu$ and a symmetric, positive-definite stiffness matrix K_{abs} which is banded according to the nearest-neighbor index set $\mathcal{N}$; specifically, we have $[\mathsf{K}_{\text{abs}}]_{\mathcal{N}^c} = 0$, where $\mathcal{N}^c$ is the index set complementary to $\mathcal{N}$. Moreover, the inverse of the model stiffness (i.e., the model covariance) $\mathsf{K}_{\text{abs}}^{-1}$ will in general be dense and, as required, will exactly coincide with the estimated covariance C within $\mathcal{N}$, that is, $[\mathsf{K}_{\text{abs}}^{-1}]_{\mathcal{N}} = [\mathsf{C}]_{\mathcal{N}}$, but with no equality necessary between any entries with indices in $\mathcal{N}^c$. In summary, an absolute entropy best-fit density ρ_{abs} as defined in (15) is known to exist and to be a shifted Gaussian with a banded stiffness matrix with parameters determined in terms of the prescribed data by

$$\mu_{\text{abs}} = \mu, \qquad [\mathsf{K}_{\text{abs}}]_{\mathcal{N}^c} = 0, \qquad [\mathsf{K}_{\text{abs}}^{-1}]_{\mathcal{N}} = [\mathsf{C}]_{\mathcal{N}}. \tag{17}$$

It is simple to calculate that, analogously to the unbanded case, if it is instead assumed a priori that the model density to be approximated is Gaussian with a banded stiffness matrix, then the same conditions (17) for the best-fit parameters could instead be obtained from the maximum likelihood principle applied to a finite ensemble of data with sample mean μ and covariance C. Similarly, and as already remarked in [13], if the model density is again assumed to be Gaussian with a banded stiffness matrix, then the necessary conditions (17) can be obtained in a third way, namely from the version of the relative entropy variational principle (12) in which the order of the model ρ_{m} and observed ρ_{o} densities are switched in the two arguments of the (nonsymmetric) functional D_{rel}. Nevertheless we will continue to describe the Gaussian parameters $(\mu_{\text{abs}}, \mathsf{K}_{\text{abs}})$ defined by (17) as the absolute entropy parameters.

Solving conditions (17) for the stiffness matrix is not entirely straightforward, and for example iterative algorithms have been proposed [6, 36]. What is less well known is that, in the case when the index set $\mathcal{N}$ corresponds to an overlapping diagonal block structure (as is the case in our *cgDNA* application), there is a simple, explicit, and local construction of the matrix K_{abs} satisfying (17) directly from the covariances $[\mathsf{C}]_{\mathcal{N}}$. There are two particularly simple special cases. First, and as previously remarked, when $\mathcal{N}$ is the entire matrix, then $\mathsf{K}_{\text{abs}} = \mathsf{C}^{-1}$. Second, in the uncoupled case where $\mathcal{N}$ has nonoverlapping diagonal blocks (not necessarily all of the same size, but containing all diagonal entries), then it is simple to verify that the full matrix case applies to the decoupled blocks, and the nonvanishing blocks of K_{abs} are just the inverses of the corresponding diagonal blocks of C, so that in addition $[\mathsf{K}_{\text{abs}}]_{\mathcal{N}^c} = 0$ is satisfied. We are interested in the general fully coupled case where the index set $\mathcal{N}$ corresponds to b overlapping diagonal blocks, with $(b-1)$ nontrivial adjacent overlaps (the special case where $\mathcal{N}$ corresponds to 18×18 diagonal blocks with 6×6 overlaps that arises in our DNA application is illustrated in the rightmost image in Figure 8 below), for which the construction is as follows. Each of the b diagonal (in our particular case 18×18) subblocks of the covariance is inverted and written to the corresponding subblock in the stiffness matrix, with all contributions being added in overlap regions between two or more blocks. Then the blocks corresponding to the $(b-1)$ (in our particular case 6×6) overlaps between adjacent subblocks are inverted and subtracted from the corresponding subblocks in the stiffness matrix. The stiffness matrix K_{abs} so constructed is the unique matrix satisfying conditions (17). The proof of this somewhat remarkable result relies on a recursive Schur complement factorization, the essence of which appears in both [22, sec. 5.3] and [19], where in each case the statement of the result is somewhat complicated by the fact that in the respective application fields of graphical models and matrix completion, there is no intrinsic ordering of the variables w so that block-bandedness is not a natural language in which to state a hypothesis. An equivalent proof in purely linear algebra terms is provided in [12]. We immediately obtain by this construction a best-fit Gaussian density $\rho_{\text{abs}}(\mathsf{w})$ with mean vector μ_{abs} and banded stiffness matrix K_{abs}.

3. Rigid-base coarse-grain oligomer models.

We next apply the theory outlined in section 2 to obtain coarse-grain, banded-Gaussian models for each oligomer in a library of DNA sequences $\{\mathsf{S}_\nu\}_{\nu=1}^N$ for which estimates of the means $\mu(\mathsf{S}_\nu)$ and covariances $\mathsf{C}(\mathsf{S}_\nu)$ are available from fine-grain MD simulations.

3.1. Coarse-grain coordinates.

We consider right-handed, double-helical DNA in which the four possible bases $\texttt{T}$, $\texttt{A}$, $\texttt{C}$, and $\texttt{G}$ are attached to two, oriented, antiparallel backbone strands and form only the standard Watson–Crick pairs $(\texttt{A},\texttt{T})$ and $(\texttt{C},\texttt{G})$. Choosing one backbone strand as a reference, which we will refer to as the Watson strand, a DNA oligomer comprising n base pairs is identified with a sequence of bases $\mathsf{S} := \texttt{X}_1\texttt{X}_2\cdots\texttt{X}_n$, listed in the $5'$ to $3'$ direction along the strand, where $\texttt{X}_a \in \{\texttt{T},\texttt{A},\texttt{C},\texttt{G}\}$ for $a = 1,\ldots,n$. The base pairs associated with this sequence are denoted by $(\texttt{X},\overline{\texttt{X}})_1,\ldots,(\texttt{X},\overline{\texttt{X}})_n$, where $\overline{\texttt{X}}$ is defined as the Watson–Crick complement of $\texttt{X}$ as illustrated in Figure 1. The notation $(\texttt{X},\overline{\texttt{X}})_a$ for a base pair indicates that base $\texttt{X}$ is attached to the reference (or Watson) strand, while $\overline{\texttt{X}}$ is attached to the complementary (or Crick) strand, and there are four possible base pairs $(\texttt{X},\overline{\texttt{X}})_a$ corresponding to the choice $\texttt{X}_a \in \{\texttt{T},\texttt{A},\texttt{C},\texttt{G}\}$. The length of a sequence S will mean the number n of base pairs in the sequence, denoted $|\mathsf{S}| := n$.

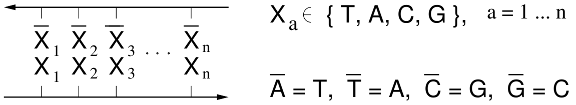
$\overline{X}_1\ \overline{X}_2\ \overline{X}_3 \ldots \overline{X}_n$
$X_1\ X_2\ X_3\ \ \ X_n$
$X_a \in \{T, A, C, G\}$, a = 1 ... n
$\overline{A} = T$, $\overline{T} = A$, $\overline{C} = G$, $\overline{G} = C$

FIG. 1. *Labeling of DNA bases.* $\texttt{X}_1\texttt{X}_2\cdots\texttt{X}_n$ *denote bases on the reference or Watson strand, while* $\overline{\texttt{X}}_1\overline{\texttt{X}}_2\cdots\overline{\texttt{X}}_n$ *denote bases on the antiparallel complementary or Crick strand, with each strand oriented according to its own* $5' \to 3'$ *direction as set by the detailed chemistry of the sugar rings.*

We adopt a coarse-grain description of DNA [7, 9, 21, 28] in which each base is modeled as a rigid entity; the backbones themselves are not considered (or observed) explicitly. Thus the configuration of an oligomer is equivalent to the configuration of all of its constituent bases as illustrated in Figure 2. The configuration of an arbitrary base is specified by giving the position of a reference point fixed in the base, and the orientation of a right-handed, orthonormal frame attached to the base. We define the reference point and frame for each base according to the *Curves+* implementation [23] of the Tsukuba convention [28]. In the model, the positions of the nonhydrogen atoms in each base 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 the nonhydrogen atoms.

The three-dimensional configuration of a DNA oligomer is defined by the relative rotation and displacement between neighboring bases both across and along the two backbone strands. To describe an arbitrary configuration, we first consider the reference point and frame for each base $\texttt{X}_a$ and each complementary base $\overline{\texttt{X}}_a$ assigned by the convention mentioned above. We then consider a reference frame for each base

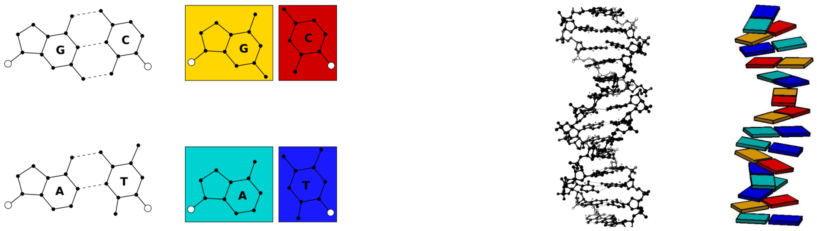

Fig. 2. *Coarse-grain rigid-base model of DNA. Each base is modeled as a rigid entity, with position and orientation determined by a reference point and frame (not shown) attached to the base; the backbone strands are not explicitly included.*

pair $(\mathtt{X},\bar{\mathtt{X}})_a$ defined via an appropriate average of the two base frames, and we also consider a reference frame for the junction between each pair of base pairs $(\mathtt{X},\bar{\mathtt{X}})_a$ and $(\mathtt{X},\bar{\mathtt{X}})_{a+1}$ defined via an appropriate average of the two base-pair frames, as illustrated in Figure 3. The relative rotation and displacement between the bases $\mathtt{X}_a$ and $\bar{\mathtt{X}}_a$ across the strands are then described in the associated base-pair frame by an *intra*-base-pair (Cayley or Gibbs) coordinate vector $y^a \in \mathbb{R}^6$ with entries comprising three rotation coordinates (Buckle-Propeller-Opening) and three displacement coordinates (Shear-Stretch-Stagger). Similarly, the relative rotation and displacement between the base pairs $(\mathtt{X},\bar{\mathtt{X}})_a$ and $(\mathtt{X},\bar{\mathtt{X}})_{a+1}$ along the strands is described in the associated junction frame by an *inter*-base-pair (Cayley or Gibbs) coordinate vector $z^a \in \mathbb{R}^6$ with entries comprising three rotation coordinates (Tilt-Roll-Twist) and three displacement coordinates (Shift-Slide-Rise), as illustrated in Figure 4. For an oligomer of n base pairs, there are a total of n intra-base-pair coordinate vectors y^a, and a total of $n-1$ inter-base-pair coordinate vectors z^a, and the collection of all coordinates is denoted by

$$\mathsf{w} := (y^1, z^1, y^2, z^2, \dots, z^{n-1}, y^n) \in \mathbb{R}^{12n-6}. \tag{18}$$

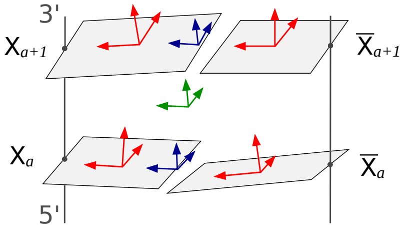

Fig. 3. *Reference frames arising in the rigid-base model of DNA. Shown is an arbitrary pair of base pairs $(\mathtt{X},\bar{\mathtt{X}})_a$ and $(\mathtt{X},\bar{\mathtt{X}})_{a+1}$: a frame is embedded in each base (four shown), a frame intermediate to the pair of base frames is embedded in each base pair (two shown), and a frame intermediate to the pair of base-pair frames is embedded in each junction or space between consecutive base pairs (one shown).*

We will use the specific nondimensional and scaled version of the above coordinates that is fully described in either [13] or [32], which includes an energy scale such that the Boltzmann factor k_BT is equal to unity. Notice that the complete configuration of a DNA oligomer is specified by introducing a vector z^0 of six additional coordinates that specify the position and orientation of the oligomer 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

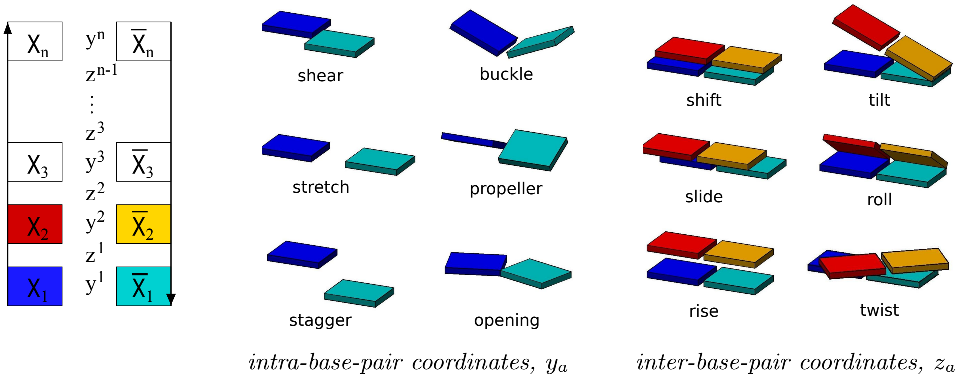

FIG. 4. *Configuration coordinates for the rigid-base model of DNA: labeling of intra- and inter-base-pair coordinate vectors (left), illustration of intra-base-pair coordinates (center), illustration of inter-base-pair coordinates (right).*

there is no external potential field. Hence the coordinate vector w is a full set of internal coordinates that completely characterizes the coarse-grain shape of a DNA oligomer. One reason for employing rotational coordinates of the Cayley type is that the configuration space for an oligomer can be taken as the entirety of $\mathbb{R}^{12n-6}$, which is mathematically convenient for the evaluation of various integrals; this would not be the case with some other types of rotational coordinates.

As described in section 2, we assume that all densities and related entropies are defined with respect to the standard measure $d\mathsf{w}$ on the configuration space $\mathbb{R}^m$. Due to the presence of rotational coordinates, it would be more precise to adopt another measure of the form $J(\mathsf{w})\,d\mathsf{w}$ that naturally arises, where $J(\mathsf{w})$ is a Jacobian factor [39] associated with the Haar measure on the rotation group $SO(3)$. However, there is increasing evidence that the effect of such a Jacobian factor on various configuration space integrals is rather small [13, 21, 27] for the length scales considered here, which is presumably due to the overall stiffness of DNA. Thus the Jacobian factor in our singularity-free, Cayley, rotational coordinates can reasonably be approximated as being constant as a simplifying assumption.

3.2. Symmetry and independence. In our later deliberations it will be important to consider the Watson–Crick symmetry associated with switching the backbone that is used as the reference strand for a given DNA oligomer. Specifically, if the sequence along the Watson strand is $\mathsf{S} = \mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$, then the sequence along the antiparallel Crick strand is the complement $\overline{\mathsf{S}} = \overline{\mathtt{X}}_n\overline{\mathtt{X}}_{n-1}\cdots\overline{\mathtt{X}}_1$. Moreover, for any given configuration of the oligomer, the coordinates w defined with respect to the Watson strand, and the coordinates $\overline{\mathsf{w}}$ defined with respect to the antiparallel Crick strand, are related by

$$\overline{\mathsf{w}} = \mathsf{E}_n\mathsf{w}, \tag{19}$$

where n is the length of the sequence S, and $\mathsf{E}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a block, trailing-diagonal matrix formed by $2n-1$ copies of the constant, diagonal matrix given by

$\mathsf{E} := \mathrm{diag}(-1, 1, 1, -1, 1, 1) \in \mathbb{R}^{6\times 6}$. Specifically, we have

$$\mathsf{E}_n := \begin{pmatrix} & & & \mathsf{E} \\ & & \mathsf{E} & \\ & \iddots & & \\ \mathsf{E} & & & \end{pmatrix} \quad \text{with the properties} \quad \mathsf{E}_n = \mathsf{E}_n^T = \mathsf{E}_n^{-1}. \tag{20}$$

As discussed in [21], the above relations are a straightforward consequence of the Watson–Crick symmetry of DNA. Their particularly simple form arises because of the introduction of the junction frame (see Figure 3) and can be described as follows. Characterize the twelve types of coordinates as being odd or even, where the odd coordinates are Buckle, Shear, Tilt, and Shift (one each of intra- and inter-base-pair, and one each of translation and rotation), and the remaining eight are all even. Then, under a change of reference strand, the coordinate blocks are reversed in order, and the odd coordinates of any configuration at any specific physical location along the oligomer change sign, whereas the even coordinates remain unaltered.

The number of sequences of a given length that are independent can now be explicitly counted. To begin, there are a total of 4^n different sequences S of a given length n. If n is odd, then it is always the case that $\overline{\mathsf{S}} \neq \mathsf{S}$, and of the 4^n total sequences, only $4^n/2$ are independent. However, when n is even, the possibility of palindromic, or self-symmetric, sequences with $\overline{\mathsf{S}} = \mathsf{S}$ arises, and the count is more complicated. For a palindromic sequence we note that the first $n/2$ bases on the Watson strand can be chosen arbitrarily, and then the remaining bases are determined by the Watson–Crick pairing rules (see Figure 1). From this we deduce that of the 4^n total sequences, there are $4^{n/2} = 2^n$ that are palindromic, and $4^n - 2^n$ that are nonpalindromic. And we note that all of the palindromic sequences, but only half of the nonpalindromic sequences, are independent. Hence, when n is even, there are $2^n + (4^n - 2^n)/2 = (2^n + 4^n)/2$ independent sequences.

The case of dinucleotide sequences ($n = 2$) is very well known in the DNA literature: there are a total of 16 possible dinucleotides, only 10 of which are independent; a complete group of independent dinucleotides consists of 4 palindromic and 6 nonpalindromic ones, where each nonpalindrome is a single representative from a pair of complementary dinucleotides. The case of tetranucleotide sequences ($n = 4$) is also of interest: there are a total of 256 possible tetranucleotides, only 136 of which are independent, and a complete group of independent tetranucleotides comprises 16 palindromic and 120 nonpalindromic ones, where as before each nonpalindrome is a single representative of a complementary pair of sequences.

Because of the Watson–Crick symmetry associated with the choice of reference strand, the density $\rho(\mathsf{w}; \mathsf{S})$ for an arbitrary sequence S is related to the density $\rho(\overline{\mathsf{w}}; \overline{\mathsf{S}})$ for the complementary sequence $\overline{\mathsf{S}}$ through the change of variable formula in (19). As a result, the corresponding means and covariances are necessarily related; specifically, we have

$$\mu_\rho(\mathsf{S}) = \mathsf{E}_n \mu_\rho(\overline{\mathsf{S}}), \qquad \mathsf{C}_\rho(\mathsf{S}) = \mathsf{E}_n \mathsf{C}_\rho(\overline{\mathsf{S}}) \mathsf{E}_n, \tag{21}$$

where n is the length of S. For nonpalindromic sequences, (21) specifies the mean and covariance for the sequence S in terms of those for the complementary sequence $\overline{\mathsf{S}}$, while for palindromes with $\overline{\mathsf{S}} = \mathsf{S}$, (21) becomes the constraints

$$\mu_\rho(\mathsf{S}) = \mathsf{E}_n \mu_\rho(\mathsf{S}), \qquad \mathsf{C}_\rho(\mathsf{S}) = \mathsf{E}_n \mathsf{C}_\rho(\mathsf{S}) \mathsf{E}_n \tag{22}$$

on the possible values of palindromic means and covariances. These constraints express the fact that for palindromic sequences the Watson and Crick strands are physically indistinguishable.

3.3. Training set data for oligomer models. The parameterization of our coarse-grain model will require data to estimate the density $\rho(\mathsf{w};\mathsf{S})$ for a variety of sequences S. The type of data that we consider will be in the form of an ensemble of observed coordinate vectors $\{\mathsf{w}^{[j]}(\mathsf{S}_\nu)\}_{j=1}^{M_\nu}$ for each sequence in a given library of sequences $\{\mathsf{S}_\nu\}_{\nu=1}^{N}$. The ensemble $\{\mathsf{w}^{[j]}(\mathsf{S}_\nu)\}$ is assumed to sample the equilibrium distribution described by $\rho(\mathsf{w};\mathsf{S}_\nu)$ for each sequence S_ν. We note that the ensemble size M_ν and the sequence length $|\mathsf{S}_\nu|$ may in general vary with ν.

We generate the coordinate ensembles $\{\mathsf{w}^{[j]}(\mathsf{S}_\nu)\}$ using extensive databases [24] of atomic-resolution MD simulations of DNA oligomers in explicit solvent. This is mathematically straightforward as the extraction of a coordinate vector at any time snapshot of the MD simulation is a nonlinear projection or fit. As a practical matter, we also filter the coordinate ensembles by eliminating outliers using various knowledge-based methods. For example, for any sequence, we typically eliminate any MD snapshot that has any inter-base hydrogen bond broken according to standard criteria [13]. In principle, the coordinate ensembles could instead be generated directly from appropriate experiments such as NMR, but there are of course serious issues of resolution in the available experimental data, and we have not yet pursued that avenue in any detail.

Once a coordinate ensemble $\{\mathsf{w}^{[j]}(\mathsf{S}_\nu)\}$ has been assembled, the estimated mean $\mu(\mathsf{S}_\nu)$ and covariance $\mathsf{C}(\mathsf{S}_\nu)$ of the ensemble are defined in the standard way:

$$\mu(\mathsf{S}_\nu) := \frac{1}{M_\nu}\sum_{j=1}^{M_\nu}\mathsf{w}^{[j]}(\mathsf{S}_\nu), \tag{23}$$

$$\mathsf{C}(\mathsf{S}_\nu) := \frac{1}{M_\nu}\sum_{j=1}^{M_\nu}(\mathsf{w}^{[j]}(\mathsf{S}_\nu)-\mu(\mathsf{S}_\nu))\otimes(\mathsf{w}^{[j]}(\mathsf{S}_\nu)-\mu(\mathsf{S}_\nu)). \tag{24}$$

Notice that $\mu(\mathsf{S}_\nu)$ and $\mathsf{C}(\mathsf{S}_\nu)$ denote known, estimated values based on a finite ensemble, whereas $\mu_\rho(\mathsf{S}_\nu)$ and $\mathsf{C}_\rho(\mathsf{S}_\nu)$ denote unknown, exact values corresponding to the underlying density. As in the exact case, the vector $\mu(\mathsf{S}_\nu)$ can be arbitrary, while the matrix $\mathsf{C}(\mathsf{S}_\nu)$ is symmetric, and can be verified to be positive-definite for each S_ν (essentially due to our large sample size).

Because of Watson–Crick symmetry, an ensemble of configurations $\{\mathsf{w}^{[j]}(\mathsf{S})\}$ for an arbitrary sequence S on the Watson strand immediately generates an ensemble $\{\overline{\mathsf{w}}^{[j]}(\overline{\mathsf{S}})\}$ for the complementary sequence $\overline{\mathsf{S}}$ on the Crick strand, where each $\mathsf{w}^{[j]}(\mathsf{S})$ and $\overline{\mathsf{w}}^{[j]}(\overline{\mathsf{S}})$ are related according to (19). For nonpalindromes with $\overline{\mathsf{S}}\neq\mathsf{S}$, the estimated means and covariances of the ensembles $\{\mathsf{w}^{[j]}(\mathsf{S})\}$ and $\{\overline{\mathsf{w}}^{[j]}(\overline{\mathsf{S}})\}$ will by construction satisfy a corresponding version of the relation in (21); specifically,

$$\mu(\mathsf{S}) = \mathsf{E}_n\mu(\overline{\mathsf{S}}), \qquad \mathsf{C}(\mathsf{S}) = \mathsf{E}_n\mathsf{C}(\overline{\mathsf{S}})\mathsf{E}_n. \tag{25}$$

However, in the case of palindromes with $\overline{\mathsf{S}}=\mathsf{S}$ there is no reason why the estimates in (23) and (24) drawn from a finite ensemble of observations $\{\mathsf{w}^{[j]}(\mathsf{S})\}$ should satisfy the conditions in (22). In fact there are two possibilities for palindromic sequences. Either the ensemble $\{\mathsf{w}^{[j]}(\mathsf{S})\}$ can be doubled in size by appending the ensemble $\{\overline{\mathsf{w}}^{[j]}(\mathsf{S})\}$ of configurations read from the physically indistinguishable second strand, in which case the symmetry conditions in (22) are automatically satisfied (or equivalently the estimates are just appropriately symmetrized), or the original ensemble $\{\mathsf{w}^{[j]}(\mathsf{S})\}$ is retained, and the degree to which the symmetry conditions in (22) are satisfied provides an indication of the convergence of the estimates in (23) and (24) to the underlying exact values.

The set of estimated means $\mu(\mathsf{S}_\nu) \in \mathbb{R}^{m_\nu}$ and covariances $\mathsf{C}(\mathsf{S}_\nu) \in \mathbb{R}^{m_\nu \times m_\nu}$ for all sequences in the ensemble $\{\mathsf{S}_\nu\}_{\nu=1}^N$ will be referred to as our training data set. In this work, we consider as an example a particular training data set based on $N = 53$ distinct sequences S_ν, each of length $n_\nu = 12$ or 18 base pairs, with corresponding configuration space of dimension $m_\nu = 138$ or 210. The coordinate ensemble $\{\mathsf{w}^{[j]}(\mathsf{S}_\nu)\}_{j=1}^{M_\nu}$ for each sequence had a size M_ν (after filtering) of the order 10^3–10^5. Our training data set was generated from a database of MD simulations of the enhanced set of ABC [1, 8, 24] sequences as described in detail in [13]. The MD simulations were performed using standard conditions and protocols, and were of 50 to 100 or more nanoseconds in duration for each sequence. The ABC sequences have the feature that they contain multiple instances of all 136 distinct tetranucleotide subsequences within their interiors, but they have only $5'$-GC and GC-$3'$ ends. Our enhancements to the sequences in the ensemble are to have access to data for a wider set of end sequences. Specifically, with the additional sequences, all 16 possible $5'$-dinucleotide-step ends and all 16 possible dinucleotide-step-$3'$ ends are represented. The sequence ensemble contains six palindromes, for which the estimates of the means and covariances were not symmetrized. For a complete listing of all the sequences in the ensemble $\{\mathsf{S}_\nu\}_{\nu=1}^N$, see [13].

It was first observed in [21] for one sequence, and later confirmed and quantified in [13] for all the sequences in our training data set, that while our observed covariances $\mathsf{C}(\mathsf{S}_\nu)$ are dense, albeit with entries decaying with distance from the diagonal, the corresponding stiffness matrices have rather small entries outside a stencil formed by 18×18 diagonal blocks, each of which overlaps (away from the oligomer ends) its two neighbors in the 6×6 diagonal blocks corresponding to every instance of the intra-base-pair coordinates y^a (cf. Figures 5 and 6). As detailed in [13, 21] and sections 4.1 and 4.2, with our choice and ordering of configuration coordinates the overlapping 18×18 block sparsity pattern in the stiffness matrices corresponds to each base being directly coupled to only its five nearest neighbors, one upstream and one downstream on the same backbone, along with the three corresponding paired base partners. Accordingly, as empirically suggested by the observed data, and physically motivated in terms of nearest-neighbor couplings between bases, we adopt the stencil formed by the overlapping 18×18 blocks as our index set $\mathcal{N}$, refer to it as the rigid-base nearest-neighbor index set, and enforce it in our oligomer-based banded Gaussian models.

We remark that various other choices of coarse-grain models, such as a rigid-base-pair description with nearest-neighbor couplings, are not similarly supported by the data observed in our training set. For instance, as noted in [21], when the submatrix of a covariance $\mathsf{C}(\mathsf{S}_\nu)$ corresponding to only the $(n-1)$ inter-base-pair coordinate vectors z^a is inverted, the resulting rigid-base-pair stiffness matrix (which is also the stiffness in the marginal of the Gaussian rigid-base model over the intra-coordinates) is very far from having the 6×6 block diagonal structure corresponding to a nearest-neighbor Gaussian model in the inter-base-pair coordinates.

3.4. Comparison of absolute and relative entropy banded models.

Now we may apply the maximum relative and absolute entropy fitting principles, as described in sections 2.4 and 2.5, to obtain two different best-fit Gaussian models with banded stiffness matrices for the specific rigid-base nearest-neighbor index set $\mathcal{N}$, and for each training set sequence S_ν. In principle, other metrics or divergences between distributions could be contemplated to generate other banded Gaussian approximations, but we will not pursue them here except to briefly consider a third, and most direct, approach based on simply cutting or truncating the nonzero entries in the stiffness matrix $\mathsf{C}(\mathsf{S}_\nu)^{-1}$. Specifically, for each sequence S_ν, we can define a Gaussian density $\rho_{\rm cut}(\mathsf{w};\mathsf{S}_\nu)$ by

$$(26)\qquad \mu_{\rm cut}(\mathsf{S}_\nu) = \mu(\mathsf{S}_\nu), \quad [\mathsf{K}_{\rm cut}(\mathsf{S}_\nu)]_{\mathcal{N}} = [\mathsf{C}(\mathsf{S}_\nu)^{-1}]_{\mathcal{N}}, \quad [\mathsf{K}_{\rm cut}(\mathsf{S}_\nu)]_{\mathcal{N}^c} = 0.$$

We will compare properties of these three different banded Gaussian approximations.

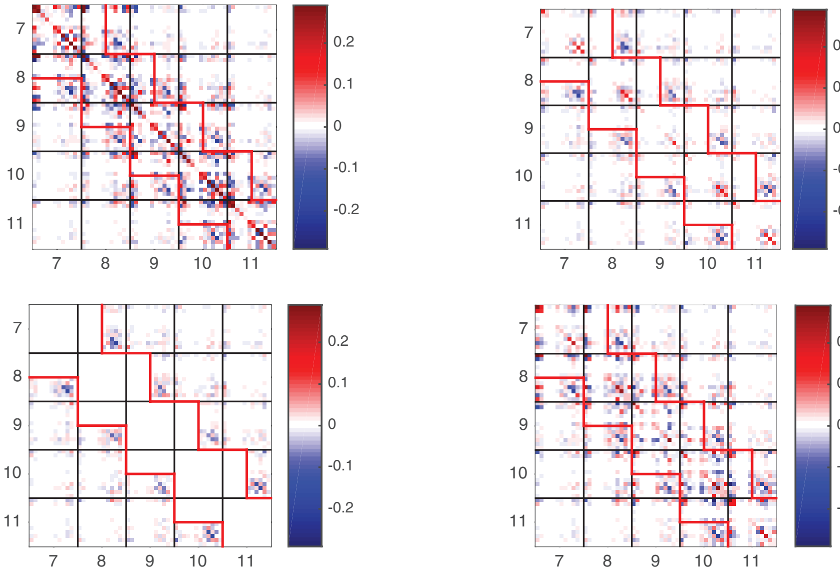

FIG. 5. *Comparisons between submatrices (corresponding to the central four of seventeen junctions) of the estimated and three different fitted covariances for the sequence* S_3*; see text. Top left:* $\mathsf{C}(\mathsf{S}_3)$*. Top right:* $\mathsf{C}(\mathsf{S}_3) - \mathsf{K}_{\rm rel}(\mathsf{S}_3)^{-1}$*. Bottom left:* $\mathsf{C}(\mathsf{S}_3) - \mathsf{K}_{\rm abs}(\mathsf{S}_3)^{-1}$*. Bottom right:* $\mathsf{C}(\mathsf{S}_3) - \mathsf{K}_{\rm cut}(\mathsf{S}_3)^{-1}$.

Figure 5 makes comparisons between the estimated covariance $\mathsf{C}(\mathsf{S}_\nu)$ and the three fitted covariances (inverse stiffnesses) $\mathsf{K}_{\rm rel}(\mathsf{S}_\nu)^{-1}$, $\mathsf{K}_{\rm abs}(\mathsf{S}_\nu)^{-1}$, and $\mathsf{K}_{\rm cut}(\mathsf{S}_\nu)^{-1}$. Results are shown for the specific training set sequence $\mathsf{S}_3 =$ `GCGCATGCATGCATGCGC`; the results are qualitatively similar for all other training set sequences S_ν. The results show that the three different banded-stiffness models generate significantly different covariance matrices. By construction, the maximum absolute entropy model has a covariance which exactly matches with the estimated covariance within the nearest-neighbor index set $\mathcal{N}$, but differs in the complementary set $\mathcal{N}^c$. In contrast, the maximum relative entropy model has a covariance that differs from the estimated covariance over the entire matrix. Nevertheless we find that these two model covariances are visually comparable over all entries. We note further that the truncation model has a covariance that is a rather poor approximation of the estimated covariance. In fact, for an overlapping stencil such as $\mathcal{N}$, there is no guarantee that the truncation operation leading to $\mathsf{K}_{\rm cut}(\mathsf{S}_\nu)$ will yield a positive-definite matrix. As it happens, for the overlapping 18×18 stencil, the truncation operation yields a positive-definite stiffness matrix for all the training set sequences, but on the same data, a similar truncation operation for an overlapping 30 × 30 stencil (corresponding to next-to-nearest-neighbor interactions) yields indefinite stiffness matrices for all the training set sequences. For both of these reasons we conclude that the fitting approach based on truncation as described here is inferior to those based on maximum relative or absolute entropy, and we consider it no further.

Figure 6 makes comparisons between the estimated stiffness (inverse covariance) $\mathsf{C}(\mathsf{S}_\nu)^{-1}$ and the fitted stiffnesses $\mathsf{K}_{\text{rel}}(\mathsf{S}_\nu)$ and $\mathsf{K}_{\text{abs}}(\mathsf{S}_\nu)$ from the two best-fit models based on entropy. Results are shown for the same sequence S_3 as before, and again the results are qualitatively similar for all other training set sequences S_ν. The two different best-fit models have noticeably different stiffness matrices. By construction, both the relative and absolute entropy stiffness matrices are banded and hence vanish outside the stencil $\mathcal{N}$. Although not shown, the truncation model stiffness $\mathsf{K}_{\text{cut}}(\mathsf{S}_3)$ is also banded and additionally agrees with $\mathsf{C}(\mathsf{S}_3)^{-1}$ inside $\mathcal{N}$ by construction, but the associated covariance is rather poor as noted above. The largest entry-by-entry absolute errors over the whole matrix are visually comparable between the relative and absolute entropy stiffness matrices, and it is interesting to note that the majority, but not all, of the eigenvalues of $[\mathsf{K}_{\text{rel}}(\mathsf{S}_3) - \mathsf{K}_{\text{abs}}(\mathsf{S}_3)]$ are positive.

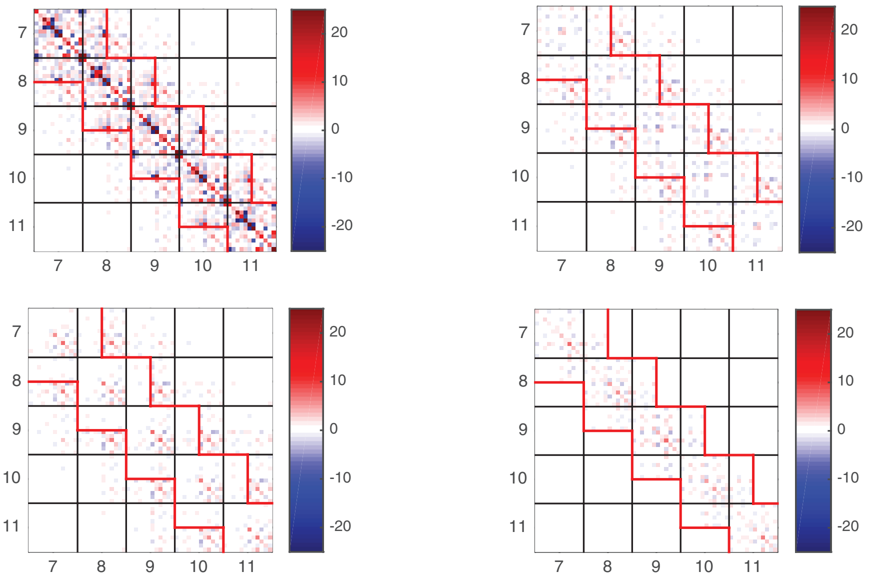

FIG. 6. *Comparisons between submatrices (corresponding to the central four of seventeen junctions) of the estimated stiffness and two stiffnesses based on absolute and relative entropy fits; see text. Top left:* $\mathsf{C}(\mathsf{S}_3)^{-1}$. *Top right:* $\mathsf{C}(\mathsf{S}_3)^{-1} - \mathsf{K}_{\text{rel}}(\mathsf{S}_3)$. *Bottom left:* $\mathsf{C}(\mathsf{S}_3)^{-1} - \mathsf{K}_{\text{abs}}(\mathsf{S}_3)$. *Bottom right:* $\mathsf{K}_{\text{rel}}(\mathsf{S}_3) - \mathsf{K}_{\text{abs}}(\mathsf{S}_3)$.

We conclude that each of the two fitting approaches based on entropy lead to reasonable, yet noticeably different, results. However, we now describe some special features which suggest that the approach based on absolute entropy is more natural for our application. We note first that the ensembles of configurations used to estimate the mean vector and covariance matrix for each sequence in our training set are generated as MD time series, and it is therefore likely that the longer the duration of the simulation, the smaller the error between the estimated and exact moments of the assumed equilibrium distribution. Moreover, as DNA is a linear polymer, we note that there is a natural ordering of the configuration variables in the covariance matrix corresponding to their positions along the molecule, and distance from the diagonal in the covariance matrix for this ordering of the variables has a physical meaning. Specifically, entries in the matrix that are close to the diagonal correspond to positions that are close together along the molecule, whereas entries far from the diagonal correspond to positions that are far apart. Furthermore, as the largest entries in the stiffness matrices are close to the diagonal, we might expect there to be high frequency, localized oscillations, which might dominate the covariances close to the diagonal, while the largest contributions to covariances far from the diagonal could be generated by lower frequency delocalized oscillations. These, purely heuristic, arguments lead us to examine the conjecture of whether for a finite time series, the estimates of the covariances closer to the diagonal have smaller errors than the estimates of the covariances further from the diagonal. The data shown in Table 1 support this conjecture.

For the specific palindromic, training-set sequence S_3, we computed eleven MD trajectories with identical simulation conditions: the original 100ns trajectory used in the training set, its extension to 1μs, and nine other independent 50ns simulations with different initial conditions for the system (including the solvent). To assess the convergence of the estimated mean $\mu(\mathsf{S}_3)$ and the estimated covariance both within $[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}$ and outside $[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}$ the nearest-neighbor index set $\mathcal{N}$, we first considered the relative errors presented in the first three rows of Table 1, where the estimates from the 1μs simulation are used in place of the unknown exact values $\mu_\rho(\mathsf{S}_3)$ and $\mathsf{C}_\rho(\mathsf{S}_3)$, so that ten relative errors can be approximated for the original 100ns simulation (column 0) and each of the additional nine independent 50ns simulations (columns 1-9).

Moreover, since the sequence S_3 is a palindrome, and because the palindromic symmetry conditions in (22) respect the stencil $\mathcal{N}$, we can additionally independently assess convergence based on the palindromic symmetry errors in $\mu(\mathsf{S}_3)$, $[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}$, and $[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}$ for each of the ten simulations. Notice that these palindromic symmetry errors are intrinsic to the estimates in the sense that they are defined independently of the unknown exact values of the quantities concerned. The palindromic symmetry errors are shown in the last three rows of Table 1.

The data for both relative and palindromic errors strongly suggest that the estimate of the mean has the smallest error, the covariances within the stencil $\mathcal{N}$ have the next smallest error, and the covariances outside $\mathcal{N}$ have the largest error. In this context the fitting approach based on maximum absolute entropy can now be seen to be more natural than the relative entropy fit: the absolute entropy fit is uniquely determined by the blocks in the estimated covariance that are within the stencil $\mathcal{N}$; in contrast, the relative entropy fit takes information from all blocks, including those in $\mathcal{N}^c$ that are far from the diagonal.

The data reported in Table 1 also suggest that longer MD time series would be desirable, which is almost always the case in practice. Indeed, the difference between estimated and exact moments of the assumed equilibrium distribution would presumably be smaller for a longer time series. However, in the remainder of this presentation we wish to focus on a better understanding of the consequences of adopting different mathematical approaches to parameter estimation starting from a common ensemble of MD data, so that we will not further discuss other MD data sets.

4. The *cgDNA* model.

The oligomer-based coarse-grain models described in the previous section are not predictive, in the sense that they can only estimate

TABLE 1

Assessments of convergence for the estimated mean vector $\mu(\mathsf{S}_3)$ and portions of the covariance matrix within $[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}$ and outside $[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}$ the nearest-neighbor index set $\mathcal{N}$. The data strongly suggests that the mean has the smallest error, the covariance within $\mathcal{N}$ the next smallest error, and the covariance outside $\mathcal{N}$ the largest error. Rows 1–3*: Relative errors. Rows* 4–6*: Relative palindromic symmetry errors. Column* 0*: Result for original* 100*ns simulation in training set. Columns* 1–9*: Results for nine additional* 50*ns, independent simulations.* $||\cdot||$ *denotes a standard Euclidean or Frobenius norm as determined by the context.*

	0	1	2	3	4	5	6	7	8	9
RE$_1$	0.03	0.04	0.03	0.03	0.07	0.03	0.04	0.07	0.03	0.07
RE$_2$	0.23	0.20	0.18	0.19	0.25	0.23	0.25	0.28	0.15	0.28
RE$_3$	0.55	0.62	0.59	0.64	0.63	0.60	0.56	0.58	0.65	0.60
PE$_1$	0.03	0.03	0.03	0.04	0.09	0.02	0.03	0.09	0.03	0.10
PE$_2$	0.17	0.20	0.11	0.11	0.32	0.08	0.14	0.31	0.19	0.28
PE$_3$	0.61	0.53	0.65	0.56	0.63	0.54	0.66	0.63	0.61	0.48

$$\mathrm{RE}_1 := \frac{||\mu(\mathsf{S}_3)-\mu_\rho(\mathsf{S}_3)||}{||\mu_\rho(\mathsf{S}_3)||} \qquad \mathrm{PE}_1 := \frac{||\mu(\mathsf{S}_3)-\mathsf{E}_n\mu(\mathsf{S}_3)||}{||\mu(\mathsf{S}_3)||}$$

$$\mathrm{RE}_2 := \frac{||[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}-[\mathsf{C}_\rho(\mathsf{S}_3)]_{\mathcal{N}}||}{||[\mathsf{C}_\rho(\mathsf{S}_3)]_{\mathcal{N}}||} \qquad \mathrm{PE}_2 := \frac{||[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}-\mathsf{E}_n[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}\mathsf{E}_n||}{||[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}}||}$$

$$\mathrm{RE}_3 := \frac{||[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}-[\mathsf{C}_\rho(\mathsf{S}_3)]_{\mathcal{N}^c}||}{||[\mathsf{C}_\rho(\mathsf{S}_3)]_{\mathcal{N}^c}||} \qquad \mathrm{PE}_3 := \frac{||[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}-\mathsf{E}_n[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}\mathsf{E}_n||}{||[\mathsf{C}(\mathsf{S}_3)]_{\mathcal{N}^c}||}$$

parameters for sequence-dependent banded Gaussian coarse-grain model of oligomers for which an MD simulation is already available to give estimates for the inputs $\mu(\mathsf{S})$ and $\mathsf{C}(\mathsf{S})$. In contrast, we now outline a model, recently introduced in [13], for predicting the underlying density $\rho(\mathsf{w};\mathsf{S})$ for an arbitrary sequence S. Assuming that the density is of the form in (1), the model provides a direct prediction of the free energy function $U(\mathsf{w};\mathsf{S})$, which by design will yield a Gaussian density with banded stiffness matrix as considered in sections 2.4 and 2.5. Later we will see how the oligomer-based models from those sections can be used to parameterize the model presented here. For more details see [13], and for descriptions of associated software and various experimental verifications see [27, 32].

4.1. Free energy.

The *cgDNA* model energy function for a given sequence S is

$$U_{\mathrm{cg}}(\mathsf{w};\mathcal{P},\mathsf{S}) = \frac{1}{2}\left[\mathsf{w}-\mu_{\mathrm{cg}}(\mathcal{P},\mathsf{S})\right]\cdot\mathsf{K}_{\mathrm{cg}}(\mathcal{P},\mathsf{S})\left[\mathsf{w}-\mu_{\mathrm{cg}}(\mathcal{P},\mathsf{S})\right]+e_{\mathrm{cg}}(\mathcal{P},\mathsf{S}). \tag{27}$$

Here $\mathcal{P}$ denotes a finite set of model parameters to be described below, $\mu_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ is the model mean vector, and $\mathsf{K}_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ is the model stiffness matrix which is symmetric, positive-definite, and banded, with a sparsity pattern corresponding to the nearest-neighbor index set $\mathcal{N}$ as considered in sections 2.4 and 2.5, and $e_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ is a model constant. From the energetic point of view, the mean vector $\mu_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ can be understood as a vector of configuration or shape parameters that defines the ground or minimum energy state of the sequence S, and consequently $e_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ represents the energy of this ground state compared to an unstressed state. Since our ultimate attention will be focused on the corresponding model density $\rho_{\mathrm{cg}}(\mathsf{w};\mathcal{P},\mathsf{S})$, we note that the constant $e_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ is inconsequential for our present purposes; it cancels out in the expression for the density. Nevertheless, the value of $e_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$ is of interest because it reflects the energy of *frustration* inherent to the *cgDNA* model.

We note that the best-fit oligomer models defined using maximum relative and absolute entropy in sections 2.4 and 2.5, with respective mean vectors and stiffness matrices $\mu_{\mathrm{rel}}(\mathsf{S})$, $\mathsf{K}_{\mathrm{rel}}(\mathsf{S})$ and $\mu_{\mathrm{abs}}(\mathsf{S})$, $\mathsf{K}_{\mathrm{abs}}(\mathsf{S})$, already provide a free energy of the form (27) where the constant term is implicitly zero. As previously described, these best-fit models can be found for any sequence S for which an ensemble of configuration coordinate vectors $\{\mathsf{w}^{[j]}(\mathsf{S})\}$, or more precisely, an estimated mean vector $\mu(\mathsf{S})$ and covariance matrix $\mathsf{C}(\mathsf{S})$, are available. However, if the sequence S is modified in any way, for example lengthened, shortened, or mutated by even one base pair, then it is not possible to determine what the modified best-fit oligomer model will be without generating an entirely new ensemble of configurations. This can be an intensive task if one wishes to explore a large number of modifications of a sequence, or equivalently, a large number of sequences. The *cgDNA* model finesses this limitation by assuming a relatively simple, and physically natural, decomposition of the free energy (27) into component parts such that the free energy of an arbitrary sequence S can be predicted from a finite, and comparatively small, parameter set $\mathcal{P}$.

In the *cgDNA* model, the free energy of an oligomer is based on a superposition of local energies that describe physically distinct interactions. Specifically, given an oligomer with sequence $\mathsf{S} = \mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$, there are two types of local energies that are considered. The first type is associated with each base pair $(\mathtt{X},\overline{\mathtt{X}})_a$, $a = 1,\ldots,n$, along the oligomer; it is referred to as a base-pair or mononucleotide energy. The second type is associated with each pair of base pairs $(\mathtt{X},\overline{\mathtt{X}})_a$ and $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$, $a = 1,\ldots,n-1$, along the oligomer; it is referred to as a junction or dinucleotide energy. The local energy associated with each base pair, with label $\mathtt{X}_a$ on the reference strand, is

$$U^{\mathtt{X}_a}(y^a) = \frac{1}{2}\left[y^a - \mu^{\mathtt{X}_a}\right]\cdot \mathsf{K}^{\mathtt{X}_a}\left[y^a - \mu^{\mathtt{X}_a}\right], \tag{28}$$

where $y^a \in \mathbb{R}^6$ is the vector of intra-base-pair coordinates that fully describes the relative translation and rotation between the two bases in the pair (see section 3.1), $\mu^{\mathtt{X}_a} \in \mathbb{R}^6$ is a vector of local shape parameters, and $\mathsf{K}^{\mathtt{X}_a} \in \mathbb{R}^{6\times 6}$ is a symmetric matrix of local stiffness parameters. The energy $U^{\mathtt{X}_a}(y^a)$ is to be interpreted as a model for the intra-base-pair interactions between the two bases of $(\mathtt{X},\overline{\mathtt{X}})_a$, as illustrated in the left panel of Figure 7.

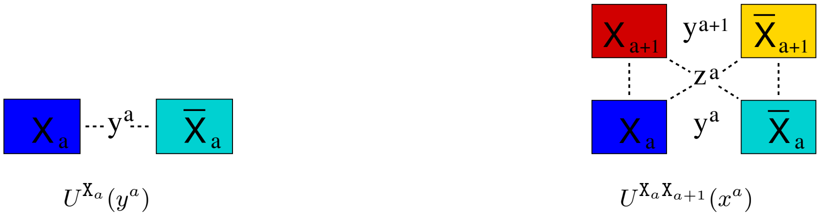

FIG. 7. *Schematic of local energies in the* cgDNA *model: The base-pair energy is a model for all the base-base interactions within a base pair (left), the junction energy is a model for all the base-base interactions across a junction (right).*

Similarly, the local energy associated with each junction or pair of base pairs, with label $\mathtt{X}_a\mathtt{X}_{a+1}$ on the reference strand, is

$$U^{\mathtt{X}_a\mathtt{X}_{a+1}}(x^a) = \frac{1}{2}\left[x^a - \mu^{\mathtt{X}_a\mathtt{X}_{a+1}}\right]\cdot \mathsf{K}^{\mathtt{X}_a\mathtt{X}_{a+1}}\left[x^a - \mu^{\mathtt{X}_a\mathtt{X}_{a+1}}\right], \tag{29}$$

where $x^a := (y^a, z^a, y^{a+1}) \in \mathbb{R}^{18}$ is the vector of coordinates that fully describes the relative translations and rotations between all four bases in the pair of base pairs, $\mu^{\mathtt{X}_a\mathtt{X}_{a+1}} \in \mathbb{R}^{18}$ is a vector of local shape parameters, and $\mathsf{K}^{\mathtt{X}_a\mathtt{X}_{a+1}} \in \mathbb{R}^{18\times 18}$ is a symmetric matrix of local stiffness parameters analogous to those mentioned before. The energy $U^{\mathtt{X}_a\mathtt{X}_{a+1}}(x^a)$ is to be interpreted as a model for all the inter-base-pair interactions involving a base of $(\mathtt{X},\overline{\mathtt{X}})_a$ and a base of $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$, in other words any nearest-neighbor, base-base interaction across the junction between the base pairs $(\mathtt{X},\overline{\mathtt{X}})_a$ and $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$, as illustrated in the right panel of Figure 7. The fact that the coordinate vector y^a appears in the expressions for $U^{\mathtt{X}_a}(y^a)$, $U^{\mathtt{X}_a\mathtt{X}_{a+1}}(x^a)$, and $U^{\mathtt{X}_{a-1}\mathtt{X}_a}(x^{a-1})$ is physically pertinent; it is associated with the phenomenon of frustration.

Notice that the number of different local energy parameters is determined by the number of different sequence composition labels. Specifically, there are four different types of base-pair energy parameters $\mu^{\mathtt{X}_a} \in \mathbb{R}^6$ and $\mathsf{K}^{\mathtt{X}_a} \in \mathbb{R}^{6\times 6}$ corresponding to the labels $\mathtt{X}_a \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$, and there are sixteen different junction energy parameters $\mu^{\mathtt{X}_a\mathtt{X}_{a+1}} \in \mathbb{R}^{18}$ and $\mathsf{K}^{\mathtt{X}_a\mathtt{X}_{a+1}} \in \mathbb{R}^{18\times 18}$ corresponding to the labels $\mathtt{X}_a, \mathtt{X}_{a+1} \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$. Hence a complete set of local energy parameters is a collection of vectors and matrices of the form

$$\mathcal{P} = \left\{\mu^{\alpha}, \mathsf{K}^{\alpha}, \mu^{\alpha\beta}, \mathsf{K}^{\alpha\beta}\right\}_{\alpha,\beta\in\{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}}. \tag{30}$$

Throughout this section we shall make various changes to this collection and exploit Watson–Crick symmetry to reduce its size. For notational simplicity, we shall continue to use the notation $\mathcal{P}$ after each change.

The overall *cgDNA* model can now be described. Specifically, given a parameter set $\mathcal{P}$, the total free energy for an oligomer with arbitrary sequence $\mathsf{S} = \mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ is defined by superposing all the local energies, namely

$$U_{\mathrm{cg}}(\mathsf{w};\mathcal{P},\mathsf{S}) = \sum_{a=1}^{n} U^{\mathtt{X}_a}(y^a) + \sum_{a=1}^{n-1} U^{\mathtt{X}_a\mathtt{X}_{a+1}}(x^a). \tag{31}$$

Notice that the total free energy is based only on nearest-neighbor interactions between neighboring bases both across and along the two backbone strands of the oligomer. Moreover, notice that this energy makes two additional and logically independent assumptions of locality in sequence, and independence of location along the oligomer, for example proximity to an end. In principle, larger parameter sets could be adopted with, for example, tetranucleotide sequence dependence of the junction energy parameters, but the examples presented in [13, 32] suggest that the level of generality outlined above already provides a rather good approximation. One could also simplify to a sequence-independent model in which the parameters μ^{α}, K^{α}, $\mu^{\alpha\beta}$, and $\mathsf{K}^{\alpha\beta}$ are independent of the labels $\alpha,\beta \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$.

4.2. Bandedness and local sequence dependence of $\mathsf{K}_{\mathrm{cg}}(\mathcal{P},\mathsf{S})$. The expressions in (27) and (31) are each ultimately a quadratic expression in the oligomer coordinate vector w so that the coefficients in one expression can be related to the coefficients in the other. To make these relations explicit, we proceed as follows. Let $\mathsf{y} = (y^1,\dots,y^n) \in \mathbb{R}^{6n}$ denote the collection of all base-pair coordinate vectors y^a, and let $\mathsf{x} = (x^1,\dots,x^{n-1}) \in \mathbb{R}^{18(n-1)}$ denote the collection of all junction coordinate vectors x^a. Then the vectors y and x are related to the vector w according to $\mathsf{y} = \mathsf{P}_{\mathsf{y}}\mathsf{w}$ and $\mathsf{x} = \mathsf{P}_{\mathsf{x}}\mathsf{w}$, where $\mathsf{P}_{\mathsf{y}} \in \mathbb{R}^{6n\times(12n-6)}$ and $\mathsf{P}_{\mathsf{x}} \in \mathbb{R}^{18(n-1)\times(12n-6)}$ are constant matrices given by

$$\mathsf{P_y} = \begin{pmatrix} I & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & 0 & 0 & I & & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & I \end{pmatrix}, \qquad \mathsf{P_x} = \begin{pmatrix} I & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & I & 0 & 0 & 0 & & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & 0 & I & 0 & & 0 \\ 0 & 0 & 0 & 0 & I & & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & I \end{pmatrix} \tag{32}$$

and where $0 \in \mathbb{R}^{6\times 6}$ and $I \in \mathbb{R}^{6\times 6}$ denote the zero and identity matrix blocks.

With the above notation, we can now write the free energy in (31) in a more convenient matrix form as the sum of two shifted quadratic forms of different dimensions, namely

$$\begin{aligned} U_{\mathrm{cg}}(\mathsf{w}; \mathcal{P}, \mathsf{S}) = \frac{1}{2}(\mathsf{P_y w} - \mu_{\mathsf{y}}) \cdot \mathsf{K_y}(\mathsf{P_y w} - \mu_{\mathsf{y}}) \\ + \frac{1}{2}(\mathsf{P_x w} - \mu_{\mathsf{x}}) \cdot \mathsf{K_x}(\mathsf{P_x w} - \mu_{\mathsf{x}}). \end{aligned} \tag{33}$$

Here $\mu_{\mathsf{y}} = (\mu^{\mathtt{X}_1}, \dots, \mu^{\mathtt{X}_n})$ and $\mu_{\mathsf{x}} = (\mu^{\mathtt{X}_1\mathtt{X}_2}, \dots, \mu^{\mathtt{X}_{n-1}\mathtt{X}_n})$ are vectors containing all the base-pair and junction shape parameters for the given sequence, and similarly $\mathsf{K_y} = \mathrm{diag}(\mathsf{K}^{\mathtt{X}_1}, \dots, \mathsf{K}^{\mathtt{X}_n})$ and $\mathsf{K_x} = \mathrm{diag}(\mathsf{K}^{\mathtt{X}_1\mathtt{X}_2}, \dots, \mathsf{K}^{\mathtt{X}_{n-1}\mathtt{X}_n})$ are block-diagonal matrices containing all the base-pair and junction stiffness matrices as their blocks. Notice that $\mu_{\mathsf{y}} \in \mathbb{R}^{6n}$ and $\mathsf{K_y} \in \mathbb{R}^{6n\times 6n}$, whereas $\mu_{\mathsf{x}} \in \mathbb{R}^{18(n-1)}$ and $\mathsf{K_x} \in \mathbb{R}^{18(n-1)\times 18(n-1)}$.

By separately comparing the coefficients of the quadratic, linear and constant terms in w, we find that the coefficients in (27) can be expressed in terms of the coefficients in (33). Specifically, from the quadratic and linear terms we get

$$\begin{aligned} \mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S}) &= \mathsf{P_y}^T \mathsf{K_y P_y} + \mathsf{P_x}^T \mathsf{K_x P_x}, \\ \mu_{\mathrm{cg}}(\mathcal{P}, \mathsf{S}) &= \mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S})^{-1} \sigma_{\mathrm{cg}}(\mathcal{P}, \mathsf{S}), \end{aligned} \tag{34}$$

where $\sigma_{\mathrm{cg}}(\mathcal{P}, \mathsf{S})$ is an auxiliary vector given by

$$\begin{aligned} \sigma_{\mathrm{cg}}(\mathcal{P}, \mathsf{S}) := \mathsf{P_y}^T \sigma_{\mathsf{y}} + \mathsf{P_x}^T \sigma_{\mathsf{x}}, \\ \sigma_{\mathsf{y}} := \mathsf{K_y} \mu_{\mathsf{y}}, \qquad \sigma_{\mathsf{x}} := \mathsf{K_x} \mu_{\mathsf{x}}. \end{aligned} \tag{35}$$

Moreover, from the constant terms we get

$$\begin{aligned} e_{\mathrm{cg}}(\mathcal{P}, \mathsf{S}) = \frac{1}{2}(\mathsf{P_y} \mu_{\mathrm{cg}} - \mu_{\mathsf{y}}) \cdot \mathsf{K_y}(\mathsf{P_y} \mu_{\mathrm{cg}} - \mu_{\mathsf{y}}) \\ + \frac{1}{2}(\mathsf{P_x} \mu_{\mathrm{cg}} - \mu_{\mathsf{x}}) \cdot \mathsf{K_x}(\mathsf{P_x} \mu_{\mathrm{cg}} - \mu_{\mathsf{x}}). \end{aligned} \tag{36}$$

We remark that σ_{y} and σ_{x} are parameter combinations that arise naturally in the algebraic manipulations. In our later developments, it will be convenient to work with weighted shape parameters defined as

$$\sigma^{\alpha} := \mathsf{K}^{\alpha} \mu^{\alpha} \in \mathbb{R}^6, \quad \sigma^{\alpha\beta} := \mathsf{K}^{\alpha\beta} \mu^{\alpha\beta} \in \mathbb{R}^{18}, \tag{37}$$

and to replace the parameter set in (30) by

$$\mathcal{P} = \left\{ \sigma^{\alpha}, \mathsf{K}^{\alpha}, \sigma^{\alpha\beta}, \mathsf{K}^{\alpha\beta} \right\}_{\alpha,\beta \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}}. \tag{38}$$

Given the above parameter set $\mathcal{P}$, the predicted stiffness matrix $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$ and shape vector $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$ for any sequence S is then given by $(34)_{1,2}$ and $(35)_1$, where σ_{y} and σ_{x} are vectors of weighted shape parameters given by $\sigma_{\mathsf{y}} = (\sigma^{\mathtt{X}_1},\dots,\sigma^{\mathtt{X}_n})$ and $\sigma_{\mathsf{x}} = (\sigma^{\mathtt{X}_1\mathtt{X}_2},\dots,\sigma^{\mathtt{X}_{n-1}\mathtt{X}_n})$. Notice that the unweighted shape parameters in μ_{y} and μ_{x} are needed explicitly only when either the oligomer frustration energy $e_{\rm cg}(\mathcal{P},\mathsf{S})$ or the local energy terms in (31) need to be evaluated explicitly.

The simple and local dependence of each subblock of the model stiffness matrix $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$ on the oligomer sequence $\mathsf{S} = \mathtt{X}_1\cdots\mathtt{X}_n$ is illustrated in Figure 8. On the left-hand side, each single number in a block denotes a dependence on the base-pair or mononucleotide composition $\mathtt{X}_a$, while each pair of numbers in a block denotes a dependence on the junction or dinucleotide composition $\mathtt{X}_a\mathtt{X}_{a+1}$. On the right-hand side, the double and triple overlapping blocks denote sums; notice that the shaded blocks with triple overlaps exhibit an effective dependence on the trinucleotide composition $\mathtt{X}_{a-1}\mathtt{X}_a\mathtt{X}_{a+1}$ corresponding to the union of the two adjacent dinucleotides steps and their common mononucleotide. By design, notice that $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$ has a banded sparsity structure corresponding to the nearest-neighbor index set $\mathcal{N}$ as considered in sections 2.4 and 2.5. Moreover, it is straightforward to show that the auxiliary vector $\sigma_{\rm cg}(\mathcal{P},\mathsf{S})$ will have an analogous dependence on sequence.

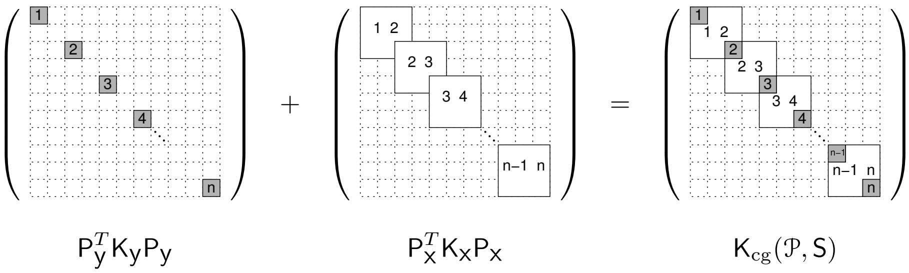
1
2
3
4
n
+
1 2
2 3
3 4
n–1 n
=
1
1 2
2
2 3
3
3 4
4
n–1
n–1 n
n
$\mathsf{P}_{\mathsf{y}}^T\mathsf{K}_{\mathsf{y}}\mathsf{P}_{\mathsf{y}}$
$\mathsf{P}_{\mathsf{x}}^T\mathsf{K}_{\mathsf{x}}\mathsf{P}_{\mathsf{x}}$
$\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$

FIG. 8. *Illustration of the sequence dependence and the nearest-neighbor sparsity structure of the model stiffness matrix* $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$.

4.3. Nonlocal sequence dependence of $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})^{-1}$ and $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$.

In contrast to the *cgDNA* model stiffness $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$, the model covariance matrix $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})^{-1}$ is dense and its entries have a nonlocal dependence on sequence. These properties follow from two simple facts from linear algebra: first, the inverse of a banded, non-block-diagonal matrix as considered here is dense; and second, the blocks in the dense inverse have a nonlocal dependence on the blocks of the banded matrix. Consequently, the model covariance has the stated properties. It is interesting to note that, if the covariance were given a priori, then the process of inversion would yield the banded and locally sequence dependent stiffness, so that the dense structure and nonlocal sequence dependence in the covariance is rather special. More remarkably, because $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$ is banded according to the index set $\mathcal{N}$, we note that it could be constructed from knowledge of only the subset of the covariance inside $\mathcal{N}$, namely $[\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})^{-1}]_{\mathcal{N}}$, via the local inversion algorithm described in section 2.5.

The entries of the model shape vector $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$ also have a nonlocal dependence on sequence. Indeed, from $(34)_2$ we see that the shape vector is given by the product of the covariance matrix $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})^{-1}$ and the auxiliary vector $\sigma_{\rm cg}(\mathcal{P},\mathsf{S})$. This relation contains two sources of nonlocality: first, due to the denseness of $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})^{-1}$, each entry of $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$ is a sum over all the entries of $\sigma_{\rm cg}(\mathcal{P},\mathsf{S})$ which collectively depend on the entire sequence; and second, the entries of $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})^{-1}$ have themselves a nonlocal dependence on sequence. We remark that such nonlocal dependence has been observed in various MD simulations (see, for example, [29]), where the strong dependence of the average dinucleotide configuration on the flanking tetranucleotide sequence context has been documented. This observation has sometimes been interpreted as implying that an accurate coarse-grain model must have a parameter set that is at least tetranucleotide dependent (which is a very large number of parameters). However, such a parameter set is not necessary. Indeed, the *cgDNA* model, which employs a parameter set $\mathcal{P}$ that is only dinucleotide dependent, can predict shape vectors $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$ with a nonlocal dependence on the sequence S that closely agree with averages observed in MD simulation [13, 32].

We remark that the expression for the frustration energy $e_{\rm cg}(\mathcal{P},\mathsf{S})$ in (36) follows directly from setting $\mathsf{w} = \mu_{\rm cg}(\mathcal{P},\mathsf{S})$ in (27) and (33); it can also be obtained from a completion-of-squares argument. Assuming the model stiffness matrix $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$ is positive-definite, a condition that will be discussed later, we see that $e_{\rm cg}(\mathcal{P},\mathsf{S})$ is the minimal accessible free energy achieved by the ground state configuration $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$ of the oligomer. The expression for $e_{\rm cg}(\mathcal{P},\mathsf{S})$ is a reflection of the fact that, in general, each base cannot simultaneously minimize the local base-pair and the two local junction energies in which it is involved. Instead, each base must find a compromise, and the ground state configuration of the oligomer is *frustrated*, which provides the physical explanation for the nonlocal sequence dependence of $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$.

4.4. Watson–Crick symmetries of $\mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S})$ and $\mu_{\rm cg}(\mathcal{P},\mathsf{S})$. When applied to the *cgDNA* model density $\rho_{\rm cg}(\mathsf{w};\mathcal{P},\mathsf{S})$, defined by the free energy $U_{\rm cg}(\mathsf{w};\mathcal{P},\mathsf{S})$, we find that the Watson–Crick symmetry relations in (21) are equivalent to

$$\mu_{\rm cg}(\mathcal{P},\mathsf{S}) = \mathsf{E}_n\mu_{\rm cg}(\mathcal{P},\overline{\mathsf{S}}), \qquad \mathsf{K}_{\rm cg}(\mathcal{P},\mathsf{S}) = \mathsf{E}_n\mathsf{K}_{\rm cg}(\mathcal{P},\overline{\mathsf{S}})\mathsf{E}_n, \tag{39}$$

where n is the length of S, and E_n is defined in (20). These relations must hold for any sequence S and its complement $\overline{\mathsf{S}}$ read from the two possible choices of reference strand for an oligomer. Moreover, in the case of a palindromic sequence with $\overline{\mathsf{S}} = \mathsf{S}$, we note that the above relations imply constraints that are equivalent to (22).

For the *cgDNA* model as described, there is no a priori reason to expect that the relations in (39) will hold for an arbitrary sequence S and its complement $\overline{\mathsf{S}}$. Hence further conditions on the model are necessary to guarantee consistency with Watson–Crick symmetry. One set of simple, sufficient conditions can be deduced from the model relations in (34) and (35), and the parameter set in (38). Specifically, in order for (39) to hold for arbitrary sequences, it is sufficient for the parameters in $\mathcal{P}$ to satisfy analogous local versions of the symmetry relations, namely

$$\begin{aligned} &\sigma^{\alpha} = \mathsf{E}_1\sigma^{\overline{\alpha}}, && \mathsf{K}^{\alpha} = \mathsf{E}_1\mathsf{K}^{\overline{\alpha}}\mathsf{E}_1, \\ &\sigma^{\alpha\beta} = \mathsf{E}_2\sigma^{\overline{\beta}\overline{\alpha}}, && \mathsf{K}^{\alpha\beta} = \mathsf{E}_2\mathsf{K}^{\overline{\beta}\overline{\alpha}}\mathsf{E}_2 \end{aligned} \qquad \forall\alpha,\beta \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}, \tag{40}$$

where $\overline{\alpha}$ denotes the complementary base to α in the sense that $\overline{\mathtt{T}} = \mathtt{A}$ and so on. Since $\overline{\overline{\alpha}} = \alpha$, $\mathsf{E}_1^{-1} = \mathsf{E}_1$, and $\mathsf{E}_2^{-1} = \mathsf{E}_2$, we note that not all of the above conditions are independent.

The conditions in (40) can be used to reduce the size of the parameter set $\mathcal{P}$. For instance, there are four parameters σ^{α}, and the conditions $\sigma^{\alpha} = \mathsf{E}_1\sigma^{\overline{\alpha}}$ provide two independent equations; hence two of the parameters can be taken as independent, for example $\sigma^{\mathtt{A}}$ and $\sigma^{\mathtt{G}}$, and the other two are then determined by Watson–Crick symmetry, namely $\sigma^{\mathtt{T}} = \mathsf{E}_1\sigma^{\mathtt{A}}$ and $\sigma^{\mathtt{C}} = \mathsf{E}_1\sigma^{\mathtt{G}}$. Similarly, there are sixteen parameters $\sigma^{\alpha\beta}$, and the conditions $\sigma^{\alpha\beta} = \mathsf{E}_2\sigma^{\overline{\beta}\overline{\alpha}}$ provide six independent equations when

$\alpha\beta \neq \overline{\beta}\overline{\alpha}$, and four independent equations when $\alpha\beta = \overline{\beta}\overline{\alpha}$. For the twelve parameters $\sigma^{\alpha\beta}$ with $\alpha\beta \neq \overline{\beta}\overline{\alpha}$, we can take six as independent, and the other six are determined by Watson–Crick symmetry. And for the four parameters $\sigma^{\alpha\beta}$ with $\alpha\beta = \overline{\beta}\overline{\alpha}$, we have four independent equations which restrict their values. Similar considerations apply to K^{α} and $\mathsf{K}^{\alpha\beta}$. Proceeding in this way, we arrive at a reduced parameter set of the form

$$\mathcal{P} = \Big\{ \{\sigma^{\alpha}, \mathsf{K}^{\alpha}, \sigma^{\gamma\beta}, \mathsf{K}^{\gamma\beta}\}_{\alpha\in\mathtt{M},\gamma\beta\in\mathtt{D}} \quad \Big| \quad \begin{aligned} &\sigma^{\gamma\beta} = \mathsf{E}_2\sigma^{\gamma\beta}, \\ &\mathsf{K}^{\gamma\beta} = \mathsf{E}_2\mathsf{K}^{\gamma\beta}\mathsf{E}_2 \quad \forall\gamma\beta \in \mathtt{D}' \Big\}. \end{aligned} \tag{41}$$

Here $\mathtt{M}$ denotes any set of two independent mononucleotides, for example $\mathtt{M} = \{\mathtt{A}, \mathtt{G}\}$, and $\mathtt{D}$ denotes any set of ten independent dinucleotides, for example $\mathtt{D} = \{\mathtt{AT}, \mathtt{GC}, \mathtt{TA}, \mathtt{CG}, \mathtt{GT}, \mathtt{TC}, \mathtt{CA}, \mathtt{TT}, \mathtt{CC}, \mathtt{CT}\}$, and $\mathtt{D}'$ denotes the set of four self-symmetric dinucleotides, which must necessarily be contained in $\mathtt{D}$, namely $\mathtt{D}' = \{\mathtt{AT}, \mathtt{GC}, \mathtt{TA}, \mathtt{CG}\}$.

4.5. Positivity of $\mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S})$. A natural requirement for the *cgDNA* model is that the predicted stiffness matrix $\mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S})$ be positive-definite for arbitrary sequences S above some minimal length n_0; that is,

$$\mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S}) > 0 \quad \forall\mathsf{S} \text{ such that } |\mathsf{S}| \geq n_0. \tag{42}$$

This requirement is natural for any model that employs a quadratic approximation of the free energy and guarantees that the corresponding density $\rho_{\mathrm{cg}}(\mathsf{w}; \mathcal{P}, \mathsf{S})$ on the configuration space of the oligomer is well defined, with a positive-definite covariance matrix, as has been observed for all the sequences in our training set. The requirement is also physical in the sense that short double-stranded sequences, such as an individual base pair (monomer), or pair (dimer), or triplet (trimer) of base pairs, are not expected to be stable in a solvent environment, whereas longer sequences are expected to be stable in the sense of exhibiting approximately stationary statistics.

For the *cgDNA* model as described, there is no reason to expect that the positivity condition in (42) will hold for arbitrary sequences. Hence further conditions on the model, specifically the parameter set $\mathcal{P}$, are necessary to guarantee positivity. In view of the model relations in $(34)_1$, we note that a rather strong set of conditions can be readily identified: namely, it is sufficient that each of the parameter matrices K^{α} and $\mathsf{K}^{\alpha\beta}$ be positive-definite. Alternatively, as described in [13], a weaker set of conditions is also sufficient: namely, each of the matrices K^{α} and $\mathsf{K}^{\alpha\beta}$ need only be positive-semidefinite, provided that a small number of overlapping sums of these matrices are positive-definite. The parameter set *cgDNAparamset*1 published as part of [13, 32], which was obtained via a numerical fit to our relative entropy training data, was seen to satisfy the weaker set of conditions, but not the stronger within numerical error, which suggests that the semidefinite conditions might be natural for fitting our data. However, the numerical procedure that delivered *cgDNAparamset*1 did not allow any of the matrices K^{α} and $\mathsf{K}^{\alpha\beta}$ to cross the semidefinite boundary, which raises the possibility that a better numerical fit might be possible with even weaker conditions on these matrices. Indeed, there is no physical reason for these matrices to be even semidefinite.

Motivated by the above observations, we now suppose that each of the parameter matrices K^{α} and $\mathsf{K}^{\alpha\beta}$ is merely symmetric, but otherwise general, and seek sufficient conditions for (42) to hold. Such sufficient conditions can be constructed as follows.

Consider auxiliary matrices $\mathsf{K}^{\alpha\beta}_{5'}$, $\mathsf{K}^{\alpha\beta}_{3'}$, and $\mathsf{K}^{\alpha\beta}_{\frac{1}{2}}$ defined for all $\alpha, \beta \in \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$ by

$$\begin{gathered}\mathsf{K}^{\alpha\beta}_{5'} := \mathsf{K}^{\alpha\beta} + \begin{pmatrix} \mathsf{K}^{\alpha} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\mathsf{K}^{\beta} \end{pmatrix}, \qquad \mathsf{K}^{\alpha\beta}_{3'} := \mathsf{K}^{\alpha\beta} + \begin{pmatrix} \frac{1}{2}\mathsf{K}^{\alpha} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathsf{K}^{\beta} \end{pmatrix}, \\ \mathsf{K}^{\alpha\beta}_{\frac{1}{2}} := \mathsf{K}^{\alpha\beta} + \begin{pmatrix} \frac{1}{2}\mathsf{K}^{\alpha} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\mathsf{K}^{\beta} \end{pmatrix},\end{gathered} \tag{43}$$

where $0 \in \mathbb{R}^{6\times 6}$ is the zero matrix. We notice that, if the parameter matrices K^{α} and $\mathsf{K}^{\alpha\beta}$ satisfy the Watson–Crick symmetry relations, then the above auxiliary matrices will satisfy analogous relations of the form

$$\mathsf{K}^{\alpha\beta}_{3'} = \mathsf{E}_2\mathsf{K}^{\overline{\beta}\overline{\alpha}}_{5'}\mathsf{E}_2, \quad \mathsf{K}^{\alpha\beta}_{\frac{1}{2}} = \mathsf{E}_2\mathsf{K}^{\overline{\beta}\overline{\alpha}}_{\frac{1}{2}}\mathsf{E}_2 \quad \forall \alpha, \beta \in \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}. \tag{44}$$

As a consequence, the sixteen matrices $\mathsf{K}^{\alpha\beta}_{5'}$ for all $\alpha, \beta \in \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$, and the ten matrices $\mathsf{K}^{\alpha\beta}_{\frac{1}{2}}$ for all $\alpha\beta \in \mathtt{D}$, may be regarded as independent; the remaining auxiliary matrices are then determined by Watson–Crick symmetry.

A simple set of sufficient conditions for the positivity of the *cgDNA* model, which allows the parameter matrices K^{α} and $\mathsf{K}^{\alpha\beta}$ to be indefinite, can now be stated. Specifically, in order for (42) to hold for arbitrary sequences S with lengths $|\mathsf{S}| \geq 3$, it is sufficient that

$$\begin{aligned} &\mathsf{K}^{\alpha\beta}_{5'} > 0 \qquad && \forall \alpha, \beta \in \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}, \\ &\mathsf{K}^{\alpha\beta}_{\frac{1}{2}} > 0 \qquad && \forall \alpha\beta \in \mathtt{D}. \end{aligned} \tag{45}$$

The proof of sufficiency is straightforward. Indeed, since the matrix E_2 is invertible, the conditions in (45) together with the symmetry relations in (44) imply that all the auxiliary matrices $\mathsf{K}^{\alpha\beta}_{5'}$, $\mathsf{K}^{\alpha\beta}_{3'}$, and $\mathsf{K}^{\alpha\beta}_{\frac{1}{2}}$ are positive-definite for all $\alpha, \beta \in \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$. The result then follows from the fact that, for an arbitrary sequence S with $|\mathsf{S}| \geq 3$, the model stiffness matrix $\mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S})$ can be written as an overlapping sum of these auxiliary matrices, specifically a matrix of type $\mathsf{K}^{\alpha\beta}_{5'}$ at the leading end, a matrix of type $\mathsf{K}^{\alpha\beta}_{3'}$ at the trailing end, and matrices of type $\mathsf{K}^{\alpha\beta}_{\frac{1}{2}}$ along the interior; see Figure 8.

Below we report an updated parameter set, *cgDNAparamset*2, with enhanced properties, which is a locally unique optimizer for our fitting procedure, contains indefinite stiffness matrices, and satisfies the sufficient conditions in (45). As we will see, the predictive capabilities of the *cgDNA* model with *cgDNAparamset*2 are noticeably improved as compared with *cgDNAparamset*1. We note that the above considerations raise the possibility of using the matrices $\mathsf{K}^{\alpha\beta}_{5'}$, $\mathsf{K}^{\alpha\beta}_{3'}$, and $\mathsf{K}^{\alpha\beta}_{\frac{1}{2}}$ as a basis for our space of stiffness parameters, but the choice is not evident. For example, because of the structure of the overlapping sums in the model stiffness matrix $\mathsf{K}_{\mathrm{cg}}(\mathcal{P}, \mathsf{S})$, it is not possible to estimate a unique set of matrices of type $\mathsf{K}^{\alpha\beta}_{5'}$, $\mathsf{K}^{\alpha\beta}_{3'}$, and $\mathsf{K}^{\alpha\beta}_{\frac{1}{2}}$ from observing only the oligomer covariance matrix.

5. *cgDNA* parameter set estimation.

Here we use our training data set to estimate the *cgDNA* model parameter set $\mathcal{P}$ in (41). Specifically, for each sequence in our ensemble $\{\mathsf{S}_\nu\}_{\nu=1}^{N}$, we have an observed density $\rho_{\mathrm{o}}(\mathsf{w}; \mathsf{S}_\nu)$ that is provided by a maximum relative or absolute entropy fit to the training data. We now seek a parameter set $\mathcal{P}$ for which the predicted density $\rho_{\mathrm{cg}}(\mathsf{w}; \mathcal{P}, \mathsf{S}_\nu)$ will fit, as closely as possible, the observed density for each sequence S_ν. Presumably, if the training set is sufficiently rich, such a best-fit parameter set should not only provide a good description of the training set sequences, but also any other sequence of arbitrary length, which is the goal of our predictive model.

5.1. Parameter space. In view of (41), a parameter set is a collection of vectors and matrices that corresponds to an element of the Euclidean space

$$\mathcal{H} = [\mathbb{R}^6]^2 \times [\mathbb{S}^6]^2 \times [\mathbb{R}^{18}]^{10} \times [\mathbb{S}^{18}]^{10}, \tag{46}$$

where $\mathbb{S}^k$ is the set of symmetric $k \times k$ matrices. A generic element of $\mathcal{H}$ is denoted by $h = \{\sigma^\alpha, \mathsf{K}^\alpha, \sigma^{\gamma\beta}, \mathsf{K}^{\gamma\beta}\}_{\alpha\in\mathtt{M},\gamma\beta\in\mathtt{D}}$. Due to the Watson–Crick symmetry relations associated with $\gamma\beta \in \mathtt{D}'$, a parameter set $\mathcal{P}$ is actually an element of a linear subspace $\mathcal{H}_{\text{self}}$ defined by

$$\mathcal{H}_{\text{self}} = \{h \in \mathcal{H} \mid \quad \sigma^{\gamma\beta} = \mathsf{E}_2\sigma^{\gamma\beta}, \quad \mathsf{K}^{\gamma\beta} = \mathsf{E}_2\mathsf{K}^{\gamma\beta}\mathsf{E}_2 \quad \forall\gamma\beta \in \mathtt{D}'\}. \tag{47}$$

The spaces $\mathcal{H}$ and $\mathcal{H}_{\text{self}}$ are large, with dimensions $\dim(\mathcal{H}) = 1944$ and $\dim(\mathcal{H}_{\text{self}}) = 1592$, which makes the parameter fitting problem rather challenging. Whereas a simple canonical basis is readily available for $\mathcal{H}$, the construction of a basis for $\mathcal{H}_{\text{self}}$ is somewhat tedious.

Other subsets of $\mathcal{H}$ naturally arise in our developments. For example, given the training ensemble $\{\mathsf{S}_\nu\}_{\nu=1}^N$, it is natural to consider the subset for which the predicted stiffness matrices over the ensemble are all positive-definite, namely

$$\mathcal{H}_{\text{train}} = \{h \in \mathcal{H} \mid \quad \mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu) > 0 \quad \forall\nu = 1, \dots, N\}, \tag{48}$$

where $\mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu)$ is defined in $(34)_1$. Thus $\mathcal{H}_{\text{self}} \cap \mathcal{H}_{\text{train}}$ is the set of all parameter sets for which the predicted densities are well defined for each sequence in the training ensemble; this is our largest possible parameter space. For purposes of discussion, the following subset corresponding to positive-semidefinite stiffness parameter matrices is also of interest:

$$\mathcal{H}_{\text{psd}} = \{h \in \mathcal{H} \mid \quad \mathsf{K}^\alpha \geq 0, \quad \mathsf{K}^{\gamma\beta} \geq 0 \quad \forall\alpha \in \mathtt{M}, \quad \forall\gamma\beta \in \mathtt{D}\}. \tag{49}$$

5.2. Objective function. Given an observed density $\rho_{\text{o}}(\mathsf{w}; \mathsf{S}_\nu)$ for each sequence in the ensemble $\{\mathsf{S}_\nu\}_{\nu=1}^N$, we consider an objective function $\mathcal{F} : \mathcal{H}_{\text{self}} \cap \mathcal{H}_{\text{train}} \to \mathbb{R}$ defined as a weighted sum of relative entropies over the ensemble, namely

$$\mathcal{F}(h) = \sum_{\nu=1}^{N} \omega_\nu D_{\text{rel}}(\rho_{\text{cg}}(\mathsf{w}; h, \mathsf{S}_\nu), \rho_{\text{o}}(\mathsf{w}; \mathsf{S}_\nu)), \tag{50}$$

where $\omega_\nu \geq 0$ are specified weights. Notice that $\mathcal{F}(h) \leq 0$ and, when all weights are positive, $\mathcal{F}(h) = 0$ if and only if $\rho_{\text{cg}}(\mathsf{w}; h, \mathsf{S}_\nu) \equiv \rho_{\text{o}}(\mathsf{w}; \mathsf{S}_\nu)$ for all $\nu = 1, \dots, N$. Hence in seeking a best-fit parameter set it is natural to maximize this function. The weights ω_ν can be chosen to reflect the relative importance or confidence in the observed densities $\rho_{\text{o}}(\mathsf{w}; \mathsf{S}_\nu)$, or more appropriately, the data from which these densities were derived. In the case when all densities are Gaussian, as considered here, we can use the relation in (8) to obtain the explicit expression

$$\begin{aligned} \mathcal{F}(h) = \sum_{\nu=1}^{N} \frac{\omega_\nu}{2} \Big[&\ln\Big(\det \mathsf{K}_{\text{o}}(\mathsf{S}_\nu) / \det \mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu)\Big) \\ &- \mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu)^{-1} : \mathsf{K}_{\text{o}}(\mathsf{S}_\nu) \ + \ I : I \\ &- [\mu_{\text{cg}}(h, \mathsf{S}_\nu) - \mu_{\text{o}}(\mathsf{S}_\nu)] \cdot \mathsf{K}_{\text{o}}(\mathsf{S}_\nu)[\mu_{\text{cg}}(h, \mathsf{S}_\nu) - \mu_{\text{o}}(\mathsf{S}_\nu)]\Big], \end{aligned} \tag{51}$$

where $\mu_{\text{cg}}(h, \mathsf{S}_\nu) = \mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu)^{-1}\sigma_{\text{cg}}(h, \mathsf{S}_\nu)$ in accordance with $(34)_2$. For each sequence S_ν, the quantities $\mathsf{K}_{\text{o}}(\mathsf{S}_\nu)$ and $\mu_{\text{o}}(\mathsf{S}_\nu)$ are given, and the quantities $\mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu)$ and $\sigma_{\text{cg}}(h, \mathsf{S}_\nu)$ are functions of $h \in \mathcal{H}_{\text{self}} \cap \mathcal{H}_{\text{train}}$ via the relations in $(34)_1$ and $(35)_1$.

5.3. Fitting problem. We will seek a best-fit parameter set $\mathcal{P}$ for the *cgDNA* model via the optimization problem

$$\mathcal{P} = \underset{h \in \mathcal{H}_{\text{self}} \cap \mathcal{H}_{\text{train}}}{\operatorname{argmax}} \; \mathcal{F}(h). \tag{52}$$

Thus a best-fit parameter set is one that maximizes and hence comes closest to achieving the upper bound for the function $\mathcal{F}(h)$ as outlined above. In view of (51), we note that any maximizers do not likely admit explicit characterizations, and a purely numerical approach is necessary.

Some insight into the maximization problem can be obtained by viewing the objective function $\mathcal{F}(h)$ as a composition:

$$h \;\xrightarrow{\mathcal{A}}\; \{\mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu), \sigma_{\text{cg}}(h, \mathsf{S}_\nu)\}_{\nu=1,\ldots,N} \;\xrightarrow{\mathcal{B}}\; \mathcal{F}(h). \tag{53}$$

Here $\mathcal{A}$ is the linear map defined by the matrix and vector assembly operations defined in $(34)_1$ and $(35)_1$, and $\mathcal{B}$ is the nonlinear map defined in (51). In the maximization problem it is desirable that any maximizer or best-fit parameter set be, at the very least, locally unique. A simple necessary condition for such local uniqueness is that the assembly map $\mathcal{A}$ be injective. In view of the structure of overlaps illustrated in Figure 8, the injectivity of $\mathcal{A}$ is dependent upon the diversity of sequences S_ν in the training set. Although sharp conditions can be derived, for brevity we note that it is sufficient for all sequences S_ν in the training set to have a length $|\mathsf{S}_\nu| \geq 4$, and that every possible dinucleotide appear in the interior of some S_ν, the end of some S_ν, and the beginning of some S_ν. The training set employed in this study has this property when data from both the reference and complementary strands is included, so that the injectivity of $\mathcal{A}$ can be guaranteed. On the other hand, if the training set lacks sufficient diversity, then the maximization problem will be degenerate and any maximizers will necessarily be nonisolated. For more details on properties of the training set, see [13].

Briefly, in order to establish injectivity, we show that the trivial parameter set (in which all parameters vanish) is the only solution of the homogeneous equation for the linear map $\mathcal{A}$. To this end, let X denote the last 6×6 diagonal block of an arbitrary parameter matrix $\mathsf{K}^{\alpha\beta}$, let Y denote an arbitrary 6×6 parameter matrix K^{β}, and let Z denote the first 6×6 diagonal block of an arbitrary parameter matrix $\mathsf{K}^{\beta\gamma}$. Then due to the structure of overlaps in the oligomer matrix $\mathsf{K}_{\text{cg}}(h, \mathsf{S}_\nu)$ illustrated in Figure 8, and the previously mentioned conditions on the training set sequences S_ν, $\nu = 1, \ldots, N$, the homogeneous equation for $\mathcal{A}$ implies that $X + Y + Z = 0$, $X + Y = 0$, and $Y + Z = 0$, from which we deduce that $X = 0$, $Y = 0$, and $Z = 0$. Moreover, the 6×6 blocks of the parameter matrices $\mathsf{K}^{\alpha\beta}$ not involved in overlaps must also necessarily vanish. The condition that $|\mathsf{S}_\nu| \geq 4$ guarantees that the overlaps follow the generic pattern in Figure 8, with a full set of leading, interior, and trailing types of overlaps for each of the parameter matrices. Similar conclusions hold for the parameter vectors $\sigma^{\alpha\beta}$ and σ^{β}. Hence the only solution of the homogeneous equation is trivial, and the linear map $\mathcal{A}$ is injective under the stated conditions.

Using exact expressions for the gradient and Hessian of the function $\mathcal{F}(h)$, we developed a numerical procedure for the maximization problem in (52). For any given set of training sequences S_ν, observed (Gaussian) densities $\rho_{\text{o}}(\mathsf{w}; \mathsf{S}_\nu)$, and weights ω_ν, the procedure uses an iterative Newton–Broyden method to solve the first-order necessary conditions on the gradient and thereby find critical points, which can then be classified using the second-order sufficient conditions on the Hessian. Since the conditions in $\mathcal{H}_{\text{train}}$ are open, we note that these first- and second-order conditions are equivalent to those for unconstrained optimization in the linear subspace $\mathcal{H}_{\text{self}}$. Moreover, since the construction of an explicit basis for this subspace is somewhat tedious, we note that all conditions can be expressed in terms of the coordinates of the ambient space $\mathcal{H}$ together with the orthogonal projection map onto $\mathcal{H}_{\text{self}}$. Beginning from an initial guess in the set $\mathcal{H}_{\text{self}}\cap\mathcal{H}_{\text{train}}$, the numerical procedure yields a sequence of iterates that remain in this set until either a critical point is found or the procedure is halted. As with any iterative method, the success of the procedure in the sense of converging to a critical point depends on the initial guess. Only those critical points that satisfy the sufficient conditions in (45) for model positivity are of interest, and these conditions can be checked a posteriori. In particular, since the positivity conditions are open, they do not affect the optimality conditions for a critical point.

5.4. Best-fit parameter sets.

Two different best-fit parameter sets $\mathcal{P}$ have been computed for the *cgDNA* model, referred to as *cgDNAparamset*1 and *cgDNAparamset*2 (both available for download from http://lcvmwww.epfl.ch/cgDNA/). These sets correspond to different choices for the observed densities $\rho_{\text{o}}(\mathsf{w};\mathsf{S}_\nu)$ and different assumptions about the stiffness parameter matrices K^α and $\mathsf{K}^{\gamma\beta}$.

*cgDNAparamset*1 is a previously computed set as detailed in [13]. This set was computed using the oligomer model density $\rho_{\text{rel}}(\mathsf{w};\mathsf{S}_\nu)$, based on maximum relative entropy, as the observed density for each sequence S_ν. Moreover, the maximization problem in (52) was considered on the more restricted space $\mathcal{H}_{\text{self}}\cap\mathcal{H}_{\text{train}}\cap\mathcal{H}_{\text{psd}}$ corresponding to positive-semidefinite stiffness parameter matrices, and employed unit weights for all sequences in the training ensemble. As already discussed, various matrices K^α and $\mathsf{K}^{\gamma\beta}$ in *cgDNAparamset*1 were on the semidefinite boundary within our numerical resolution, which suggests that a better fit may be possible with a larger parameter space, and moreover, there is no physical reason for demanding that the matrices K^α and $\mathsf{K}^{\gamma\beta}$ be semidefinite.

*cgDNAparamset*2 is a newly computed set. In contrast to the previous case, this set was computed using the oligomer model density $\rho_{\text{abs}}(\mathsf{w};\mathsf{S}_\nu)$, based on maximum absolute entropy, as the observed density for each sequence S_ν. And moreover, the maximization problem in (52) was considered on the parameter space $\mathcal{H}_{\text{self}}\cap\mathcal{H}_{\text{train}}$, which allows the stiffness parameter matrices to be indefinite. As before, unit weights were used for all sequences in the training ensemble. This set was computed using a Newton–Broyden method as described above and was found to satisfy the first-order necessary and second-order sufficient conditions for an isolated local maximum of the function $\mathcal{F}(h)$ on the space $\mathcal{H}_{\text{self}}\cap\mathcal{H}_{\text{train}}$. We remark that various parameter matrices K^α and $\mathsf{K}^{\gamma\beta}$ in *cgDNAparamset*2 are noticeably indefinite, and nevertheless the sufficient conditions in (45) are robustly satisfied for positivity of the model stiffness matrix for arbitrary sequences.

6. Results.

Here we compare results for the parameter sets *cgDNAparamset*1 and *cgDNAparamset*2. As we will see, the predictive capabilities of the *cgDNA* model with the second set are noticeably improved as compared with the first.

6.1. Shapes, stiffnesses.

The capabilities of the *cgDNA* model with *cgDNAparamset*1 to quantitatively predict the sequence-dependent, ground-state properties of various different oligomers have been discussed in [13], and also in [32] where comparisons to various experimental data and results from other works are made. Here we illustrate the effect of the parameter set on such predictions.

Figure 9 contains comparisons of the predicted ground-state configuration or shape vector $\mu_{\text{cg}}(\mathcal{P}, \mathsf{S}_\nu)$ and stiffness matrix $\mathsf{K}_{\text{cg}}(\mathcal{P}, \mathsf{S}_\nu)$ obtained with the two different parameter sets $\mathcal{P}$; for brevity we use the notation $\mathcal{P}_1$ and $\mathcal{P}_2$ to denote the sets *cgDNAparamset*1 and *cgDNAparamset*2. For each sequence in the training ensemble $\{\mathsf{S}_\nu\}_{\nu=1}^N$, the plot shows the relative differences

$$\text{RD}_{\text{Shape}} := \frac{||\mu_{\text{cg}}(\mathcal{P}_1, \mathsf{S}_\nu) - \mu_{\text{cg}}(\mathcal{P}_2, \mathsf{S}_\nu)||}{||\mu_{\text{cg}}(\mathcal{P}_1, \mathsf{S}_\nu)||},$$

$$\text{RD}_{\text{Stiff}} := \frac{||\mathsf{K}_{\text{cg}}(\mathcal{P}_1, \mathsf{S}_\nu) - \mathsf{K}_{\text{cg}}(\mathcal{P}_2, \mathsf{S}_\nu)||}{||\mathsf{K}_{\text{cg}}(\mathcal{P}_1, \mathsf{S}_\nu)||},$$

where $||\cdot||$ denotes a standard Euclidean or Frobenius norm as determined by the context. To avoid end effects, the entries in $\mu_{\text{cg}}(\mathcal{P}, \mathsf{S}_\nu)$ and $\mathsf{K}_{\text{cg}}(\mathcal{P}, \mathsf{S}_\nu)$ corresponding to the first and last three base pairs of each sequence S_ν are clipped, so that only differences in shape and stiffness corresponding to the interior portion of S_ν are compared.

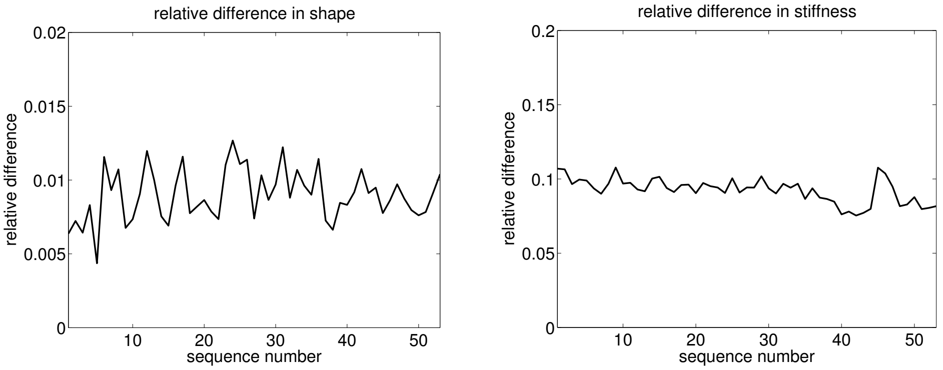

FIG. 9. *Relative differences in the* cgDNA *model shape vector* $\mu_{\text{cg}}(\mathcal{P}, \mathsf{S}_\nu)$ *and stiffness matrix* $\mathsf{K}_{\text{cg}}(\mathcal{P}, \mathsf{S}_\nu)$ *due to parameter sets* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *over all sequences* S_ν *in the training ensemble; the change in unstressed shape is only* 1% *or so, while the change in stiffnesses is around* 10%*; see text for further details. Left:* RD_{Shape} *versus* ν*. Right:* RD_{Stiff} *versus* ν*.*

Our results show that *cgDNAparamset*1 and *cgDNAparamset*2 yield predictions of ground-state shapes that closely agree, but yield predictions of ground-state stiffnesses that differ more significantly. Indeed, the relative difference is about 1% for the shape vector, and about 10% for the stiffness matrix, over all the sequences in the training ensemble. Hence the different choices for the observed densities and parameter space in the fitting problem associated with *cgDNAparamset*1 or *cgDNAparamset*2 have a significant and systematic effect on the predicted stiffnesses, but a comparatively smaller effect on the predicted shapes. Further numerical experiments reveal that about half of the difference in the predicted stiffnesses is attributable to the change of observed density (from a maximum relative entropy to a maximum absolute entropy fit at the oligomer level), whereas the other half is attributable to the change in parameter space (from positive-semidefinite to indefinite stiffness parameter matrices). As we will see below, the stiffness predictions from *cgDNAparamset*2 are significantly better than those from *cgDNAparamset*1 when compared against an accepted quantitative estimate from the literature.

6.2. Persistence lengths. One of the frequently mentioned standard statistical properties of DNA is its persistence length. There are in fact several related, but distinct, precise definitions of persistence length, and it is often not entirely clear which experimental data is being used to estimate persistence length in precisely what sense. And persistence length is believed to depend on solvent conditions and ion concentration, among other things. Nevertheless, there is a consensus in the literature that the sequence-averaged persistence length of DNA is around 150 base pairs (or bp), but with a quite large variation, as estimates between 140–180bp also appear.

We will consider here only one notion of persistence length, namely as a fit to ensemble averages of the form

$$\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle, \tag{54}$$

where $\langle \cdot \rangle$ denotes expectation with respect to a given ensemble of configurations, $\mathbf{t}_0$ is a unit vector associated with a specific base pair labeled with index 0 (usually taken to be away from the physical end of the DNA to avoid any possible end effect), and $\mathbf{t}_i$ is the analogous unit vector at the ith base pair along the DNA. Usually $\mathbf{t}_i$ is taken as some approximation to a unit tangent to the DNA, so that (54) is often described as a tangent-tangent correlation function. We will in fact take $\mathbf{t}_i$ to be the unit normal to each base pair. In simple polymer models the expectations (54) can be proven to decay exponentially with the index i and the characteristic length of this exponential decay (in number of base pairs) is one notion of persistence length ℓ_p. Here we use this notion to further illustrate the differences between our parameter sets.

A direct Monte Carlo sampling code has recently been written [27] to generate ensembles of configurations corresponding to the *cgDNA* model density $\rho_{\text{cg}}(\mathsf{w}; \mathcal{P}, \mathsf{S})$ for any given sequence S, and the expectation $\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle$ over this ensemble can be efficiently evaluated for each base pair index i along S. To illustrate the effect of the parameter set $\mathcal{P}$ on predictions of persistence length we performed two numerical experiments. We first generated an ensemble $\{\mathsf{S}_k\}_{k=1}^{1000}$ of random sequences, each of length 220bp, with equal probability of each base composition at each base pair. Then for each sequence S_k we generated two different ensembles of configurations (each with one million members): one ensemble was sampled from the density $\rho_{\text{cg}}(\mathcal{P}_1, \mathsf{S}_k, \mathsf{w})$, and the other from $\rho_{\text{cg}}(\mathcal{P}_2, \mathsf{S}_k, \mathsf{w})$, where as before $\mathcal{P}_1$ and $\mathcal{P}_2$ denote the parameter sets *cgDNAparamset*1 and *cgDNAparamset*2. From the ensemble of configurations generated using $\mathcal{P}_1$ we can compute a persistence length $\ell_p(\mathcal{P}_1, \mathsf{S}_k)$ for sequence S_k, and similarly from the ensemble generated using $\mathcal{P}_2$ we can compute $\ell_p(\mathcal{P}_2, \mathsf{S}_k)$. In each case, the persistence length is computed from a plot of $\ln\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle$ versus i; specifically, it is given by the negative of the slope of the best linear fit through the origin. (In fact we took $i = 0$ to be the 11th base pair from one end, and then sampled $\mathbf{t}_i \cdot \mathbf{t}_0$ until $i = 200$ in order to avoid end effects.)

Figure 10 shows the results of our persistence length experiments. The left panel shows the normalized histogram of the persistence length values $\ell_p(\mathcal{P}_1, \mathsf{S}_k)$ for $k = 1, \ldots, 1000$ obtained with *cgDNAparamset*1, while the right panel is the analogous normalized histogram of $\ell_p(\mathcal{P}_2, \mathsf{S}_k)$ obtained with *cgDNAparamset*2. We refer the reader to [27] for a discussion of the reasons underlying the rather wide spread of values in these histograms in each case. Here we merely observe that the apparently mathematically abstruse changes from the semidefinite, relative entropy parameters *cgDNAparamset*1 to the indefinite, maximum entropy parameters *cgDNAparamset*2, both extracted from the same MD training set data, implies a drop in the physically important sequence-averaged persistence length from 188bp to 161bp, with the latter estimate being really rather close to the consensus experimental value of 150bp. Hence the stiffness predictions from *cgDNAparamset*2 are significantly better than those from *cgDNAparamset*1 in this sense of sequence-averaged persistence length.

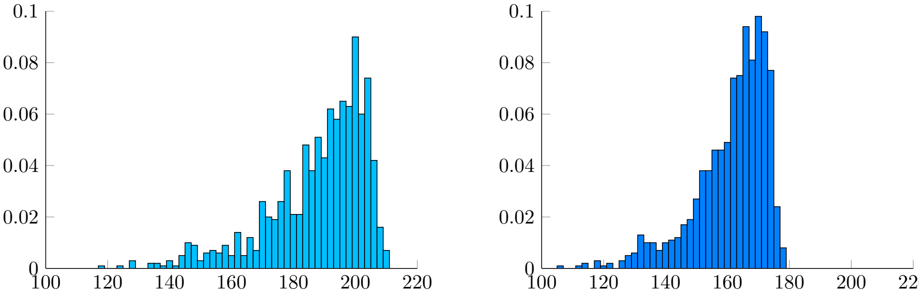

FIG. 10. *Normalized histograms of persistence length values* $\ell_p(\mathcal{P}, \mathsf{S}_k)$ *for* $k = 1, \ldots, 1000$ *in units of base pair (bp) obtained with the two parameter sets* $\mathcal{P}_1$ *and* $\mathcal{P}_2$*; see text. Left: Histogram of* $\ell_p(\mathcal{P}_1, \mathsf{S}_k)$ *with average at* 188*bp. Right: Histogram of* $\ell_p(\mathcal{P}_2, \mathsf{S}_k)$ *with average at* 161*bp.*

7. Summary and conclusions.

We have described a procedure for estimating the material parameters in a coarse-grain rigid-base model of DNA, with sequence-dependent nearest-neighbor interactions, referred to as the *cgDNA* model. Beginning from an extensive database of atomic-resolution MD simulations of a set of training oligomers in explicit solvent, the procedure delivers a complete parameter set for the *cgDNA* model, which can then be used to predict the ground-state configuration, stiffness, and other properties of arbitrary oligomers.

In the central step of our procedure, an estimated configurational mean vector and covariance matrix for each training oligomer is fit by a descriptive Gaussian model with an assumed banded structure in the stiffness matrix that expresses a nearest-neighbor interaction assumption. For this step, we compared two fitting strategies based on maximizing either an absolute or relative entropy. Due to the chain-like structure of DNA and the convergence characteristics of MD time series data, we argued that a fit based on maximum absolute entropy was more natural than maximum relative entropy for a model with nearest-neighbor type interactions. Specifically, the approach based on absolute entropy employs data from only a band about the diagonal of the estimated covariance matrix, whereas the approach based on relative entropy employs data from the entire estimated covariance matrix, and there is numerical evidence to suggest that the data that is close to the diagonal has a smaller error with respect to its assumed equilibrium or stationary value than the data that is far away. Moreover, the maximum absolute entropy fit can be constructed using a simple, local inversion algorithm, whereas the relative entropy fit requires numerical optimization techniques. And furthermore, the approach based on maximum absolute entropy can be adapted to fit higher-order, beyond-Gaussian models in a natural way that may be more convenient than an approach based on relative entropy.

In the final step of our procedure, we estimated parameters for the *cgDNA* model by fitting the Gaussian description of each training oligomer. In this step, we examined various assumptions on the choice of parameter space. An important requirement on the *cgDNA* model is that it produce a positive-definite stiffness matrix for any arbitrary oligomer. This requirement can be guaranteed under various different restrictions on the stiffness parameter matrices. In previous work [13], these parameter matrices were assumed to be positive-semidefinite, which complicates the parameter space and the associated numerical treatment of the parameter fitting problem. In contrast, here we showed that the positivity condition on the *cgDNA* model can be achieved with parameter matrices that are indefinite. The lifting of the positive-semidefinite restriction simplifies the parameter space and allows for a faster and more efficient numerical treatment of the fitting problem.

We compared two best-fit parameter sets for the *cgDNA* model referred to as *cgDNAparamset*1 and *cgDNAparamset*2. Whereas *cgDNAparamset*1 is a previously computed set [13] based on a maximum relative entropy description of each training oligomer and positive-semidefinite restrictions on the parameter stiffness matrices, *cgDNAparamset*2 is a newly computed set based on a maximum absolute entropy description without any such restrictions on definiteness. The set *cgDNAparamset*2 is a locally unique optimizer for our fitting procedure, contains indefinite stiffness matrices, and satisfies the sufficient conditions for model positivity. The predictive capabilities of the *cgDNA* model with *cgDNAparamset*2 are noticeably improved as compared with *cgDNAparamset*1. Specifically, while the two parameter sets can each predict the sequence-dependent variations in shape within and between oligomers rather well, we find that *cgDNAparamset*2 is a significant improvement over *cgDNAparamset*1 in predicting the stiffness properties of oligomers in the sense of persistence length.

With the improved parameter estimation procedure outlined here it becomes practical to pursue various types of studies, which in turn give rise to further mathematical problems to be resolved. For instance, the impact of different MD force fields, or different solvent and ion conditions, on the coarse-grain model parameters could be studied; it is just necessary to run the appropriate set of MD simulations, thereby modify the training data set, and reapply the parameter estimation procedure described here. Similarly, if a coarse-grain model of methylated bases is desired, this can be done provided that an appropriate set of MD simulations is available, as has already been carried out for a rigid-base-pair model [30]; in this case there will be a larger parameter set to allow for methylated and unmethylated bases. In addition to a richer alphabet of bases, it would also be of interest to extend the coarse-grain model and its parameterization to explicitly include information from the two backbones, such as the phosphate groups. In each of these types of studies, the coarse-grain parameters would be estimated from a training data set generated by MD simulation. In view of the computational expense associated with MD simulations, it would also be of interest to mathematically characterize the smallest possible training data set from which a complete set of coarse-grain parameters could be robustly estimated in a locally unique way.

Acknowledgments. It is a pleasure for the authors to thank G. Strang and C. Uhler for stimulating discussions on maximum entropy estimation.

REFERENCES

[1] 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, *Molecular dynamics simulations of the* 136 *unique tetranucleotide sequences of DNA oligonucleotides.* I: *Research design and results on d(CpG) steps*, Biophys. J., 87 (2004), pp. 3799–3813.

[2] M. Buehner, *Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation*, Monthly Weather Rev., 140 (2012), pp. 617–636.

[3] J. Burg, D. Luenberger, and D. Wenger, *Estimation of structured covariance matrices*, Proc. IEEE, 70 (1982), pp. 963–974.

[4] C. Calladine, H. Drew, B. F. Luisi, and A. A. Travers, *Understanding DNA*, 3rd ed., Elsevier, New York, 2004.

[5] C. Danforth, E. Kalnay, and T. Miyoshi, *Estimating and correcting global weather model error*, Monthly Weather Rev., 135 (2007), pp. 281–299.

[6] A. Dempster, *Covariance selection*, Biometrics, 28 (1972), pp. 157–175.

[7] 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, *Definitions and nomenclature of nucleic acid structure parameters*, J. Mol. Biol., 205 (1989), pp. 787–791.

[8] 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, *Molecular dynamics simulations of the* 136 *unique tetranucleotide sequences of DNA oligonucleotides.* II: *Sequence context effects on the dynamical structures of the* 10 *unique dinucleotide steps*, Biophys. J., 89 (2005), pp. 3721–3740.

[9] M. El Hassan and C. Calladine, *The assessment of the geometry of dinucleotide steps in double-helical DNA: A new local calculation scheme*, J. Mol. Biol., 251 (1995), pp. 648–664.

[10] A. Ferrante and M. Pavon, *Matrix completion á la Dempster by the principle of parsimony*, IEEE Trans. Inform. Theory, 57 (2011), pp. 3925–3931.

[11] R. Furrer, M. Genton, and D. Nychka, *Covariance tapering for interpolation of large spatial datasets*, J. Comput. Graph. Statist., 15 (2006), pp. 502–523.

[12] J. Glowacki, *Computation and Visualization in Multiscale Modelling of DNA Mechanics*, Ph.D. Thesis 7062, EPFL, Lausanne, Switzerland, 2016.

[13] O. Gonzalez, D. Petkevičiūtė, and J. Maddocks, *A sequence-dependent rigid-base model of DNA*, J. Chem. Phys., 138 (2013), 055102.

[14] R. Grone, C. Johnson, E. Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[15] R. Hogg and A. Craig, *Introduction to Mathematical Statistics*, 3rd ed., Macmillan, New York, 1970.

[16] E. Jaynes, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), pp. 620–630.

[17] E. Jaynes, *Information theory and statistical mechanics* II, Phys. Rev., 108 (1957), pp. 171–190.

[18] E. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[19] C. Johnson and M. Lundquist, *Local inversion of matrices with sparse inverses*, Linear Algebra Appl., 277 (1998), pp. 33–39.

[20] S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.

[21] F. Lankas, O. Gonzalez, L. Heffler, G. Stoll, M. Moakher, and J. Maddocks, *On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations*, Phys. Chem. Chem. Phys., 11 (2009), pp. 10565–10588.

[22] S. Lauritzen, *Graphical Models*, Clarendon Press, Oxford, UK, 1996.

[23] R. Lavery, M. Moakher, J. Maddocks, D. Petkevičiūtė, and K. Zakrzewska, *Conformational analysis of nucleic acids revisited: Curves+*, Nucleic Acids Res., 37 (2009), pp. 5917–5929.

[24] 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, *A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA*, Nucleic Acids Res., 38 (2010), pp. 299–313.

[25] A. Majda and X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*, Cambridge University Press, Cambridge, UK, 2006.

[26] F. Michor, J. Liphardt, M. Ferrari, and J. Widom, *What does physics have to do with cancer?*, Nature Reviews Cancer, 11 (2011), pp. 657–670.

[27] J. Mitchell, J. Glowacki, A. Grandchamp, R. Manning, and J. Maddocks, *Sequence-dependent persistence lengths of DNA*, J. Chem. Theory Comput., 13 (2017), pp. 1539–1555, https://doi.org/10.1021/acs.jctc.6b00904.

[28] 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, *A standard reference frame for the description of nucleic acid base-pair geometry*, J. Mol. Biol., 313 (2001), pp. 229–237.

[29] M. Pasi, J. Maddocks, D. Beveridge, T. Bishop, D. Case, T. Cheatham, III, P. Dans, B. Jayaram, F. Lankas, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Perez, D. Petkevičiūtė, N. Spackova, J. Sponer, K. Zakrzewska, and R. Lavery, *μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA*, Nucleic Acids Res., 42 (2014), pp. 12272–12283.

[30] A. Perez, C. Castellazzi, F. Battistini, K. Collinet, O. Flores, O. Deniz, M. Ruiz, D. Torrents, R. Eritja, M. Soler-Lopez, and M. Orozco, *Impact of methylation on the physical properties of DNA*, Biophys. J., 102 (2012), pp. 2140–2148, https://doi.org/10.1016/j.bpj.2012.03.056.

[31] D. Petkevičiūtė, *A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulations*, Ph.D. Thesis 5520, EPFL, Lausanne, Switzerland, 2012.

[32] D. Petkevičiūtė, M. Pasi, O. Gonzalez, and J. Maddocks, *cgDNA: A software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA*, Nucleic Acids Res., 42 (2014), e153, https://doi.org/10.1093/nar/gku825.

[33] H. Sang, M. Jun, and J. Huang, *Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors*, Ann. Appl. Stat., 5 (2011), pp. 2519–2548.

[34] R. Schleif, *DNA looping*, Ann. Rev. Biochem., 61 (1992), pp. 199–223.

[35] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. Moore, J. Wang, and J. Widom, *A genomic code for nucleosome positioning*, Nature, 442 (2006), pp. 772–778.

[36] T. Speed and H. Kiiveri, *Gaussian Markov distributions over finite graphs*, Ann. Statist., 14 (1986), pp. 138–150.

[37] J. Stroud and T. Bengtsson, *Sequential state and variance estimation within the ensemble Kalman filter*, Monthly Weather Rev., 135 (2007), pp. 3194–3208.

[38] B. Sturmfels and C. Uhler, *Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry*, Ann. Inst. Statist. Math., 62 (2010), pp. 603–638.

[39] J. Walter, O. Gonzalez, and J. Maddocks, *On the stochastic modeling of rigid body systems with application to polymer dynamics*, Multiscale Model. Simul., 8 (2010), pp. 1018–1053, https://doi.org/10.1137/090765705.