Supplementary material to "A sequence-dependent rigid-base model of DNA" by O. Gonzalez, D. Petkeviciute, and J.H. Maddocks, J. Chemical Physics, 138 (5), 2013.

Supplement to Section II.A: A full description of the oligomer configuration coordinates

Origins and frames for bases, basepairs, and junctions

As described in the main article, a DNA oligomer comprising n basepairs is identified with a sequence of bases $\mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$, with $\mathtt{X}_a \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$, listed in the $5'$ to $3'$ direction along a chosen reference backbone. The basepairs associated with this sequence are denoted by $(\mathtt{X},\overline{\mathtt{X}})_a$ where $\overline{\mathtt{X}}$ is defined as the Watson-Crick complement of $\mathtt{X}$, and the notation implies that the base $\mathtt{X}$ is attached to the reference backbone, while $\overline{\mathtt{X}}$ is on the complementary backbone. We use the *Curves+* [S10] implementation of the Tsukuba convention [S15] to assign a reference point $\boldsymbol{r}^a$ and a right-handed, orthonormal frame $\{\boldsymbol{d}_i^a\}$ $(i=1,2,3)$ to each base $\mathtt{X}_a$. Likewise, a point $\overline{\boldsymbol{r}}^a$ and frame $\{\overline{\boldsymbol{d}}_i^a\}$ are assigned to each complementary base $\overline{\mathtt{X}}_a$, but, because the two strands are anti-parallel, an additional half turn (about $\overline{\boldsymbol{d}}_1^a$) is included in the definition of $\{\overline{\boldsymbol{d}}_i^a\}$ so that the two frames $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$ are close to parallel when the basepair is close to its undeformed conformation. An illustration of the base origins and frames is provided in Figure S1.

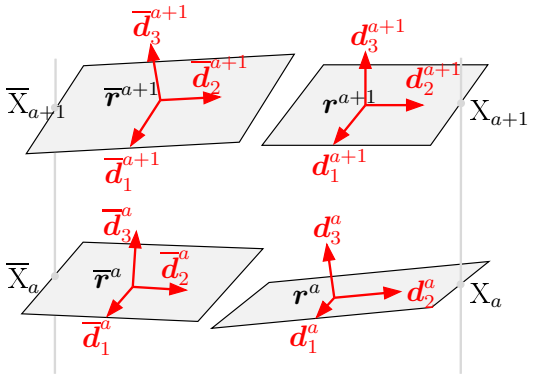


Figure S1: A schematic view of rigid bases with their reference points or origins $\boldsymbol{r}^a$ and $\overline{\boldsymbol{r}}^a$, and frames $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$, $i=1,2,3$. By construction, and despite the two backbones of DNA being anti-parallel, the two frames $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$ are close to parallel for small deformations of the B-form double-helix.

In a rigid-base model, the positions of the non-hydrogen atoms in each base relative to the associated reference point and frame are considered to be constant. As a consequence, once the reference point and frame of each base are specified, so too are the positions of all the non-hydrogen atoms. Estimated values for these idealized atomic positions in each basepair in their ideal rigid form are tabulated in [S15], while the positioning of atoms with respect to each of our base frames for each of the four possible rigid bases $\mathtt{X}_a \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$ is explained in [S10]. Independent of the details, for our purposes the important point is that in our model the configuration

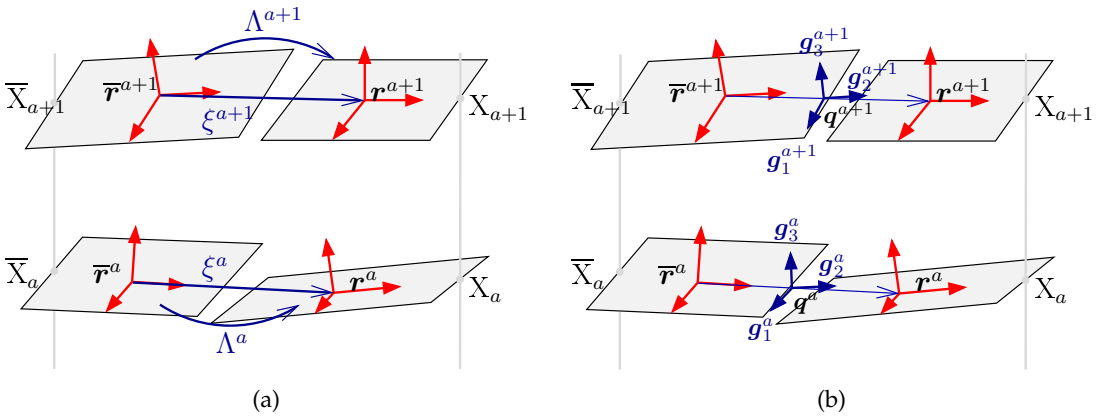
of a DNA molecule consisting of n basepairs is completely defined by the reference points $\boldsymbol{r}^a$ and $\overline{\boldsymbol{r}}^a$ and the frames $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$ $(a = 1, \ldots, n)$. These points and frames are in turn uniquely defined by component vectors $r^a, \overline{r}^a \in \mathbb{R}^3$ and rotation matrices $D^a, \overline{D}^a \in \mathbb{R}^{3\times 3}$, where $r_i^a = \boldsymbol{e}_i \cdot \boldsymbol{r}^a$, $\overline{r}_i^a = \boldsymbol{e}_i \cdot \overline{\boldsymbol{r}}^a$, $D_{ij}^a = \boldsymbol{e}_i \cdot \boldsymbol{d}_j^a$ and $\overline{D}_{ij}^a = \boldsymbol{e}_i \cdot \overline{\boldsymbol{d}}_j^a$. Here $\{\boldsymbol{e}_i\}$ denotes an arbitrary, but fixed laboratory reference frame. In terms of these components, we have

$$\boldsymbol{d}_j^a = \sum_{i=1}^{3} D_{ij}^a \boldsymbol{e}_i, \qquad \boldsymbol{r}^a = \sum_{i=1}^{3} r_i^a \boldsymbol{e}_i, \qquad \overline{\boldsymbol{d}}_j^a = \sum_{i=1}^{3} \overline{D}_{ij}^a \boldsymbol{e}_i, \qquad \overline{\boldsymbol{r}}^a = \sum_{i=1}^{3} \overline{r}_i^a \boldsymbol{e}_i. \tag{S1}$$

Then a set of internal coordinates determining the three-dimensional shape of a DNA molecule, but not its absolute position and orientation in space, is given by the relative rotation and displacement between neighboring bases both across and along the two backbone strands. The relative rotation and displacement between the bases $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$ across the strands can be described in the general form

$$\boldsymbol{d}_j^a = \sum_{i=1}^{3} \Lambda_{ij}^a \overline{\boldsymbol{d}}_i^a, \qquad \boldsymbol{r}^a = \overline{\boldsymbol{r}}^a + \sum_{i=1}^{3} \xi_i^a \boldsymbol{g}_i^a, \tag{S2}$$

where $\Lambda^a \in \mathbb{R}^{3\times 3}$ is a rotation matrix that describes the orientation of frame $\{\boldsymbol{d}_i^a\}$ with respect to $\{\overline{\boldsymbol{d}}_i^a\}$, $\xi^a \in \mathbb{R}^3$ is a vector of intra-basepair translational coordinates which describes the position of $\boldsymbol{r}^a$ with respect to $\overline{\boldsymbol{r}}^a$, and $\{\boldsymbol{g}_i^a\}$ is a right-handed, orthonormal frame intermediate to the base frames $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$; $\{\boldsymbol{g}_i^a\}$ is defined to be the rotational average of the base frames $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$ and is referred to as the basepair frame associated with $(\mathtt{X}, \overline{\mathtt{X}})_a$. We give an explicit formula for it below. We also introduce the basepair reference point $\boldsymbol{q}^a$ that is defined to be the Euclidean average of the base reference points $\boldsymbol{r}^a$ and $\overline{\boldsymbol{r}}^a$.



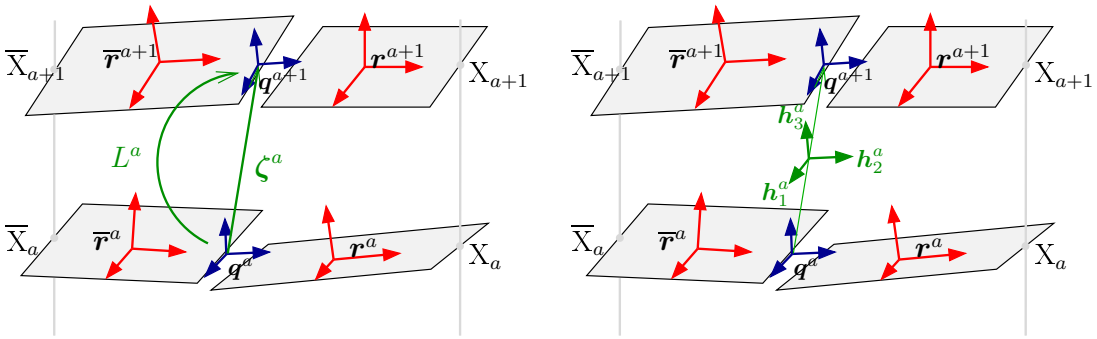
(a) (b)

Figure S2: (a) Illustration of intra-basepair translation and rotation. (b) Illustration of the basepair frame (blue) and the two base frames (red) in each basepair. The intra-basepair translation and rotation coordinates are given as components in the basepair frame.

The relative displacement and rotation along the strands between the basepair origins and frames associated with $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ can be described in the general form

$$\boldsymbol{g}_j^{a+1} = \sum_{i=1}^{3} L_{ij}^a \boldsymbol{g}_i^a, \qquad \boldsymbol{q}^{a+1} = \boldsymbol{q}^a + \sum_{i=1}^{3} \zeta_i^a \boldsymbol{h}_i^a, \tag{S3}$$

where $L^a \in \mathbb{R}^{3\times3}$ is a rotation matrix that describes the orientation of frame $\{\boldsymbol{g}_i^{a+1}\}$ with respect to $\{\boldsymbol{g}_i^a\}$, $\zeta^a \in \mathbb{R}^3$ is a vector of inter-basepair translational coordinates that describes the position of $\boldsymbol{q}^{a+1}$ with respect to $\boldsymbol{q}^a$, and $\{\boldsymbol{h}_i^a\}$ is a right-handed, orthonormal frame intermediate to the two basepair frames $\{\boldsymbol{g}_i^a\}$ and $\{\boldsymbol{g}_i^{a+1}\}$; $\{\boldsymbol{h}_i^a\}$ is referred to as the junction frame associated with $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ and is defined to be the rotational average of the frames $\{\boldsymbol{g}_i^a\}$ and $\{\boldsymbol{g}_i^{a+1}\}$. Again it is defined explicitly below.



(a) (b)

Figure S3: (a) Illustration of inter-basepair translation and rotation. (b) Illustration of the junction frame (green) and the two basepair frames (blue) in a pair of adjacent basepairs. The inter-basepair translation and rotation coordinates are given as components in the junction frame.

The Cayley parameterization of proper rotation matrices

To describe rotations we will use Cayley (or Euler-Rodrigues) parameters. This parameterization of the group of proper rotations is the same as that used in the program *Curves+* [S10] and in [S11] up to different scalings. Versions of these parameters are described in many places, e.g. [S1] and [S4]. Another popular choice of rotation matrix parameters is Euler angles, used, for example, in the *3DNA* software package [S12]. The main difference between the two parameterizations is that in the case of Cayley parameters every rotation is represented by a single, but variable, rotation axis with one associated rotation angle, while in the Euler angle case every rotation is decomposed into three elementary rotations, i.e. into consecutive rotations around three specific axes through three variable angles. The difference is well illustrated in [S1] for example.

The Cayley parameters for a rotation matrix arise as a consequence of Euler's Theorem on rotations, which states that any proper, or right-handed, rotation matrix $Q \in \mathbb{R}^{3\times3}$ can be represented as a pure, or simple, rotation about some axis (or unit vector) $k \in \mathbb{R}^3$ through an angle $\varphi \in [0, \pi]$. That is, for any vector $v \in \mathbb{R}^3$, the transformed vector $Qv \in \mathbb{R}^3$ corresponds to a right-handed rotation of v about k through an angle φ. For angles in the interval $\varphi \in (0, \pi)$ the relation between Q and the axis-angle pair (k, φ) is uniquely invertible, whereas for the two angles $\varphi = 0$ and $\varphi = \pi$ it is not. Given the pair (k, φ), the corresponding Q is given by the Euler-Rodrigues formula

$$Q = \cos\varphi\, I + (1 - \cos\varphi)\, k \otimes k + \sin\varphi\, k^{\times}, \tag{S4}$$

where $I \in \mathbb{R}^{3\times3}$ is the identity matrix, $k \otimes k = kk^T \in \mathbb{R}^{3\times3}$ is the rank-one outer-product matrix,

and $k^\times \in \mathbb{R}^{3\times 3}$ is the skew matrix

$$k^\times = \begin{pmatrix} 0 & -k_3 & k_2 \\ k_3 & 0 & -k_1 \\ -k_2 & k_1 & 0 \end{pmatrix}, \tag{S5}$$

which is defined such that $(k^\times)v = k \times v$ for any $v \in \mathbb{R}^3$. Conversely, given a rotation matrix Q, the corresponding pair (k, φ) is given by

$$\varphi = \arccos\left(\frac{\operatorname{tr} Q - 1}{2}\right), \qquad k = \frac{1}{2\sin\varphi}\begin{pmatrix} Q_{32} - Q_{23} \\ Q_{13} - Q_{31} \\ Q_{21} - Q_{12} \end{pmatrix} = \frac{1}{2\sin\varphi}\operatorname{vec}[Q - Q^T]. \tag{S6}$$

Here $\operatorname{vec}[A] \in \mathbb{R}^3$ denotes the axial vector associated with a skew matrix $A \in \mathbb{R}^{3\times 3}$, which is defined as $\operatorname{vec}[A] = (A_{32}, A_{13}, A_{21})^T$. We remark that the relations in (S6) follow from (S4) upon noting that $Q - Q^T = 2\sin\varphi\, k^\times$, $\operatorname{tr} Q = 1 + 2\cos\varphi$ and $\operatorname{vec}[k^\times] = k$.

The relation between Q and the pair (k, φ) is special in either of the two cases $\varphi = 0$ or $\varphi = \pi$. As we are assuming that Q is a proper, or right-handed, rotation matrix, it satisfies $Q^{-1} = Q^T$ and $\det Q = +1$. In this case, the three eigenvalues of Q are $(+1, e^{-i\varphi}, e^{i\varphi})$, where φ is the rotation angle, and the rotation axis k is a real eigenvector of Q with unit eigenvalue. When Q is symmetric, which happens precisely in either of the two cases of rotation angle $\varphi = 0$ or $\varphi = \pi$, we have $\sin\varphi = 0$ and $Q - Q^T = 0$, so that the expression for k in (S6) is undefined. Specifically, when $\varphi = 0$, we have $Q = I$, and the rotation axis k is completely arbitrary; that is, for any choice of k, a rotation through zero angle about k will give $Q = I$. When $\varphi = \pi$, the rotation axis k is uniquely defined up to sign as a dominant (unit) eigenvector of the matrix $Q + I$, which has eigenvalues $(0, 0, 2)$. For rotations through π either choice of sign for k yields the same rotation.

While rotations through an angle of $\varphi = \pi$ are unlikely to occur between neighboring bases in our application to B-form DNA, and accordingly are not of significant interest here, it is certainly desirable to seek a representation of Q that is well-behaved for rotation angles close to $\varphi = 0$. By a Cayley parameterization of Q we mean one in which the axis-angle pair (k, φ) is represented by a single vector $\eta \in \mathbb{R}^3$. Specifically, given any non-negative, invertible function $f(\varphi)$, we can consider a mapping of the form $\eta = f(\varphi)k$, which can be inverted to yield $k = \eta/|\eta|$ and $\varphi = f^{-1}(|\eta|)$, where $|\cdot|$ denotes the standard Euclidean norm. Notice that the direction of η encodes the rotation axis, while its magnitude encodes the rotation angle. When these relations are substituted into (S4) and (S6) we obtain a representation of Q in terms of the single vector η, rather than the pair (k, φ). Moreover, the function f can be chosen such that the relation between Q and η has better mathematical properties than that between Q and (k, φ). In this work, as in [S11], we use the choice $f(\varphi) = 2\tan(\varphi/2)$. Then we obtain the relations

$$Q = (I + \frac{1}{2}\eta^\times)(I - \frac{1}{2}\eta^\times)^{-1} =: \operatorname{cay}(\eta), \qquad \eta = \frac{1}{1 + \operatorname{tr} Q}\operatorname{vec}[Q - Q^T] =: \operatorname{cay}^{-1}(Q). \tag{S7}$$

The above relations provide a one-to-one correspondence between a rotation matrix Q and a single vector η. In contrast to the axis-angle parameterization in (S4) and (S6), the Cayley parameterization in (S7) is well-defined for small rotation angles near $\varphi = 0$, which corresponds to $|\eta| = 0$. Similar to the axis-angle parameterization, the Cayley parameterization becomes undefined at a rotation angle of $\varphi = \pi$ or equivalently when $\operatorname{tr} Q = -1$, which corresponds to $|\eta| = \infty$. Thus the Cayley parameterization in (S7) based on the choice $f(\varphi) = 2\tan(\varphi/2)$ provides a one-to-one relation between rotation matrices Q with rotation angles $0 \le \varphi < \pi$ and vectors $\eta \in \mathbb{R}^3$ with magnitudes $0 \le |\eta| < \infty$. The advantages of the above Cayley parameterization over the axis-angle

parameterization are now apparent: it is well-defined for all rotations with angles $\varphi \in [0, \pi)$ and parameter vectors $\eta \in \mathbb{R}^3$, and the problematic case of rotations with angle $\varphi = \pi$ occurs only at $|\eta| = \infty$. We remark that one way to interpret the non-dimensionalization described in Section II.D of the main article is as a simple re-scaling of the form $f(\varphi) = 10 \tan(\varphi/2)$.

Different choices can certainly be made for the function f, which lead to different types of Cayley parameterizations. For example, the choice $f(\varphi) = 180\varphi/\pi$ (the value of φ in degrees) is used in *Curves+* [S10]. The choice $f(\varphi) = \sin(\varphi/2)$ gives the vector part of a unit quaternion (or three of the four scalar Euler parameters) with $\cos(\varphi/2)$ being the scalar part of the quaternion (or the fourth Euler parameter). Geometrically this choice can be regarded as a stereographic projection of the Cayley parameter vector in $\mathbb{R}^3$ onto the upper unit hemisphere in four dimensions with the singularity for rotations through π, or the point at infinity, mapped to the equator. Adding the lower hemisphere in the appropriate way then leads to the classic quaternion, or Euler parameter, singularity-free, double covering of the rotation group, which can be useful for smoothly tracking the large absolute orientations of basepair frames that can arise for a long oligomer. However, in this article we are only concerned with internal oligomer coordinates, or relative rotations, where rotations through π are unimportant. Consequently, the Cayley parameter vector $\eta \in \mathbb{R}^3$ is for us a more convenient choice of internal rotational coordinates.

Cayley parameters and mid-frames

In our application to B-form DNA, we consider reference frames attached to each base of an oligomer, so that it is necessary to describe the rotation that transforms the basis vectors of one frame to those of another. In particular we want a description that transforms simply under the Watson-Crick symmetry of reversing the roles of the reference $\mathtt{X}$ and complementary backbone $\overline{\mathtt{X}}$. To this end we will be interested in constructing an intermediate, or middle frame, between any given two frames. If the matrix of components (or direction cosine matrix) of a right-handed, orthonormal frame is $\overline{D} = [\overline{d}_1\ \overline{d}_2\ \overline{d}_3] \in \mathbb{R}^{3\times3}$, where the $\overline{d}_i \in \mathbb{R}^3$ (column vectors) are the components of the three basis vectors of that frame, and $D = [d_1\ d_2\ d_3] \in \mathbb{R}^{3\times3}$ is the matrix of components of another right-handed, orthonormal frame, then there exists a unique proper rotation matrix $Q \in \mathbb{R}^{3\times3}$ with the property that $d_i = Q\overline{d}_i$. Specifically, in matrix notation we have

$$D = Q\overline{D} \qquad \text{where} \qquad Q = D\overline{D}^T. \tag{S8}$$

By the middle frame $G = [g_1\ g_2\ g_3] \in \mathbb{R}^{3\times3}$ between $\overline{D}$ and D we mean the frame defined by

$$G = \sqrt{Q}\,\overline{D}, \tag{S9}$$

where $\sqrt{Q}$ is the half-rotation matrix from $\overline{D}$ to D, i.e. the rotation about the same axis k, but by the half-angle $\varphi/2$, so that $Q = \sqrt{Q}\sqrt{Q}$. In this way, we have $G = \sqrt{Q}\,\overline{D}$ and $D = \sqrt{Q}\,G$.

The relation in (S8) can be expressed in an alternative form. Specifically, it can be written as

$$D = \overline{D}\Lambda = \left[\sum_{i=1}^{3} \Lambda_{i1}\overline{d}_i\ \ \sum_{i=1}^{3} \Lambda_{i2}\overline{d}_i\ \ \sum_{i=1}^{3} \Lambda_{i3}\overline{d}_i\right], \tag{S10}$$

where $\Lambda \in \mathbb{R}^{3\times3}$ is a rotation matrix defined by

$$\Lambda = \overline{D}^T D = \overline{D}^T Q \overline{D} = D^T Q D. \tag{S11}$$

If Λ has a unique rotation axis k_Λ, which by definition satisfies $\Lambda k_\Lambda = k_\Lambda$, and Q has a unique rotation axis k_Q, which by definition satisfies $Qk_Q = k_Q$, then from the above relations we deduce that

$$k_\Lambda = \overline{D}^T k_Q = D^T k_Q, \tag{S12}$$

which shows that k_Λ and k_Q are the components of the same geometric vector $\boldsymbol{k}$, but expressed in different frames. Indeed, whereas k_Q are the components of $\boldsymbol{k}$ in the lab frame $\{\boldsymbol{e}_i\}$, $\overline{D}^T k_Q$ are the components of $\boldsymbol{k}$ in the frame $\{\overline{\boldsymbol{d}}_i\}$, and $D^T k_Q$ are the components of $\boldsymbol{k}$ in the frame $\{\boldsymbol{d}_i\}$. Notice that, since Q rotates the frame $\{\overline{\boldsymbol{d}}_i\}$ onto the frame $\{\boldsymbol{d}_i\}$ about the axis defined by $\boldsymbol{k}$, it follows that the components of $\boldsymbol{k}$ in these two frames are the same, as expressed in (S12). Moreover, because the matrices Λ and Q are related through a similarity transformation, their eigenvalues are also the same. From these considerations we deduce that the two matrices Λ and Q represent the same geometric rotation, but expressed in different frames. We call Λ the relative rotation matrix and Q the absolute rotation matrix from $\overline{D}$ to D. In this work, we use relative rotation matrices to define the internal coordinates for our rigid-base model of B-form DNA.

Explicit formulæ for the rigid-base double-chain topology

The intra-basepair relative rotation and displacement between $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$ are defined in (S2). From (S2) and (S1) we have $\Lambda^a = (\overline{D}^a)^T D^a$, and from this rotation matrix we extract an intra-basepair rotation vector $\vartheta^a = \mathrm{cay}^{-1}(\Lambda^a) \in \mathbb{R}^3$ and an intra-basepair rotation angle $\varphi^a = f^{-1}(|\vartheta^a|)$ using the Cayley parameterization described in (S7). From these quantities, and with the notation that the frame $\{\boldsymbol{g}_i^a\}$ has component matrix $G^a \in \mathbb{R}^{3\times 3}$ where $G^a_{ij} = \boldsymbol{e}_i \cdot \boldsymbol{g}_j^a$, we can construct the half-rotation matrix $\sqrt{\Lambda^a}$, and then the basepair frame matrix is given by $G^a = \overline{D}^a\sqrt{\Lambda^a}$. In view of (S2) and (S1), the intra-basepair translation components are then given by $\xi^a = (G^a)^T(r^a - \overline{r}^a)$. An illustration of intra-basepair rotational and translational parameters is shown in Figure S2.

To describe the relative rotation and displacement between neighboring bases along the strands we consider the basepair frame $\{\boldsymbol{g}_i^a\}$ and basepair reference point $\boldsymbol{q}^a = (\overline{\boldsymbol{r}}^a + \boldsymbol{r}^a)/2$ associated with $(\mathtt{X}, \overline{\mathtt{X}})_a$, and the analogous frame $\{\boldsymbol{g}_i^{a+1}\}$ and point $\boldsymbol{q}^{a+1}$ associated with $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$. The relative rotation and displacement between $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ along the strands is described in (S3). The frame $\{\boldsymbol{h}_i^a\}$ has component matrix $H^a \in \mathbb{R}^{3\times 3}$ where $H^a_{ij} = \boldsymbol{e}_i \cdot \boldsymbol{h}_j^a$, and the points $\boldsymbol{q}^a$ and $\boldsymbol{q}^{a+1}$ have component vectors $q^a, q^{a+1} \in \mathbb{R}^3$ where $q^a = (\overline{r}^a + r^a)/2$ and $q^{a+1} = (\overline{r}^{a+1} + r^{a+1})/2$. Similar to before, we have $L^a = (G^a)^T G^{a+1}$, and from this rotation matrix we extract an inter-basepair rotation vector $\theta^a = \mathrm{cay}^{-1}(L^a) \in \mathbb{R}^3$ and an inter-basepair rotation angle $\phi^a = f^{-1}(|\theta^a|)$ using the Cayley parameterization described in (S7). From these quantities, we can construct the half-rotation matrix $\sqrt{L^a}$, and then the junction frame matrix is given by $H^a = G^a\sqrt{L^a}$. Similar to before, the inter-basepair translation components are then given by $\zeta^a = (H^a)^T(q^{a+1} - q^a)$. An illustration of the inter-basepair rotational and translational parameters is shown in Figure S3.

The choice of writing intra- and inter-basepair rotational and translational parameters with respect to coordinates in the mid (respectively basepair and junction) frames, when combined with the additional half-turn about $\overline{\boldsymbol{d}}_1^a$ in the definition of the frame $\{\overline{\boldsymbol{d}}_i^a\}$, implies that the transformation of the coordinates under the Watson-Crick symmetry of exchange of roles of the reference and complementary backbones is particularly simple. Specifically the four 1-components of each set of intra- and inter-basepair translational and rotational coordinates, i.e. Buckle, Shear, Tilt and Shift, all change sign, while the remaining eight other components are all unchanged. The particularly simple manifestation of the Watson-Crick symmetry implies the simple transformation rules between complementary parameter sets that are discussed in the main article, and the symmetry that must arise in oligomers with palindromic sequences.

The above formulas give the intra-basepair coordinates $y^a = (\vartheta, \xi)^a$ in terms of the relative rotation and displacement between bases $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$ across the strands, whereas the inter-basepair coordinates $z^a = (\theta, \zeta)^a$ are given by the relative rotation and displacement between the pairs $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ along the strands. Contrariwise, the origins and frames in a complete rigid-base description of the configuration of a DNA oligomer can be constructed from the intra- and inter-basepair coordinates provided that six, additional, external coordinates $z^0 = (\theta, \zeta)^0 \in \mathbb{R}^6$ for the first basepair frame and reference point with respect to the lab frame $\{\boldsymbol{e}_i\}$ are provided. Explicitly, all the frames and reference points for the bases $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$ ($a = 1, \ldots, n$) can be constructed using the recursive formulas

$$\boldsymbol{g}_j^a = \sum_{i=1}^{3} L_{ij}^{a-1} \boldsymbol{g}_i{}^{a-1}, \qquad \boldsymbol{q}^a = \boldsymbol{q}^{a-1} + \sum_{i=1}^{3} \zeta_i{}^{a-1} \boldsymbol{h}_i{}^{a-1}, \tag{S13}$$

$$\overline{\boldsymbol{d}}_j^a = \sum_{i=1}^{3} (\sqrt{\Lambda^a})_{ji} \boldsymbol{g}_i^a, \qquad \overline{\boldsymbol{r}}^a = \boldsymbol{q}^a - \frac{1}{2} \sum_{i=1}^{3} \xi_i^a \boldsymbol{g}_i^a, \tag{S14}$$

$$\boldsymbol{d}_j^a = \sum_{i=1}^{3} (\sqrt{\Lambda^a})_{ij} \boldsymbol{g}_i^a, \qquad \boldsymbol{r}^a = \boldsymbol{q}^a + \frac{1}{2} \sum_{i=1}^{3} \xi_i^a \boldsymbol{g}_i^a. \tag{S15}$$

Here $\Lambda^a = \mathrm{cay}[\vartheta^a]$ is the rotation matrix corresponding to ϑ^a, $L^a = \mathrm{cay}[\theta^a]$ is the rotation matrix corresponding to θ^a, $\{\boldsymbol{h}_i^a\}$ is the junction frame with component matrix $H^a = G^a \sqrt{L^a}$ described above, $z^0 = (\theta, \zeta)^0$ are the external coordinates of the first basepair frame, and we adopt the convention that $\{\boldsymbol{g}_i^0\} = \{\boldsymbol{e}_i\}$ and $\boldsymbol{q}^0 = \boldsymbol{0}$.

Cayley vectors and statistical mechanics on the proper rotation group

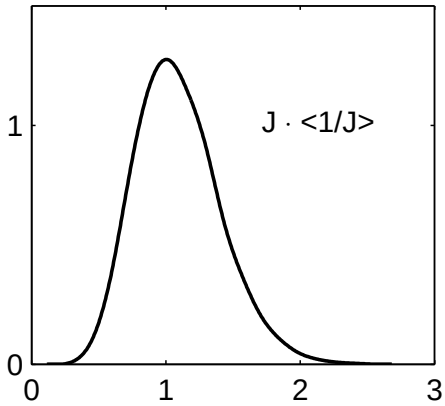


Figure S4: Histogram of the scaled Jacobian factor $J\langle 1/J \rangle$ along the time series for the training set oligomer S_1.

The proper rotation group is compact and so has a natural, associated Haar measure, which is unique up to an unimportant constant. Moreover, as discussed for example in [S17, S18], the statistical mechanics of rigid bodies naturally leads to equilibrium distributions that are Boltzmann, or pure exponential, with respect to this Haar measure. Consequently, for any particular parameterization of the rotation group, equilibrium distributions of the form of equation (6) in the main article arise, namely a Boltzmann term with an additional coordinate dependent Jacobian factor. In this regard, the particular choice of the Cayley parameterization of rotations has two associated and desirable features. First, the domain of definition of each Cayley parameter vector is the whole of $\mathbb{R}^3$ which is convenient for the explicit evaluation of Gaussian integrals. Second, the Jacobian for the Cayley parameterization has the rather simple explicit form detailed in equation (7) of the main article. In contrast, we are unaware of the analogous expression for the Jacobian associated with the *3DNA* [S12] rotational coordinates.

Even with the simple explicit expression for the Jacobian of the Cayley parameterization, our parameter extraction methodology approximates the Jacobian factor to be constant in order to be

able to benefit from various closed-form expressions for Gaussian integrals. A previous analysis [S11] of MD simulation data for one oligomer under similar conditions and using a Cayley parameterization as considered here indicates that the error associated with this approximation is rather small: various averages computed with and without the Jacobian factor differed by less than 3%. For reference, Figure S4 shows a histogram of the Jacobian factor (scaled by the average of its reciprocal) along the time series for one oligomer in our training set. Although the precise nature of the error associated with the constant approximation remains an open question, we notice that the distribution of the scaled Jacobian is rather peaked, which further suggests that the constant approximation should be reasonable. We remark that the locality of the distribution of the Jacobian depends both on the physical property of the DNA oligomer being relatively stiff, so that the high probability (or low energy) regions of configuration space are relatively localized, and on the singularity of the rotational coordinate system being far from the high probability regions. In particular, for other coordinate systems, the distribution of the associated Jacobian could be rather different.

Supplement to Section II.F: The Kullback-Leibler divergence

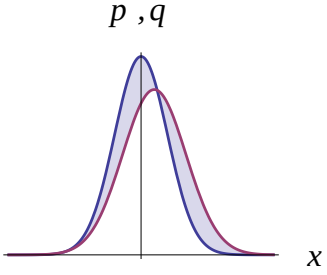
In probability theory, a divergence $D(p,q)$ is a function that measures the difference between two normalized probability density functions $p(x)$ and $q(x)$, which for our purposes are assumed to be smooth, positive functions of a variable $x \in \mathbb{R}^d$. By definition, a divergence is non-negative in the sense that $D(p,q) \geq 0$ for all density functions $p(x)$ and $q(x)$, and non-degenerate in the sense that $D(p,q) = 0$ if and only if $p(x) = q(x)$ for all $x \in \mathbb{R}^d$. Because it need not satisfy the symmetry condition $D(p,q) = D(q,p)$, nor the triangle inequality $D(p,q) \leq D(p,r) + D(r,q)$, for all density functions $p(x)$, $q(x)$ and $r(x)$, a divergence is more general than a distance or metric on the set of normalized probability densities. In this work, we employ the Kullback-Leibler divergence function defined as

$$D(p,q) := \int_{\mathbb{R}^d} p(x) \ln\left[\frac{p(x)}{q(x)}\right] \, dx = \left\langle \frac{p(x)}{q(x)} \ln\left[\frac{p(x)}{q(x)}\right] \right\rangle_q , \tag{S16}$$

where $\langle\cdot\rangle_q$ denotes expectation with respect to $q(x)$. This function, which originated in the work of Kullback and Leibler [S8, S9], provides a convenient measure of the difference between probability densities and has been employed in a number of different applications [S2, S3, S13]. It is intimately related to the notion of (relative) entropy in statistical mechanics and information theory, and provides the basis for the maximum entropy principle of statistical inference [S5, S6, S7].

The case of one-dimensional distributions

The divergence $D(p,q)$ can be illustrated simply in the case when $d = 1$, where $p(x)$ and $q(x)$ are functions of a single variable $x \in \mathbb{R}$. In this case, the divergence $D(p,q)$ provides a measure of the (unsigned) area A between the two densities as shown: $D(p,q) > 0$ if and only if $A > 0$, and $D(p,q) = 0$ if and only if $A = 0$. Although the area A itself also provides a measure of the difference between $p(x)$ and $q(x)$, the divergence has various more desirable properties due to its connection with the notion of entropy; in general, there is no simple, closed-form relation between the divergence D and the area A. Further insight can be gained in the special case when the densities are Gaussian. Specif-



ically, consider

$$p(x) = \frac{1}{\sigma_p\sqrt{2\pi}} e^{-(x-\mu_p)^2/2\sigma_p^2}, \quad q(x) = \frac{1}{\sigma_q\sqrt{2\pi}} e^{-(x-\mu_q)^2/2\sigma_q^2}, \tag{S17}$$

where μ_p and μ_q are the means and $\sigma_p > 0$ and $\sigma_q > 0$ are the standard deviations of the two densities. In the case when the two densities have arbitrary means but the same standard deviation, so that $\sigma_p = \sigma_q = \sigma$, we find by direct integration that

$$D(p,q) = \frac{(\mu_p - \mu_q)^2}{2\sigma^2}. \tag{S18}$$

Hence in this case the divergence provides a measure of the difference between means, which vanishes only when they coincide. In the case when the two densities have arbitrary standard deviations but the same mean, so that $\mu_p = \mu_q = \mu$, we find by direct integration that

$$D(p,q) = \frac{1}{2}\left(\frac{\sigma_p^2}{\sigma_q^2} - \ln\left[\frac{\sigma_p^2}{\sigma_q^2}\right] - 1\right). \tag{S19}$$

Hence in this case the divergence provides a measure of the difference between the standard deviations, which vanishes only when they coincide. More generally, in the case when the two densities have arbitrary means and standard deviations we find

$$D(p,q) = \frac{1}{2}\left(\frac{\sigma_p^2}{\sigma_q^2} - \ln\left[\frac{\sigma_p^2}{\sigma_q^2}\right] - 1\right) + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2}. \tag{S20}$$

The case of multivariate Gaussian distributions

The expression (S20) for the divergence between two scalar Gaussian distributions generalizes straightforwardly to the multivariate case. As given in equation (17) of the main article, the Kullback-Leibler divergence between two Gaussian densities ρ_{m} and ρ_{o} can be evaluated to be

$$D(\rho_{\mathrm{m}}, \rho_{\mathrm{o}}) = \frac{1}{2}\left[\mathsf{K}_{\mathrm{m}}^{-1} : \mathsf{K}_{\mathrm{o}} - \ln(\det \mathsf{K}_{\mathrm{o}}/\det \mathsf{K}_{\mathrm{m}}) - I : I\right] + \frac{1}{2}(\widehat{\mathsf{w}}_{\mathrm{m}} - \widehat{\mathsf{w}}_{\mathrm{o}}) \cdot \mathsf{K}_{\mathrm{o}}(\widehat{\mathsf{w}}_{\mathrm{m}} - \widehat{\mathsf{w}}_{\mathrm{o}}), \tag{S21}$$

where a colon denotes the standard Euclidean inner product for square matrices and I denotes the identity matrix of the same dimension as K_{m} and K_{o}. Here $\widehat{\mathsf{w}}_{\mathrm{m}}$ and $\widehat{\mathsf{w}}_{\mathrm{o}}$ are the means and K_{m} and K_{o} are the stiffness (or inverse covariance) matrices of the multivariate densities ρ_{m} and ρ_{o}. The term in brackets involves only the two stiffness matrices, and can be rewritten in the form

$$\begin{aligned} D^{\dagger}(\mathsf{K}_{\mathrm{m}}, \mathsf{K}_{\mathrm{o}}) &:= \frac{1}{2}\left[\mathsf{K}_{\mathrm{m}}^{-1} : \mathsf{K}_{\mathrm{o}} - \ln(\det \mathsf{K}_{\mathrm{o}}/\det \mathsf{K}_{\mathrm{m}}) - I : I\right] \\ &= \frac{1}{2}\sum_{i=1}^{12n-6} (\mu_i - \ln \mu_i - 1), \end{aligned} \tag{S22}$$

where μ_i are the eigenvalues of the symmetric, positive-definite, generalized eigenvalue problem

$$\mathsf{K}_{\mathrm{o}}\mathsf{v}_i = \mu_i \mathsf{K}_{\mathrm{m}}\mathsf{v}_i, \quad i = 1, \ldots, 12n-6. \tag{S23}$$

It is evident that $D^{\dagger}(\mathsf{K}_{\mathrm{m}}, \mathsf{K}_{\mathrm{o}})$ defined in (S22) is non-negative and vanishes only when $\mu_i = 1$ for all i, which implies that $\mathsf{K}_{\mathrm{o}} = \mathsf{K}_{\mathrm{m}}$, so that it is an appropriate measure of the difference between two

symmetric, positive-definite matrices [S14]. Similarly, it is evident that the eigenvalues defined in (S23) are dimensionless and so independent of the choice of length, rotation and energy scales, and that

$$D^\dagger(\mathsf{K}_\mathrm{m}, \mathsf{K}_\mathrm{o}) = D^\dagger(\mathsf{K}_\mathrm{o}^{-1}, \mathsf{K}_\mathrm{m}^{-1}). \tag{S24}$$

Analogously, the second term in (S21) is non-negative and vanishes only when the means $\widehat{\mathsf{w}}_\mathrm{m}$ and $\widehat{\mathsf{w}}_\mathrm{o}$ coincide. It is also independent of the length and rotation scales, but depends on the energy scale (which was absorbed into the stiffness matrix in the non-dimensionalization procedure). Thus the divergence in (S21) is a linear combination of the differences in the stiffnesses and means of two Gaussians, with the relative weighting dependent on the energy scale, or equivalently temperature.

Combining equation $(41)_2$ of the main article with equation (S22) above we find that the stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{M}}$ must satisfy the optimization problem

$$\begin{aligned} \mathsf{K}^*_{\mu,\mathrm{M}} &= \underset{\mathsf{K}_{\mu,\mathrm{M}}}{\operatorname{argmin}} \ \frac{1}{2}\left[\mathsf{K}^{-1}_{\mu,\mathrm{M}} : \mathsf{K}_{\mu,\mathrm{o}} - \ln(\det \mathsf{K}_{\mu,\mathrm{o}}/\det \mathsf{K}_{\mu,\mathrm{M}}) - I : I\right] \\ &= \underset{\mathsf{K}_{\mu,\mathrm{M}}}{\operatorname{argmin}} \ D^\dagger(\mathsf{K}_{\mu,\mathrm{M}}, \mathsf{K}_{\mu,\mathrm{o}}), \end{aligned} \tag{S25}$$

where the minimum is taken over the set of symmetric matrices of the specified sparsity. Equivalently, in view of equation (S24) above, we have

$$\mathsf{K}^*_{\mu,\mathrm{M}} = \underset{\mathsf{K}_{\mu,\mathrm{M}}}{\operatorname{argmin}} \ D^\dagger(\mathsf{K}^{-1}_{\mu,\mathrm{o}}, \mathsf{K}^{-1}_{\mu,\mathrm{M}}). \tag{S26}$$

Hence the matrix optimization problem can also be regarded as that of finding a stiffness matrix $\mathsf{K}_{\mu,\mathrm{M}}$ of specified sparsity such that the associated (and in general dense) model covariance matrix $\mathsf{K}^{-1}_{\mu,\mathrm{M}}$ has a minimum distance, in an appropriate sense, to the observed covariance matrix $\mathsf{K}^{-1}_{\mu,\mathrm{o}}$.

A scale for the Kullback-Leibler divergence

The Kullback-Leibler divergence is a non-dimensional and natural measure of the difference between probability densities. However, to address the question of whether the divergence between any two probability densities is large or small we need to set a scale. As described in the main article, we introduce a Kullback-Leibler divergence scale D_o for 18-mers by

$$D_\mathrm{o} = \underset{\substack{n_\mu = 18 \\ \mu_1 \neq \mu_2}}{\operatorname{avg}} D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}). \tag{S27}$$

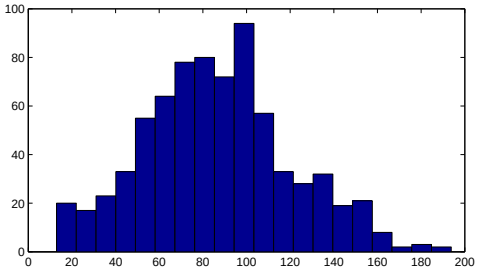


Figure S5: A histogram (frequencies versus bins) of the symmetrized Kullback-Leibler divergence $D^\mathrm{sym}(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}})$ over all distinct pairs of 18-mer sequences S_{μ_1} and S_{μ_2} in the data set of Table SI.

Because $D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}})$ and $D(\rho_{\mu_2,\mathrm{o}}, \rho_{\mu_1,\mathrm{o}})$ are both counted, the above average is actually determined by a symmetrized version of the divergence defined as

$$D^\mathrm{sym}(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}) = \frac{1}{2}[D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}) + D(\rho_{\mu_2,\mathrm{o}}, \rho_{\mu_1,\mathrm{o}})]. \tag{S28}$$

Specifically, the average in (S27) is equivalent to

$$D_\mathrm{o} = \underset{\substack{n_\mu = 18 \\ \mu_1 \neq \mu_2}}{\operatorname{avg}} D^\mathrm{sym}(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}). \tag{S29}$$

The scale D_{o} is characteristic of the differences between the probability densities in our training data set, which we attribute to differences in sequence composition. It is therefore reasonable to accept modeling errors in our reconstruction rules, as measured by divergences, that are small compared to this scale, as we then expect to still be able to resolve sequence variation within our model. Figure S5 is a distribution, or histogram, of $D^{\mathrm{sym}}(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}})$ over all distinct pairs of 18-mers S_μ in the training set detailed in Table SI. A direct computation of the average of this distribution gives the value $D_{\mathrm{o}} \doteq 85$.

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

Table SI: Sequences S_μ contained in the MD data set. For the reasons described in the main text, the last three sequences were dropped from the training set.

Supplement to Section IV.B: Filtering data on disrupted hydrogen bonds

In some of our simulations, especially for oligomers ending with AT basepairs, intra-basepair hydrogen bonds at the oligomer ends were broken and the basepairs were open for a significant portion of the simulation time. Since such open basepairs are outside the scope of our

quadratic model, we decided not to use any snapshot with a broken hydrogen bond in our training data set. Following previous work [S11, S12], we considered a hydrogen bond to be broken if the distance between donor and acceptor was greater than 4 Å. To motivate and justify this choice, we plotted histograms of distances between atoms connected by intra-basepair hydrogen bonds in each simulation. One example is provided in Figure S6, with analogous plots for each intra-basepair hydrogen bond within each oligomer in our training data set online at `http://lcvmwww.epfl.ch/cgDNA`. One can notice that the distributions of the distances between pairs of atoms forming a hydrogen bond are close to Gaussians centered around 3 Å and their standard deviation is around 0.1-0.2 Å for most of the oligomers. Therefore setting a threshold for filtering the outliers at 4 Å (or around 5 standard deviations away from the mean) gives robust statistics in the remaining data, and structures that are significantly outside the scope of our quadratic model are explicitly excluded.

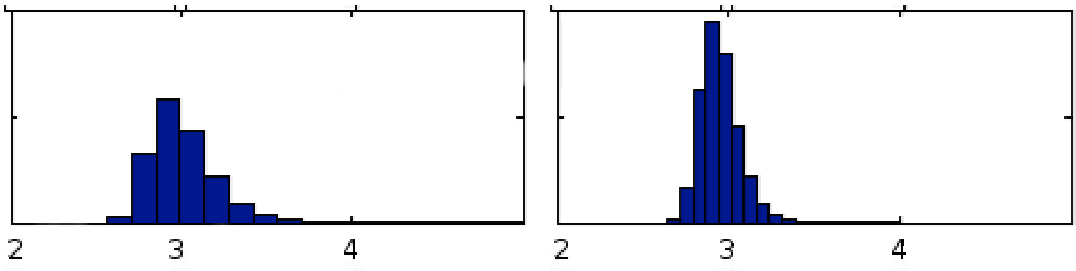


Figure S6: Histograms of the two hydrogen bond lengths at sequence position $\mathtt{X}_7 = \mathtt{T}$ during the MD simulation of the sequence S_1 of the training data set.

Supplement to Section V.B: Analysis of the least-squares system

To generate an initial approximation for a maximal relative entropy best-fit parameter set, we seek to construct a least-squares solution to the over-determined system of linear equations

$$\left.\begin{array}{rcl} \mathsf{K}_{\mu,\mathrm{m}} & = & \mathsf{K}^*_{\mu,\mathrm{M}} \\ \sigma_{\mu,\mathrm{m}} & = & \sigma^*_{\mu,\mathrm{M}} \end{array}\right\}, \quad \mu = 1, \ldots, 53. \tag{S30}$$

It is the unknown parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ that is to be estimated from this system. As detailed in equations (32)–(33) of the main article, the matrices $\mathsf{K}_{\mu,\mathrm{m}}$ on the left-hand side of the above equation are explicitly-known, linear functions of only the stiffnesses in the parameter set, while the vectors $\sigma_{\mu,\mathrm{m}}$ are explicitly-known, linear functions of only the weighted shape parameters. The matrices $\mathsf{K}^*_{\mu,\mathrm{M}}$ on the right-hand side of the above equation are observed data from the training set, as determined in the initial prescribed-sparsity, oligomer-based fit. Similarly, the vectors $\sigma^*_{\mu,\mathrm{M}} = \mathsf{K}^*_{\mu,\mathrm{M}}\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$ on the right-hand side are most simply defined as a matrix-vector product computed between the oligomer-based quantities fit to observed data. However, due to the decoupled nature of the system, we observe that there is some freedom in how to compute a least-squares solution. Specifically, we found that the following approach gave an initial approximation of the parameter set that yielded noticeably better reconstructions. The system $(\mathrm{S30})_1$ can first be solved in the least-squares sense for only the stiffness parameters, which yields specific values

$\mathsf{K}^{\#}_{\mu,\mathrm{m}}$ for the functions $\mathsf{K}_{\mu,\mathrm{m}}$. Then in the remaining least-squares system $(\mathrm{S30})_2$ for the weighted shape parameters, we used the right-hand side data $\sigma^{*}_{\mu,\mathrm{M}} := \mathsf{K}^{\#}_{\mu,\mathrm{m}}\widehat{\mathsf{w}}^{*}_{\mu,\mathrm{M}}$.

Several considerations arise in the least-squares treatment of (S30). Further details of the associated normal equations and their explicit solution can be found in §7 of [S16]. First, in view of (32)–(33), we see that each non-zero 6×6 block of each matrix $\mathsf{K}_{\mu,\mathrm{m}}$ depends on either one, two or three of the parameters $\{\mathsf{K}^{\alpha}_{1}, \mathsf{K}^{\alpha\beta}_{2}\}$, and each 6×1 block of each vector $\sigma_{\mu,\mathrm{m}}$ depends on either one, two or three of the parameters $\{\sigma^{\alpha}_{1}, \sigma^{\alpha\beta}_{2}\}$. Moreover, in view of (S30), we see that the equations for the entries of each such block are decoupled and of a similar form. Hence the equations in (S30) can be reduced to a collection of sparse, independent equations for each entry of the unknown parameter matrices $\{\mathsf{K}^{\alpha}_{1}, \mathsf{K}^{\alpha\beta}_{2}\}$ and the unknown parameter vectors $\{\sigma^{\alpha}_{1}, \sigma^{\alpha\beta}_{2}\}$. This entry-by-entry decoupling greatly reduces the computational effort required to solve the system. Second, in seeking a least-squares solution of (S30), we are free to consider only a subset of all blocks that appear, or to assign different weights to different blocks, reflecting differences in either importance of the fit or confidence in the data. For both reasons, it is quite natural to first consider only the subset of blocks associated with the interior of each oligomer and thereby ignore the data at the ends. If all blocks associated with ends are ignored, then the resulting interior system has a simple, explicit least-squares solution involving table averages over all instances of dimer and trimer sequences. However, due to the overlapping structure illustrated in (34)–(35), the least-squares solution of this interior system is not unique; in fact, the associated normal equations have a rather high-dimensional nullspace. On the other hand, if the blocks associated with ends are included, then the resulting system can be expected to have a unique least-squares solution provided that the training data set contains all possible dimer ends. This was why the original ABC set of oligomers was extended as described in Section IV.B. In the extended set of oligomers, considering both the reference and complementary strands, there are many instances of each of the $5'$-GC and GC-$3'$ ends, but only one instance for most of the other $5'$-dimer and dimer-$3'$ ends. Hence the different dimer ends are not equally represented in our training set, and the ability to control the weighting of the least-squares system at the ends is convenient.

In view of the considerations above, we adopted the following approach in our least-squares treatment of (S30). We assigned a unit weight to all interior 6×6 and 6×1 blocks, and a small, variable end-weight to all blocks associated with the leading and trailing ends of each oligomer. Since all dimer ends are included in the training set, we can obtain a unique least-squares solution for any positive end-weight. We then consider the limit in which the end-weight vanishes, and choose this as our least-squares solution. This choice is justified by the fact that all possible dimer ends are not equally represented in our training set and hence it is desirable to attempt to minimize any biases at the ends. Moreover, the vanishing-end-weight solution is rather simple and is available in closed-form in terms of table averages; it merely selects a unique element in the nullspace of the interior system in which all blocks associated with the ends are ignored. While other choices of a least-squares solution could be made, we prefer the rationale for the choice described here, although it is not crucial.

The procedure described above takes no account of the admissibility constraint on the stiffness parameter matrices $\{\mathsf{K}^{\alpha}_{1}, \mathsf{K}^{\alpha\beta}_{2}\}$ described in Section III.B.2. Specifically, it appears natural to require that each of these parameter matrices should be at least semi-positive-definite, and that the constructed oligomer matrices for any ten independent sequences of length two (physical dimers) should be positive-definite. The vanishing-end-weight, least-squares procedure described above gave an inadmissible parameter set according to these criteria: some of the stiffness parameter matrices had some negative eigenvalues, although they did give reasonable constructions of some oligomers away from the ends. Curiously, we considered a variety of intuitive choices for selecting a particular least-squares solution within the null-space of all such solutions, but found

that all choices were inadmissible due to the presence of some negative eigenvalues. For this reason, we developed a numerical procedure to explore the high-dimensional nullspace associated with the interior system described above and search for an admissible least-squares solution. Working with $(S30)_1$ and its vanishing-end-weight solution, the procedure incrementally adjusted the free variables in the corresponding nullspace so as to increase the negative eigenvalues in the stiffness matrices. In this way, we obtained an admissible least-squares solution of $(S30)_1$ in which the stiffness matrices were all at least semi-positive-definite. Actually, the matrices were all positive-definite, but some had some extremely small eigenvalues. We then considered $(S30)_2$ and its vanishing-end-weight solution, and adjusted the free variables in the corresponding nullspace so as to obtain a least-squares solution of $(S30)_2$ in which the weighted shape parameter vectors were orthogonal to the eigenspaces of the small eigenvalues of the associated stiffness parameter matrices. This orthogonality condition was imposed to avoid potential ill-conditioning problems associated with these eigenvalues.

Our least-squares treatment of the linear system in (S30) thus provided a rational, admissible parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ to be used as an initial guess in our numerical minimization of the nonlinear, Kullback-Leibler objective functional defined in (43) of the main article.

Supplement to Section V.C: Properties of the $\mathcal{P}^*$ parameter set

Visualization of the parameter set

The data in Figures S7, S8 and S9 provide a visual illustration of the entire best-fit parameter set $\mathcal{P}^* = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$. Figure S7 provides color plots related to the 1-mer stiffness parameter matrices $\mathsf{K}_1^\alpha \in \mathbb{R}^{6\times 6}$ and weighted shape parameter vectors $\sigma_1^\alpha \in \mathbb{R}^6$. For the stiffness parameters, we plot the Euclidean average $\mathsf{K}_1^{\text{avg}}$ and the standard deviation $\mathsf{K}_1^{\text{dev}}$ over all 4 possible monomers $\alpha = \mathtt{T}, \mathtt{A}, \mathtt{C}$ and $\mathtt{G}$, along with the differences $\mathsf{K}_1^{\Delta\alpha} = \mathsf{K}_1^\alpha - \mathsf{K}_1^{\text{avg}}$ for the 2 independent monomers $\alpha = \mathtt{A}$ and $\mathtt{G}$. Analogous plots are also presented for the weighted shape parameters. As can be seen, there are marked sequence-dependent variations among the stiffness and the weighted shape parameters, which suggests that they are successfully capturing differences in the intra-basepair interactions within the 2 independent monomers shown here, and hence by objectivity between all 4 possible monomers.

Analogous plots are made for the 2-mer stiffness parameter matrices $\mathsf{K}_2^{\alpha\beta} \in \mathbb{R}^{18\times 18}$ in Figure S8, and for the 2-mer weighted shape parameter vectors $\sigma_2^{\alpha\beta} \in \mathbb{R}^{18}$ in Figure S9. For the stiffness parameters, we plot the Euclidean average $\mathsf{K}_2^{\text{avg}}$ and the standard deviation $\mathsf{K}_2^{\text{dev}}$ over all 16 possible dimers $\alpha\beta$, along with in column one the differences from the average $\mathsf{K}_2^{\Delta\alpha\beta} = \mathsf{K}_2^{\alpha\beta} - \mathsf{K}_2^{\text{avg}}$ for the 3 independent purine-pyrimidine dimers $\alpha\beta = \mathtt{AT}, \mathtt{GC}$ and $\mathtt{GT}$, and analogously in column two for the 3 independent pyrimidine-purine dimers $\alpha\beta = \mathtt{TA}, \mathtt{CG}$ and $\mathtt{TG}$, and in column three for the 4 independent purine-purine dimers $\alpha\beta = \mathtt{AA}, \mathtt{GG}, \mathtt{AG}$ and $\mathtt{GA}$, with the same groupings of results for the weighted shape parameters in Figure S9. While there is some sequence-dependent variation within the columns for both the stiffness and the weighted shape parameters, as there should be, there are more striking patterns that are common within, and distinct between, each column. As before, this observation suggests that the parameter set is successfully capturing differences in the inter-basepair or stacking interactions within the 10 independent dimers shown here, and hence by objectivity all 16 possible dimers.

Eigenvalues of the parameter set stiffness matrices

Figure S10 shows the logarithm of the eigenvalues of a) the 1-mer stiffness parameter matrices K_1^α for 2 independent monomers α, b) the 2-mer stiffness parameter matrices $\mathsf{K}_2^{\alpha\beta}$ for 10 independent dimers $\alpha\beta$, and c) the constructed oligomer stiffness matrices $\mathsf{K}_{\mu,\mathrm{m}}^*$ for the same 10 independent oligomers S_μ of length two (i.e. physical dimers); these short oligomers are not part of the training data set, but we retain the notation for convenience. While all the eigenvalues, denoted by the black symbols, of the stiffness parameter matrices K_1^α and $\mathsf{K}_2^{\alpha\beta}$ are positive, some are extremely small, of the order 10^{-6} or less. Consequently, some of the stiffness parameter matrices could be reasonably approximated by lower-rank matrices with some eigenvalues set exactly equal to zero. For instance, both 1-mer stiffness parameter matrices K_1^α could be approximated by lower-rank matrices with 2, 3 or 4 eigenvalues equal to zero, and some of the 2-mer stiffness parameter matrices $\mathsf{K}_2^{\alpha\beta}$ could be approximated by lower-rank matrices with 1 or 2 eigenvalues equal to zero. We remark that the matrices and hence the eigenvalues presented here are expressed in dimensionless units according to the scales introduced in Section II.D. Precisely the same magnitudes and conclusions would be obtained in dimensional units when lengths are expressed in units of Å, angles are expressed in units of $1/5$-radians (approximately 11-degrees), and energies are expressed in units of k_BT. Specifically, the eigenvalues of small magnitudes shown here are not artifacts of the non-dimensionalization, but are characteristic properties of the 1-mer and 2-mer interaction energy models at these scales. When the parameter matrices K_1^α and $\mathsf{K}_2^{\alpha\beta}$ are combined, as illustrated in (34), to form the model stiffness matrices $\mathsf{K}_{\mu,\mathrm{m}}^*$ for the oligomers of length two described above, the resulting matrices $\mathsf{K}_{\mu,\mathrm{m}}^*$, without exception, have all eigenvalues greater than 10^{-1} or so; these eigenvalues are denoted by the red symbols. Hence the individual 1-mer and 2-mer interaction energies, with rather soft modes, stabilize each other when superposed to yield a dimer (or length two oligomer) energy that is appropriately stiff.

End effects in the reconstruction of a homogeneous, sequence-averaged oligomer

As discussed in the main article, it is of interest to consider the sequence-averaged, best-fit parameter set $\mathcal{P}^{*,\mathrm{avg}} = \{\sigma_1^{\mathrm{avg}}, \mathsf{K}_1^{\mathrm{avg}}, \sigma_2^{\mathrm{avg}}, \mathsf{K}_2^{\mathrm{avg}}\}$, where σ_1^{avg}, $\mathsf{K}_1^{\mathrm{avg}}$ and so on denote the Euclidean averages illustrated in Figures S7, S8 and S9. The set $\mathcal{P}^{*,\mathrm{avg}}$ can be interpreted as providing a homogeneous, nearest-neighbor model of DNA in which the occurrence of each of the four possible basepairs is assumed to be equally likely at each position in an oligomer. Using this parameter set, a model shape vector $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$ and stiffness matrix $\mathsf{K}_{\mathrm{h,m}}^*$ can then be constructed for a homogeneous oligomer of arbitrary length. Figure S11 below shows entries of the constructed shape vector $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$ and stiffness matrix $\mathsf{K}_{\mathrm{h,m}}^*$ as a function of position along a homogeneous, 18-basepair oligomer. The top four panels of the figure show the different entries of the shape vector $\widehat{\mathsf{w}}_{\mathrm{h,m}}^*$ versus oligomer position, with discrete point values visualized using linear interpolation. Each of the four panels contains plots of three of the twelve types of parameters, grouped by intra- and inter-basepair types, and by translational and rotational types; the numerical scale on the ordinate is different in each panel to suit the pertinent data. Despite the fact that the oligomer is homogeneous, with a uniform parameter set, significant end effects are visible: the constant value of each parameter in the interior of the oligomer is only approached sufficiently far from the ends. For some parameters, for example Propeller, the end effects are visible on the scale of the plot to a depth of penetration of 4 basepairs. Such nonlocal end effects are typical of all the shape parameters: the magnitude and depth are similar for both intra- and inter-basepair parameters, but are more evident in the former because of differences in the scales between the panels. Even with end effects, the required palindromic symmetry of the homogeneous model is evident, with oddness of

Buckle, Shear, Tilt and Shift (all plotted in black), and evenness of the remaining parameters.

The bottom four panels of Figure S11 are analogous and show the different diagonal entries of the stiffness matrix $\mathsf{K}^*_{\mathrm{h,m}}$. Although the stiffness matrix has many non-zero entries, we choose to plot only the diagonal entries for brevity. In contrast to the nonlocal end effects in the shape vector $\widehat{\mathsf{w}}^*_{\mathrm{h,m}}$, it is a consequence of our nearest-neighbor model that the end effects in the stiffness matrix $\mathsf{K}^*_{\mathrm{h,m}}$ are localized precisely to the first and last 6×6, intra-basepair blocks as reflected in the left two panels, while the inter-basepair stiffnesses do not change at all as reflected in the right two panels. Notice that the localized end effects in the intra-basepair stiffnesses are significant: some parameters change by approximately 50% at the oligomer ends. The palindromic symmetry of the homogeneous model is again evident: palindromy implies that all the diagonal stiffness parameters should be even functions of position about the middle of the oligomer, as is visible in each panel.

Interior values of shape parameters in homogeneous, sequence-averaged oligomers

Figure S11 illustrates that, sufficiently far from the ends, the shape parameters of a homogeneous oligomer approach the constant values $\widehat{\mathsf{w}}^*_{\mathrm{h,m}}$ that are reported in Table II of the main text in both *Curves+* and *3DNA* coordinates. The computation of a *3DNA* version of our homogeneous shapes $\widehat{\mathsf{w}}^*_{\mathrm{h,m}}$ is not entirely straightforward, and involves various choices. We adopted the following procedure: we first reconstructed absolute coordinates (reference points and frames) of a sequence-averaged rigid-base DNA configuration, using the parameter set $\mathcal{P}^{*,\mathrm{avg}}$. Then for this coarse-grain configuration, we reconstructed 5 sets of absolute coordinates of all the non-hydrogen atoms in an idealized base using the sequences S_μ of Table SI for $\mu = 1,\, 3,\, 5,\, 9,\, 17$ and the *Curves+* base embedding rules [S10]. We then ran the program *3DNA* [S12] with these 5 sets of reconstructed atomic coordinates as inputs, and averaged the corresponding intra- and inter-basepair *3DNA* coarse-grain parameter outputs over all basepairs and junctions of all the five sequences, computed along both strands while staying three basepairs away from the ends.

Supplement to Section VI: Further comparisons of constructed models

Here we extend the discussion of the main article and compare results for the four training set oligomers S_μ with $\mu = 1, 3, 8, 42$ as detailed in Table SI. Although oligomers S_1 and S_8 are also discussed in the main article, here we provide additional information on these oligomers. Regarding sequence composition, we note that S_1 is a palindromic 18-mer with a period two sequence in its interior, S_3 is a palindromic 18-mer with a period four sequence in its interior, S_8 is a non-palindromic 18-mer also with a period four sequence in its interior, and S_{42} is a non-palindromic 12-mer with a non-periodic sequence in its interior. The three oligomers S_1, S_3 and S_8 all have $5'$-GC and GC-$3'$ dimer ends, whereas S_{42} has $5'$-GG and GG-$3'$ dimer ends.

Divergences between different oligomers

Table SII shows various relative divergences between the observed internal configuration density $\rho_{\mu,\mathrm{o}}$, the oligomer-based model density $\rho^*_{\mu,\mathrm{M}}$, and the constructed dimer-based model density $\rho^*_{\mu,\mathrm{m}}$ for the 18-mers S_μ, $\mu = 1, 3, 8, 1'$. Here the oligomer $\mathsf{S}_{1'}$ has a single point mutation from the sequence of S_1 as discussed in Section VII. With the Kullback-Leibler scale D_o for 18-mers introduced in Section IV.E, the diagonal cells of the table report the relative divergence $D(\rho^*_{\mu,\mathrm{M}},\, \rho_{\mu,\mathrm{o}})/D_\mathrm{o}$ between the observed and oligomer-based model densities of oligomer S_μ in the top entry of the

	S_1	S_3	S_8	$\mathsf{S}_{1'}$
S_1	**0.087**	0.676	0.567	0.116
	0.076	0.619	0.407	0.053
S_3	0.616	**0.095**	0.541	0.621
	0.489	**0.075**	0.297	0.466
S_8	0.423	0.422	**0.091**	0.418
	0.310	0.314	**0.088**	0.286
$\mathsf{S}_{1'}$	0.139	0.679	0.588	**0.089**
	0.040	0.574	0.364	**0.082**

Table SII: Relative pairwise Kullback-Leibler divergences between the internal configuration densities $\rho_{\mu,\mathrm{o}}$, $\rho^*_{\mu,\mathrm{M}}$, and $\rho^*_{\mu,\mathrm{m}}$ for the 18-mers S_μ, $\mu = 1, 3, 8, 1'$, where $\mathsf{S}_{1'}$ is a single point mutation of S_1. Diagonal cells top: $D(\rho^*_{\mu,\mathrm{M}}, \rho_{\mu,\mathrm{o}})/D_\mathrm{o}$ for oligomer S_μ. Diagonal cells bottom: $D(\rho^*_{\mu,\mathrm{m}}, \rho^*_{\mu,\mathrm{M}})/D_\mathrm{o}$ for oligomer S_μ. Off-diagonal cells top: $D(\rho_{\mu,\mathrm{o}}, \rho_{\nu,\mathrm{o}})/D_\mathrm{o}$ for distinct oligomers S_μ and S_ν. Off-diagonal cells bottom: $D(\rho^*_{\mu,\mathrm{m}}, \rho^*_{\nu,\mathrm{m}})/D_\mathrm{o}$ for distinct oligomers S_μ and S_ν.

cell, and the relative divergence $D(\rho^*_{\mu,\mathrm{m}}, \rho^*_{\mu,\mathrm{M}})/D_\mathrm{o}$ between the oligomer-based and dimer-based model densities of S_μ in the bottom entry. The off-diagonal cells of the table show the relative divergence $D(\rho_{\mu,\mathrm{o}}, \rho_{\nu,\mathrm{o}})/D_\mathrm{o}$ between the observed densities of the distinct oligomers S_μ and S_ν in the top entry of the cell, and the relative divergence $D(\rho^*_{\mu,\mathrm{m}}, \rho^*_{\nu,\mathrm{m}})/D_\mathrm{o}$ in the dimer-based model densities of S_μ and S_ν in the bottom entry. In the diagonal cells, the fact that each entry is less than 10% indicates that the error incurred at each stage of modeling, from the observed to the oligomer-based to the dimer-based model, is less than 10% for each of the four oligomers S_μ. In the off-diagonal cells, the top entries quantify differences in the observed densities due to differences in sequence, whereas the bottom entries quantify the same sequence dependence but for the dimer-based model densities. With the exception of the outermost off-diagonal cells, we see that the bottom entries are of the same order as the top, which indicates that the dimer-based model is reasonably capturing the variation in density due to the variation in sequence. In the outermost off-diagonal cells, the oligomers S_1 and $\mathsf{S}_{1'}$ differ by just a single point mutation. While the dimer-based model can still capture the variation in density in this case, it gives a variation that is noticeably smaller than observed. Thus the dimer-based model can resolve differences in the probability density on the 210-dimensional internal configuration space of an 18-mer due to differences in sequence, even when the difference in sequence is in a single basepair.

Quality of shape and stiffness reconstructions

Figures S12, S13, S14 and S15 show entries of the shape vector and stiffness matrix as a function of position for the four training set oligomers S_1, S_3, S_8 and S_{42}. The figures illustrate pointwise comparisons between the observed and constructed dimer-based model parameters along the different oligomers. Figures S12 and S14 are identical to Figures 4 and 5, and are repeated here for convenience. The data in Figures S12–S15 illustrate the quality of the dimer-based model constructions. As noted in the main article, the differences between the observed and the constructed quantities are rather small, and with very few exceptions, the pointwise differences in the quantities are less than the variation with sequence. There is a tendency for the constructed quantities to exhibit larger errors at the ends, particularly for the oligomer S_{42}, which may indicate a lack of sampling of GG dimer ends in the training data set. Visually, the errors in the intra-basepair shape

and stiffness parameters appear larger, but the scales in the plots of the intra- and inter-basepair parameters are different. For both intra- and inter-basepair shape parameters, and away from the ends, rather few errors are larger than 0.1Å in translational variables and 2° in rotational variables. All constructed parameters shown for oligomers S_1, S_3 and S_8 are clearly consistent with the periodicity of their interior sequences. By design, the constructed parameters exactly satisfy the requisite symmetry conditions for the palindromic oligomers S_1 and S_3. The observed parameters computed directly from the MD time series data, and shown in the solid lines, for the most part also closely satisfy the requisite symmetries, but errors can arise from a lack of convergence of the MD simulation of the relevant oligomer. For example, the breaking of evenness in the plot of the observed shape parameter Stagger and the observed stiffness parameter Twist-Twist in Figure S12 violates the palindromic symmetry of oligomer S_1, and must reflect a lack of convergence of the MD time series.

Comparison of marginals

Figures S16, S17, S18, S19 and S20 show various one-dimensional marginal distributions (or histograms) for each type of intra- and inter-basepair coordinate at each location along the four training set oligomers S_1, S_3, S_8 and S_{42}. These marginal distributions provide a way to assess the quality of the Gaussian assumption in our modeling approach and further illustrate sequence and end effects. For each type of coordinate, at each location, along each oligomer S_μ, we compare four different marginal distributions as described in the main article. Figures S16, S17 and S19 are identical to Figures 6, 7 and 8 and are repeated here for convenience. Figure S16 shows the marginal distributions for the intra-basepair coordinates along oligomer S_8. As noted in the main article, the four distributions for each coordinate at each position are practically indistinguishable. Similar results hold for the distributions of intra-basepair coordinates along oligomers S_1, S_3 and S_{42}; the four distributions at each position on each of these oligomers are even closer than those for oligomer S_8. Figures S17–S20 provide analogous plots for the inter-basepair coordinates. Now it can be seen that there are cases where the actual marginal distribution obtained from the MD data is noticeably non-Gaussian. For example, see the marginals of Slide for the various `TA` dimers in Figure S17, the marginals of Twist for the various `CG` dimers in Figures S18, S19 and S20, and the marginals of Slide, Shift and Twist for the various `GG` dimers in Figure S20. The marginal of Slide in the $5'$-`GG` dimer end in Figure S20 is particularly far from Gaussian. While such bi-modal and otherwise non-Gaussian behavior is beyond the scope of the Gaussian approach considered here, the results show that the dimer-based model with the best-fit parameter set can capture the dominant features of sequence variation in a satisfactory way. Specifically, when comparing the dimer-based model construction to the MD data, the error in the mean and width of the constructed marginal of any coordinate is almost always qualitatively smaller than the variation in these quantities due to sequence.

Further comparisons

Plots analogous to those presented in Figures S12–S15, and Figures S16–S20, but for all the oligomers in our training data set are available online at `http://lcvmwww.epfl.ch/cgDNA`.

References

[S1] M.S. Babcock, E. P. D. Pednault and W. K. Olson, Nucleic Acid Structure Analysis. Mathematics for Local Cartesian and Helical Structure Parameters That Are Truly Comparable Between

Structures, *J. Mol. Biol* **237** (1994) 125-156.

[S2] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Second Edition, Springer, New York (2002).

[S3] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York (1991).

[S4] P. Hughes, *Spacecraft Attitude Dynamics*, Wiley, Boston (1983).

[S5] E.T. Jaynes, Information Theory and Statistical Mechanics, *Physical Review* **106** (1957) 620–630.

[S6] E.T. Jaynes, Information Theory and Statistical Mechanics II, *Physical Review* **108** (1957) 171–190.

[S7] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, London (2003).

[S8] S. Kullback and R.A. Leibler, On Information and Sufficiency, *Annals of Mathematical Statistics* **22** (1951) 79-86.

[S9] S. Kullback, *Information Theory and Statistics*, Wiley, New York (1959).

[S10] R. Lavery, M. Moakher, J. Maddocks, D. Petkeviciute and K. Zakrzewska, Conformational analysis of nucleic acids revisited: *Curves+*. *Nucleic Acids Res. 37* (2009), 5917–5929.

[S11] F. Lankaš, 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. *Physical Chemistry Chemical Physics 11* (2009), 10565–10588.

[S12] X.-J. Lu and W. Olson, *3DNA*: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res. 31*, 17 (2003), 5108–5121.

[S13] A.J. Majda and X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*, Cambridge University Press (2006).

[S14] M. Moakher, On the Averaging of Symmetric Positive-Definite Tensors, *J. Elasticity* **82** (2006) 273-296.

[S15] 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), 229–237.

[S16] D. Petkeviciute, *A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulation*, PhD Thesis no **5520** EPFL (2012)

[S17] J. Walter, O. Gonzalez, J.H. Maddocks, On the stochastic modeling of rigid body systems with application to polymer dynamics, *SIAM Multiscale Modeling and Simulation* **8** (2010)

[S18] J. Walter, C. Hartmann, J.H. Maddocks , Ambient space formulations and statistical mechanics of holonomically constrained Langevin systems, *Eur. Phys. J. Special Topics* **200** (2011)

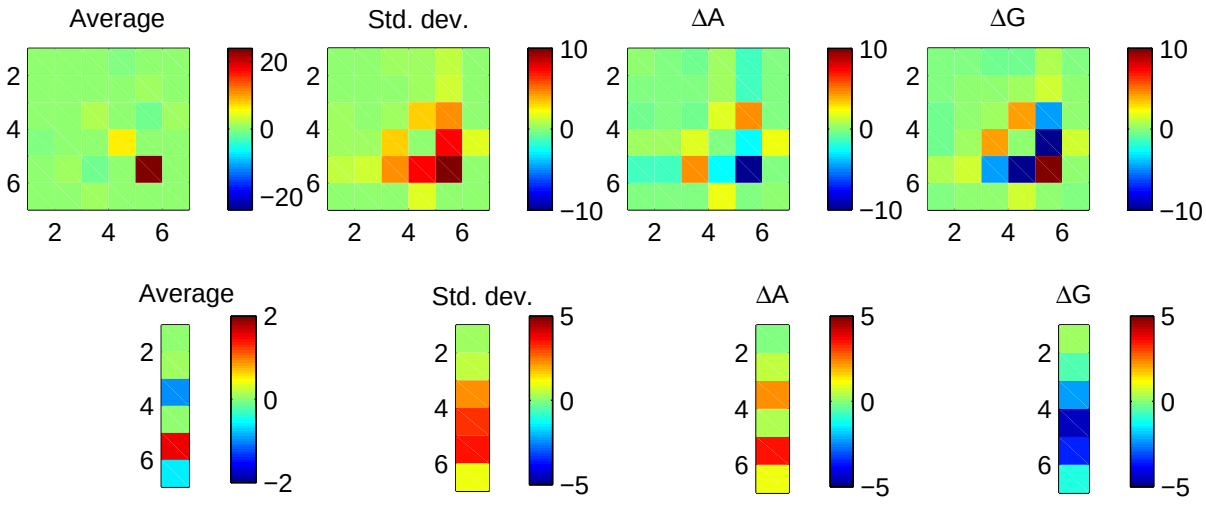


Figure S7: Averages, standard deviations and differences in dimensionless units of the 1-mer parameters K_1^α and σ_1^α of the best-fit parameter set $\mathcal{P}^*$ of the dimer-based model. The indices $1, \ldots, 6$ correspond to the variables Buckle, Propeller, Opening, Shear, Stretch, Stagger. Top row: plot of $\mathsf{K}_1^{\text{avg}}$, $\mathsf{K}_1^{\text{dev}}$, $\mathsf{K}_1^{\Delta\alpha} = \mathsf{K}_1^\alpha - \mathsf{K}_1^{\text{avg}}$ for the two independent monomers $\alpha = \mathtt{A}$ and $\mathtt{G}$. Bottom row: plot of σ_1^{avg}, σ_1^{dev}, $\sigma_1^{\Delta\alpha} = \sigma_1^\alpha - \sigma_1^{\text{avg}}$ for the two independent monomers as above.

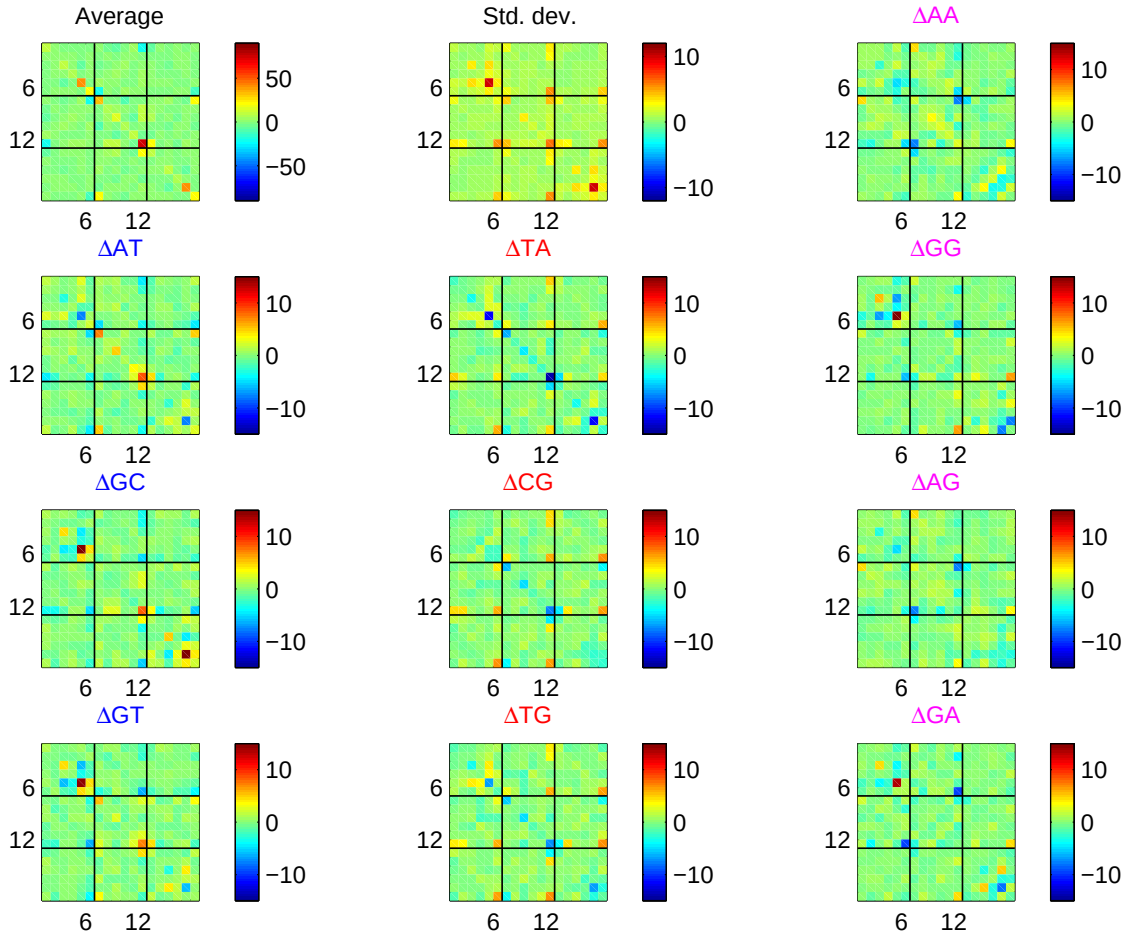


Figure S8: Averages, standard deviations and differences in dimensionless units of the 2-mer stiffness parameters $\mathsf{K}_2^{\alpha\beta}$ of the best-fit parameter set $\mathcal{P}^*$ of the dimer-based model. Each group of indices $1,\ldots,6$ and $13,\ldots,18$ corresponds to the variables Buckle, Propeller, Opening, Shear, Stretch, Stagger. The group of indices $7,\ldots,12$ corresponds to the variables Tilt, Roll, Twist, Shift, Slide, Rise. The plots are of $\mathsf{K}_2^{\text{avg}}$, $\mathsf{K}_2^{\text{dev}}$ and $\mathsf{K}_2^{\Delta\alpha\beta} = \mathsf{K}_2^{\text{avg}} - \mathsf{K}_2^{\alpha\beta}$ for the 3 independent purine-pyrimidine dimers (first column), the 3 independent pyrimidine-purine dimers (second column) and the 4 independent purine-purine dimers (third column).

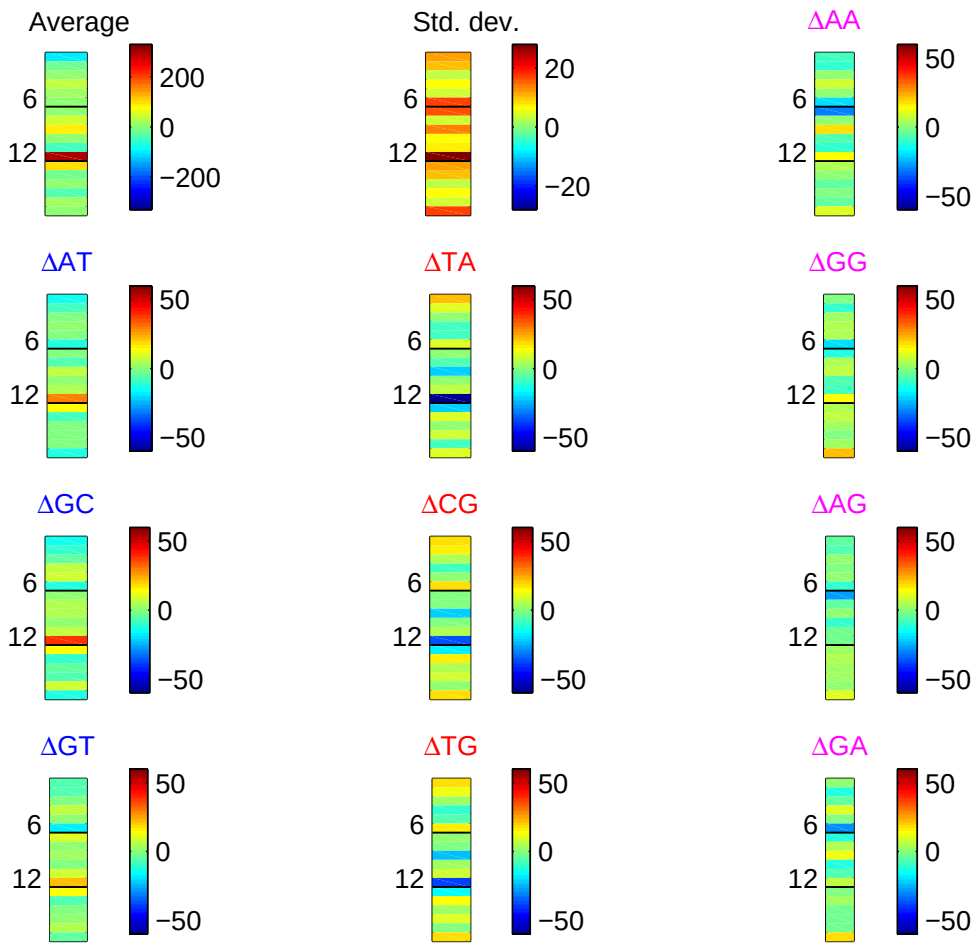


Figure S9: Averages, standard deviations and differences in dimensionless units of the 2-mer parameters $\sigma_2^{\alpha\beta}$ of the best-fit parameter set $\mathcal{P}^*$ of the dimer-based model. Each group of indices $1,\ldots,6$ and $13,\ldots,18$ corresponds to the variables Buckle, Propeller, Opening, Shear, Stretch, Stagger. The group of indices $7,\ldots,12$ corresponds to the variables Tilt, Roll, Twist, Shift, Slide, Rise. Plot are of σ_2^{avg}, σ_2^{dev} and $\sigma_2^{\Delta\alpha\beta} = \sigma_2^{\text{avg}} - \sigma_2^{\alpha\beta}$ for the 3 independent purine-pyrimidine dimers (first column), the 3 independent pyrimidine-purine dimers (second column) and the 4 independent purine-purine dimers (third column).

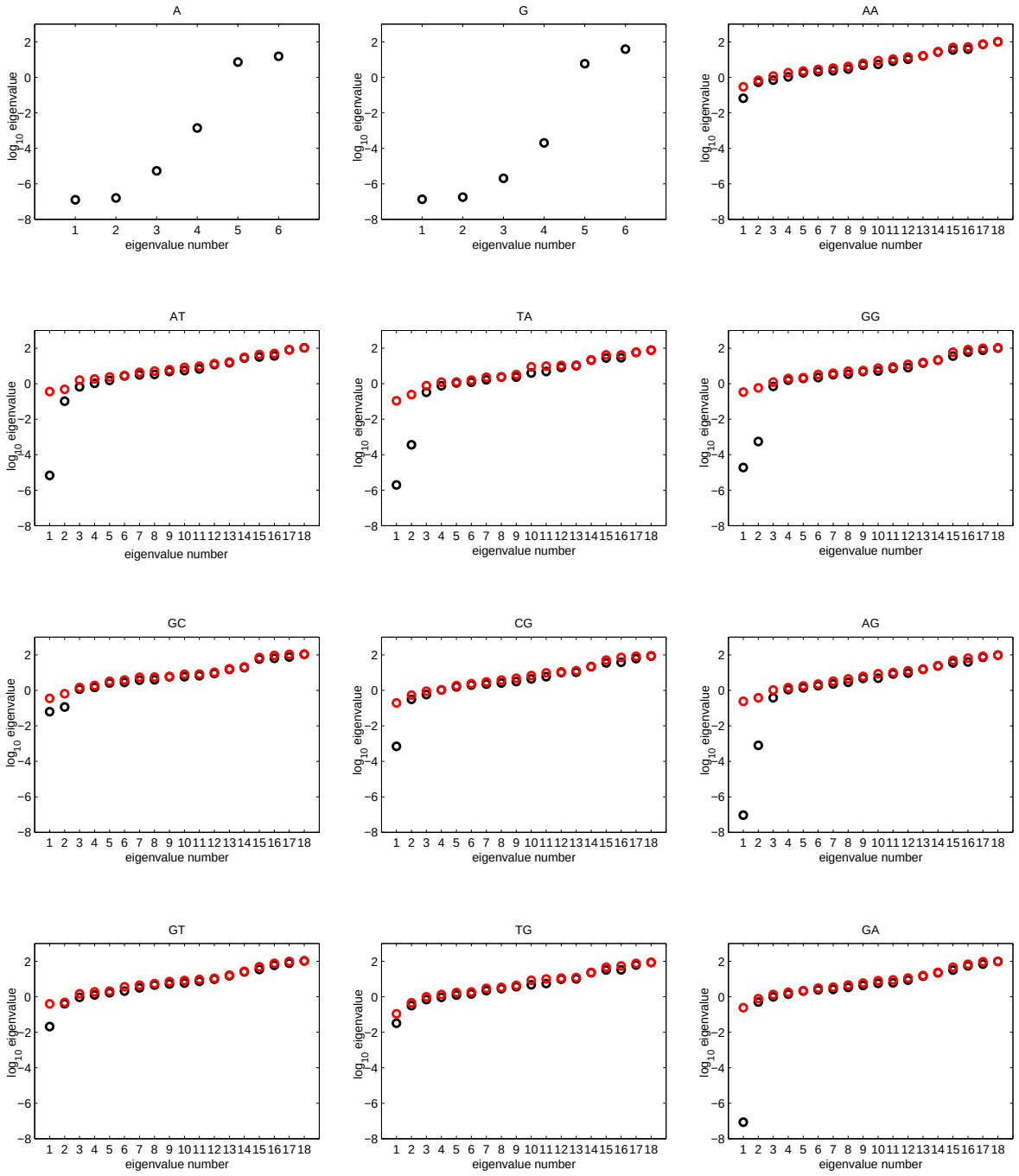


Figure S10: Logarithm of stiffness matrix eigenvalues in dimensionless units associated with the best-fit parameter set $\mathcal{P}^*$ of the dimer-based model. Black symbols: eigenvalues of the 1-mer parameter matrices K_1^{α} for 2 independent monomers α, and the 2-mer parameter matrices $\mathsf{K}_2^{\alpha\beta}$ for 10 independent dimers $\alpha\beta$. Red symbols: eigenvalues of the constructed oligomer stiffness matrices $\mathsf{K}^*_{\mu,\mathrm{m}}$ for the corresponding 10 independent oligomers of length two (physical dimers).

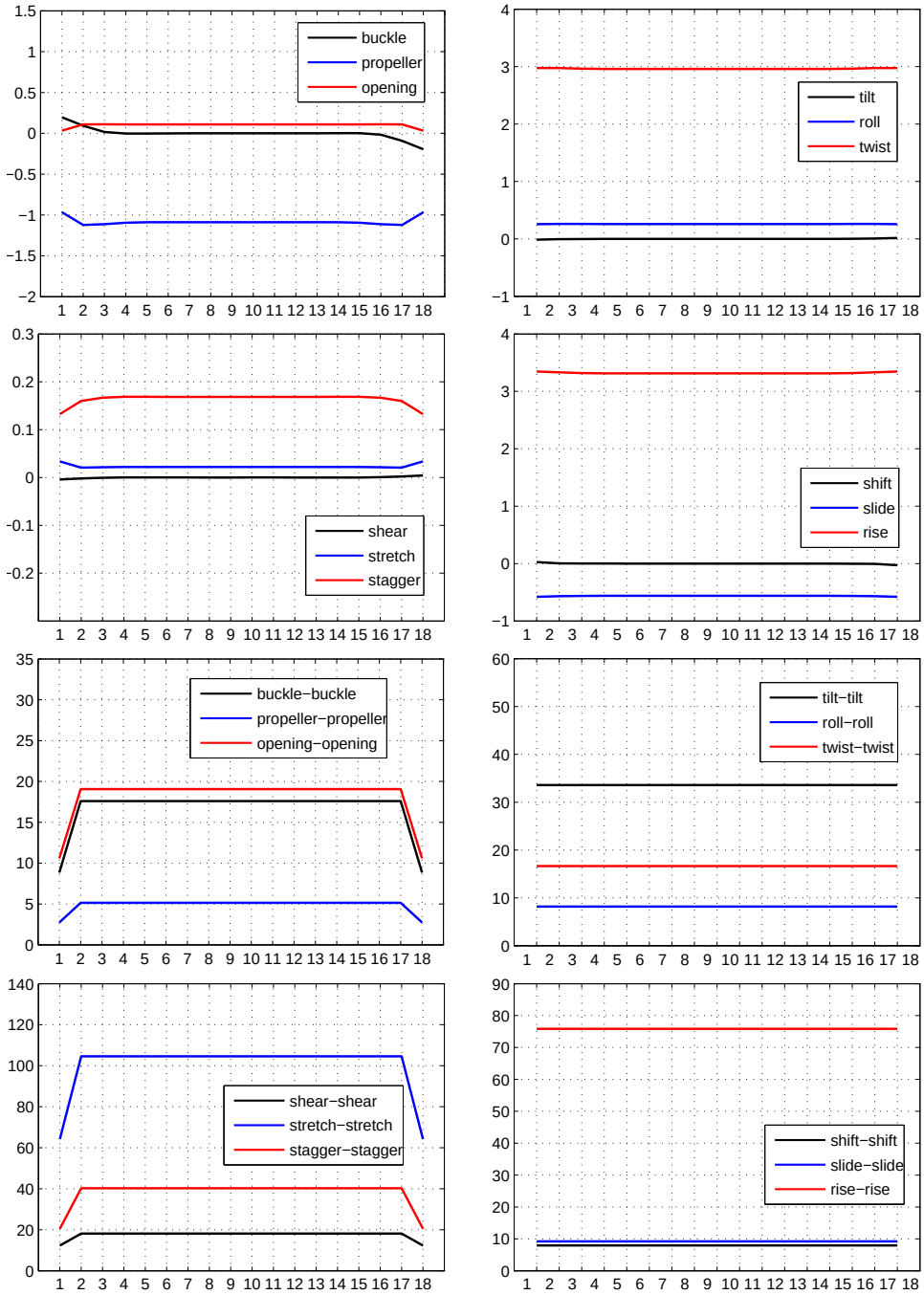


Figure S11: Entries of the shape vector $\widehat{w}^{*}_{h,m}$ and stiffness matrix $K^{*}_{h,m}$, constructed from the sequence-averaged, best-fit parameter set $\mathcal{P}^{*,avg}$, in dimensionless units as a function of position along a homogeneous, 18-basepair oligomer. Values at successive positions are joined by a piecewise linear curve. Top four panels: each of the twelve types of entries of $\widehat{w}^{*}_{h,m}$ versus position. Bottom four panels: each of the twelve types of diagonal entries of $K^{*}_{h,m}$ versus position. Both intra- and inter-basepair shape parameters exhibit nonlocal end effects, whereas the intra-basepair stiffnesses exhibit only local end effects, and the inter-basepair stiffnesses exhibit no end effects.

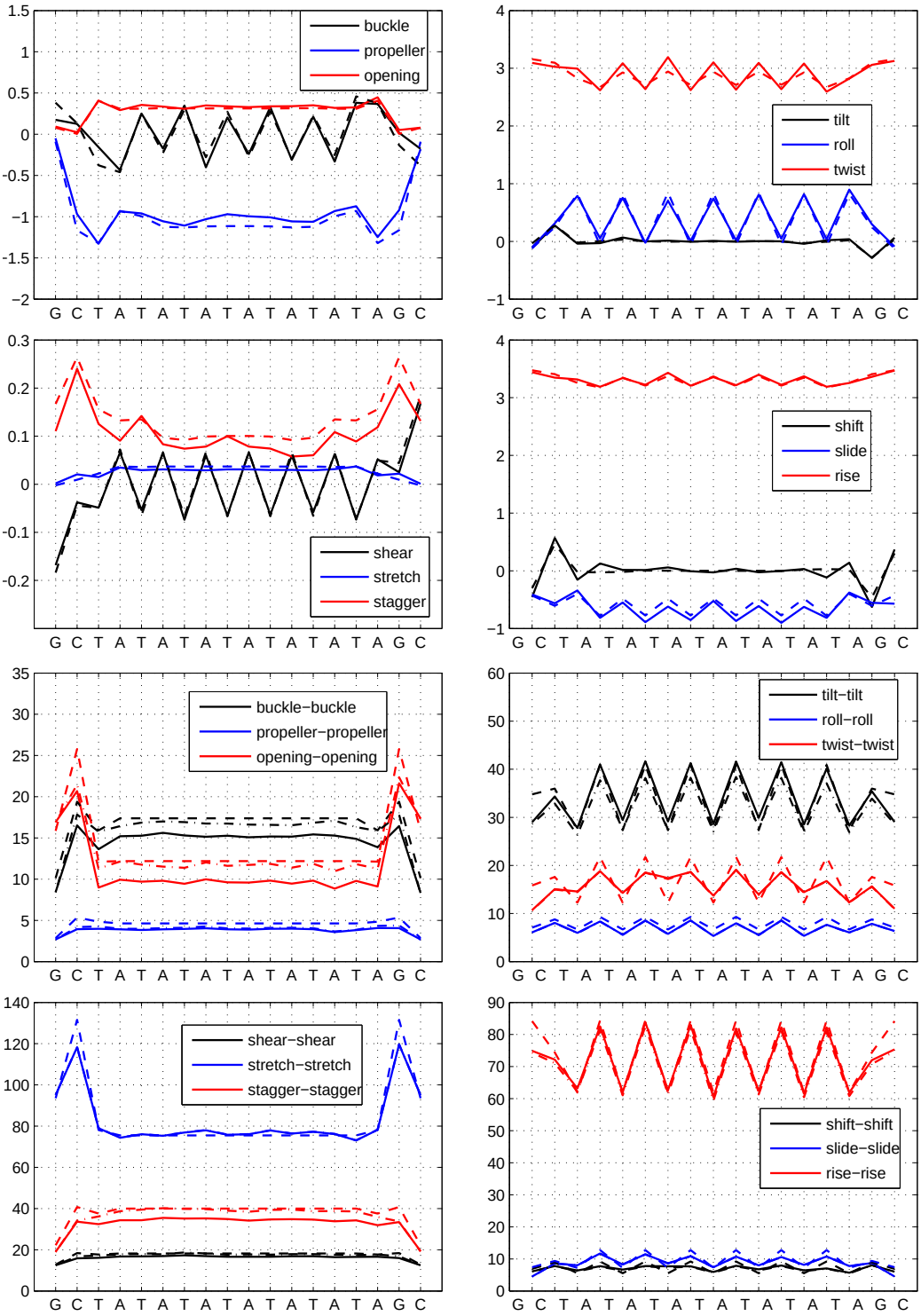


Figure S12: Entries of shape vectors and stiffness matrices in dimensionless units for the palindromic, interior period two, 18-mer S_1 from the training set. Top four panels: entries of observed vector $\widehat{w}_{1,o}$ (solid) and constructed dimer-based model vector $\widehat{w}^*_{1,m}$ (dashed). Bottom four panels: diagonal entries of observed matrix $K_{1,o}$ (solid), constructed dimer-based model matrix $K^*_{1,m}$ (dashed) and nearest-neighbor oligomer-based model matrix $K^*_{1,M}$ (dash-dot).

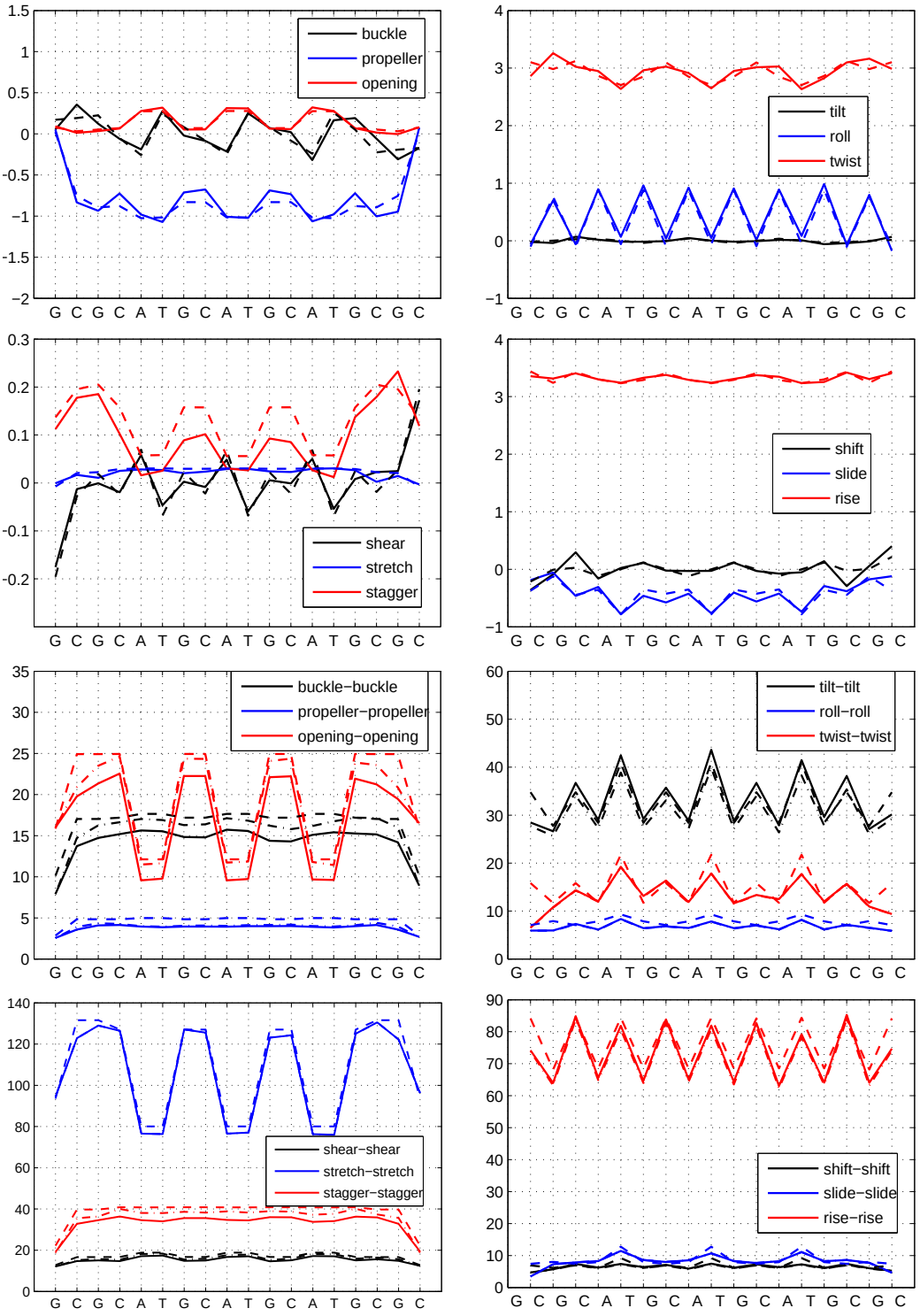


Figure S13: Entries of shape vectors and stiffness matrices in dimensionless units for the palindromic, interior period four, 18-mer S_3 from the training set. Top four panels: entries of observed vector $\widehat{\mathsf{w}}_{3,\mathrm{o}}$ (solid) and constructed dimer-based model vector $\widehat{\mathsf{w}}^*_{3,\mathrm{m}}$ (dashed). Bottom four panels: diagonal entries of observed matrix $\mathsf{K}_{3,\mathrm{o}}$ (solid), constructed dimer-based model matrix $\mathsf{K}^*_{3,\mathrm{m}}$ (dashed) and nearest-neighbor oligomer-based model matrix $\mathsf{K}^*_{3,\mathrm{M}}$ (dash-dot).

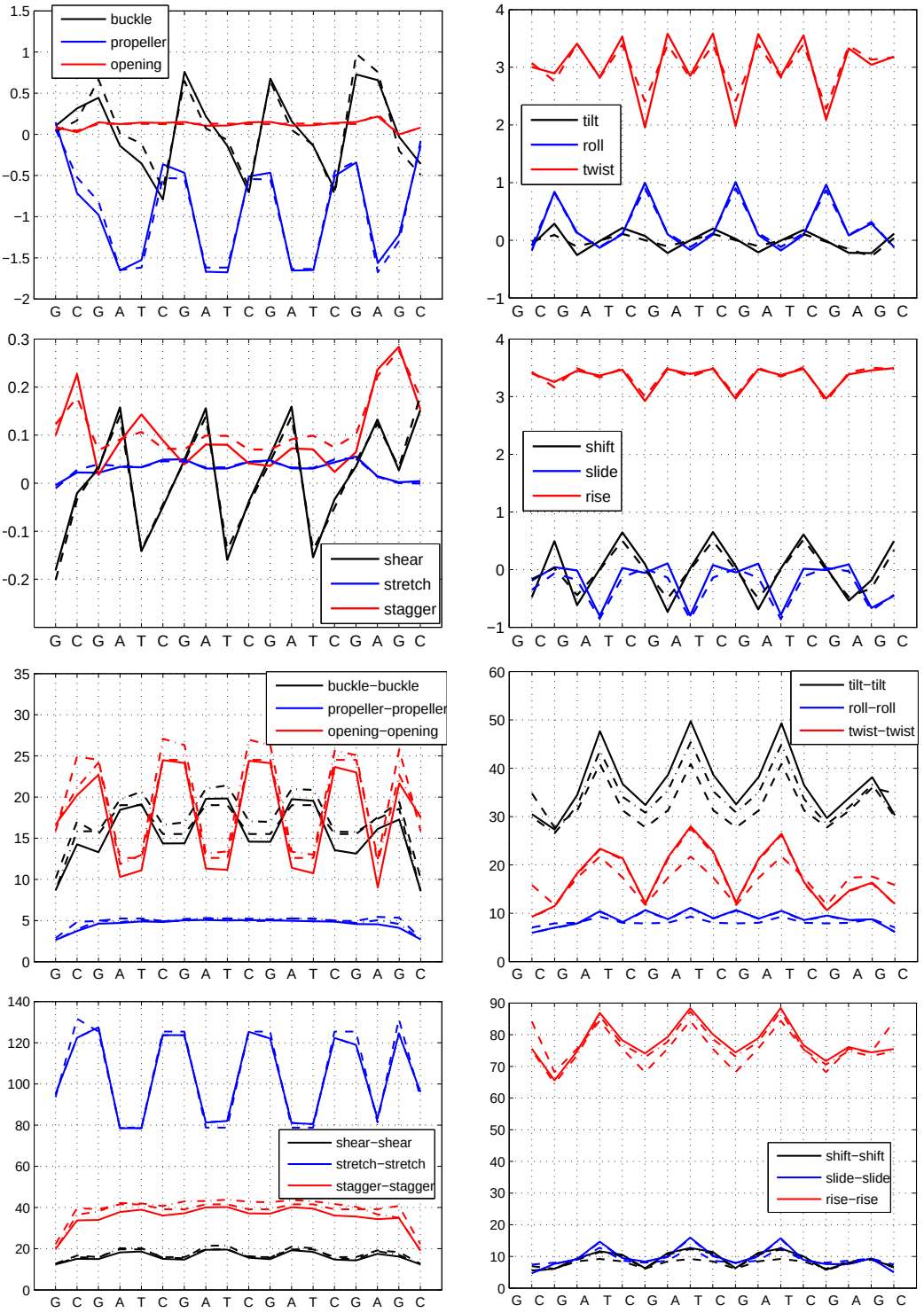


Figure S14: Entries of shape vectors and stiffness matrices in dimensionless units for the non-palindromic, interior period four, 18-mer S_8 from the training set. Top four panels: entries of observed vector $\widehat{w}_{8,o}$ (solid) and constructed dimer-based model vector $\widehat{w}^*_{8,m}$ (dashed). Bottom four panels: diagonal entries of observed matrix $K_{8,o}$ (solid), constructed dimer-based model matrix $K^*_{8,m}$ (dashed) and nearest-neighbor oligomer-based model matrix $K^*_{8,M}$ (dash-dot).

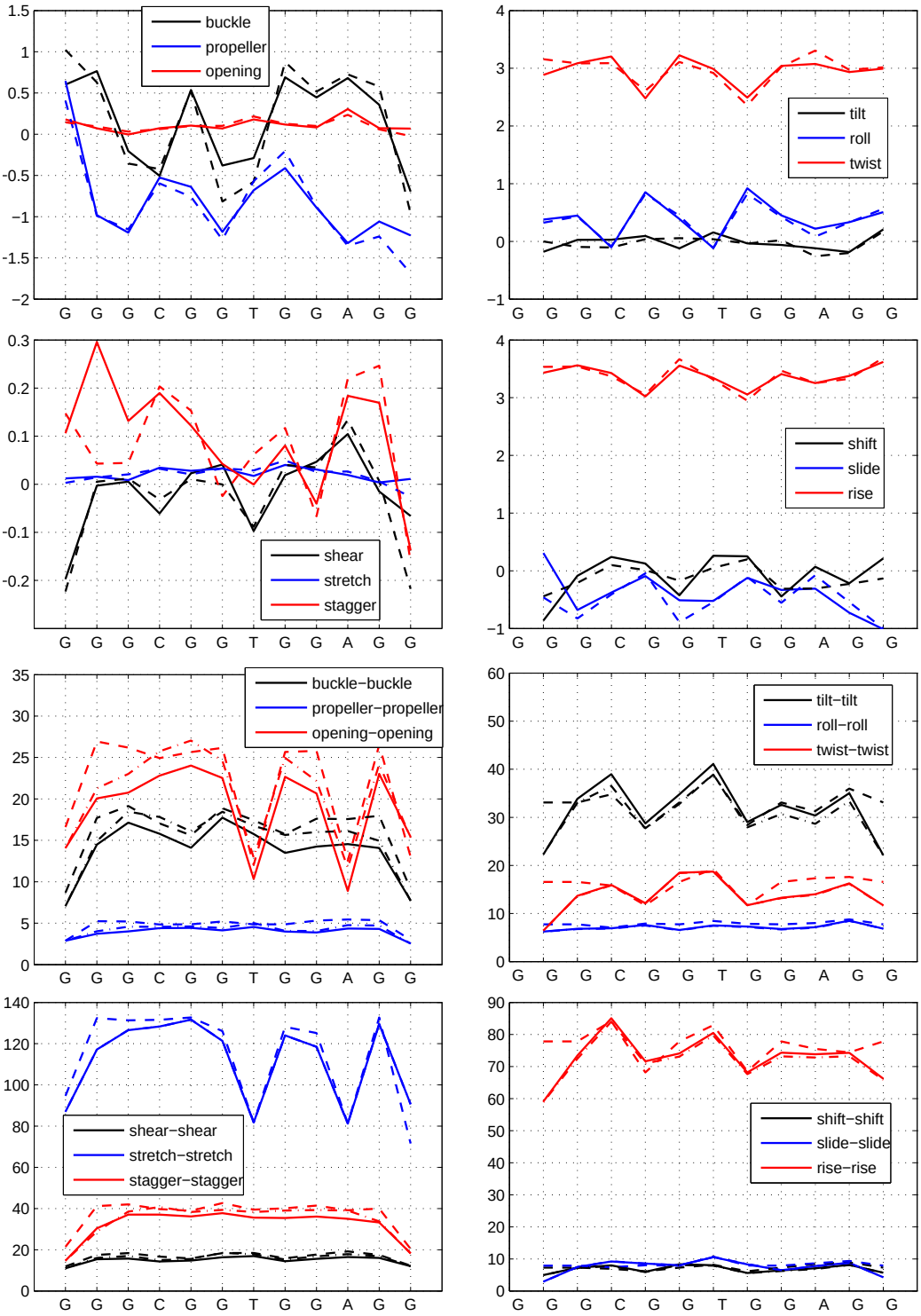


Figure S15: Entries of shape vectors and stiffness matrices in dimensionless units for the non-palindromic, interior non-periodic, 12-mer S_{42} from the training set. Top four panels: entries of observed vector $\widehat{w}_{42,o}$ (solid) and constructed dimer-based model vector $\widehat{w}^{*}_{42,m}$ (dashed). Bottom four panels: diagonal entries of observed matrix $K_{42,o}$ (solid), constructed dimer-based model matrix $K^{*}_{42,m}$ (dashed) and nearest-neighbor oligomer-based model matrix $K^{*}_{42,M}$ (dash-dot).

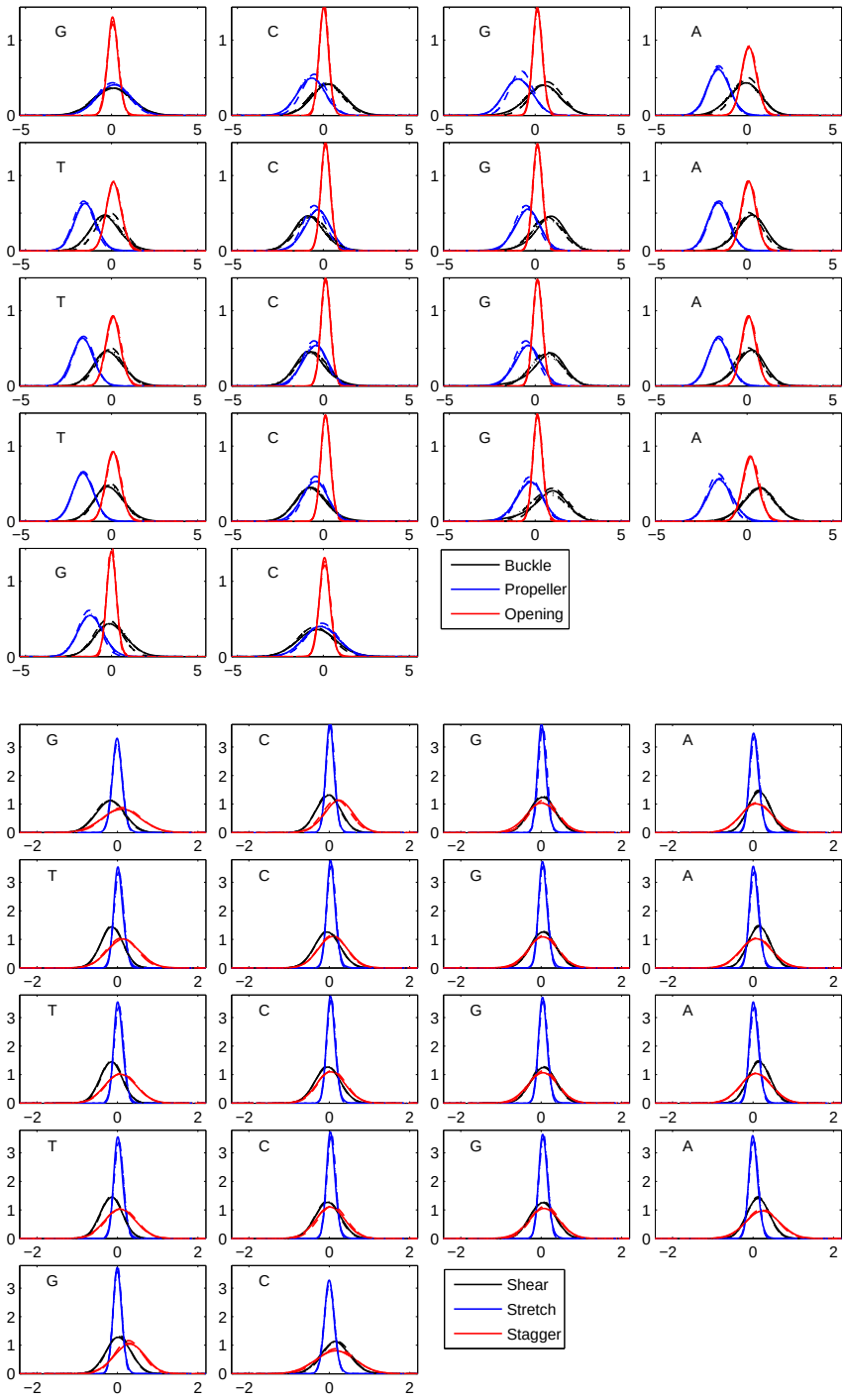


Figure S16: Normalized marginal distributions for intra-basepair coordinates at each position along oligomer S_8. Positions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each position shows the monomer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{8,\mathrm{o}}$, dotted; density $\rho^*_{8,\mathrm{M}}$, dashed-dotted; density $\rho^*_{8,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units. The marginals from the different sources are virtually indistinguishable.

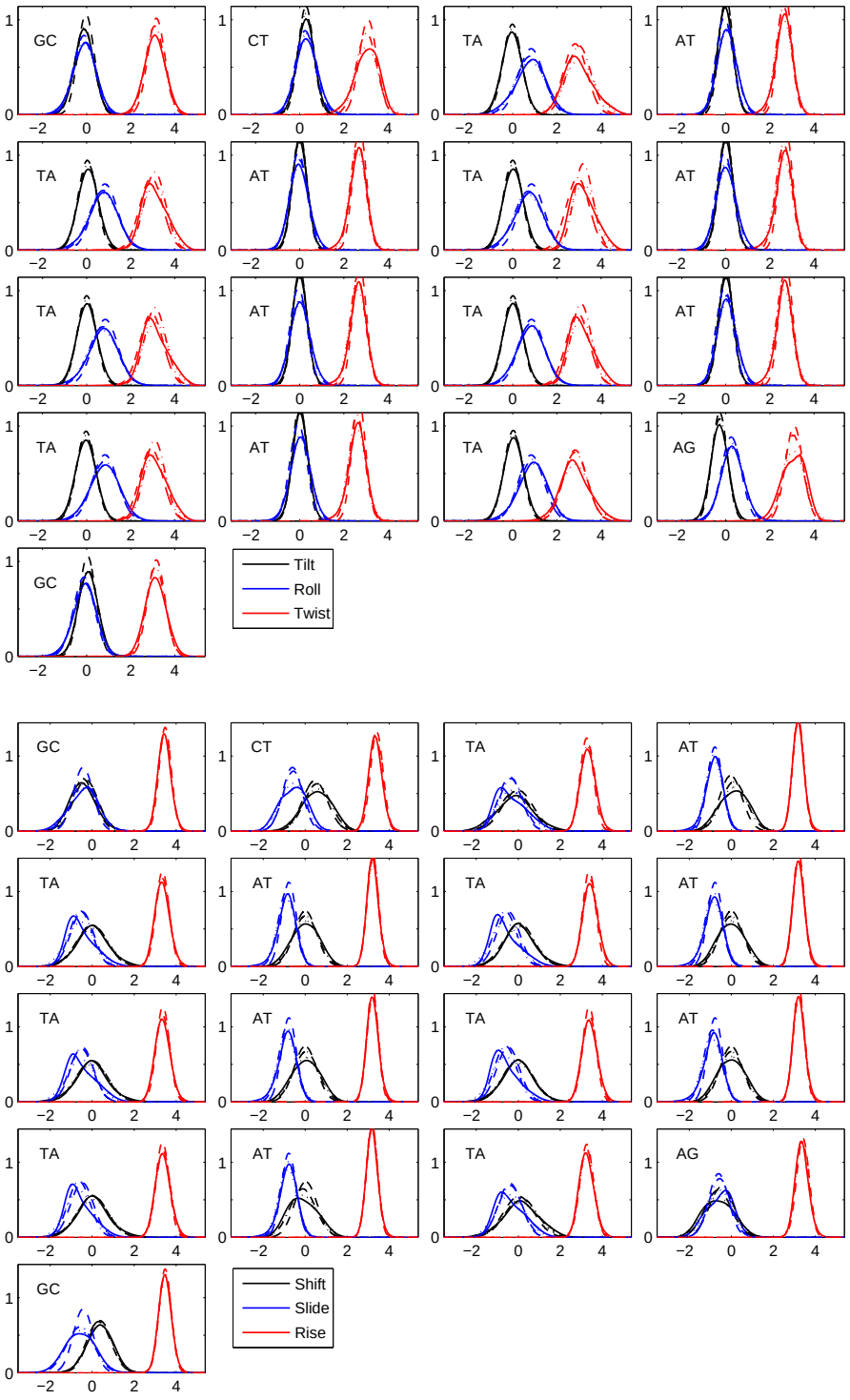


Figure S17: Normalized marginal distributions for inter-basepair coordinates at each junction along oligomer S_1. Junctions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each junction shows the dimer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{1,\mathrm{o}}$, dotted; density $\rho^*_{1,\mathrm{M}}$, dashed-dotted; density $\rho^*_{1,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units.

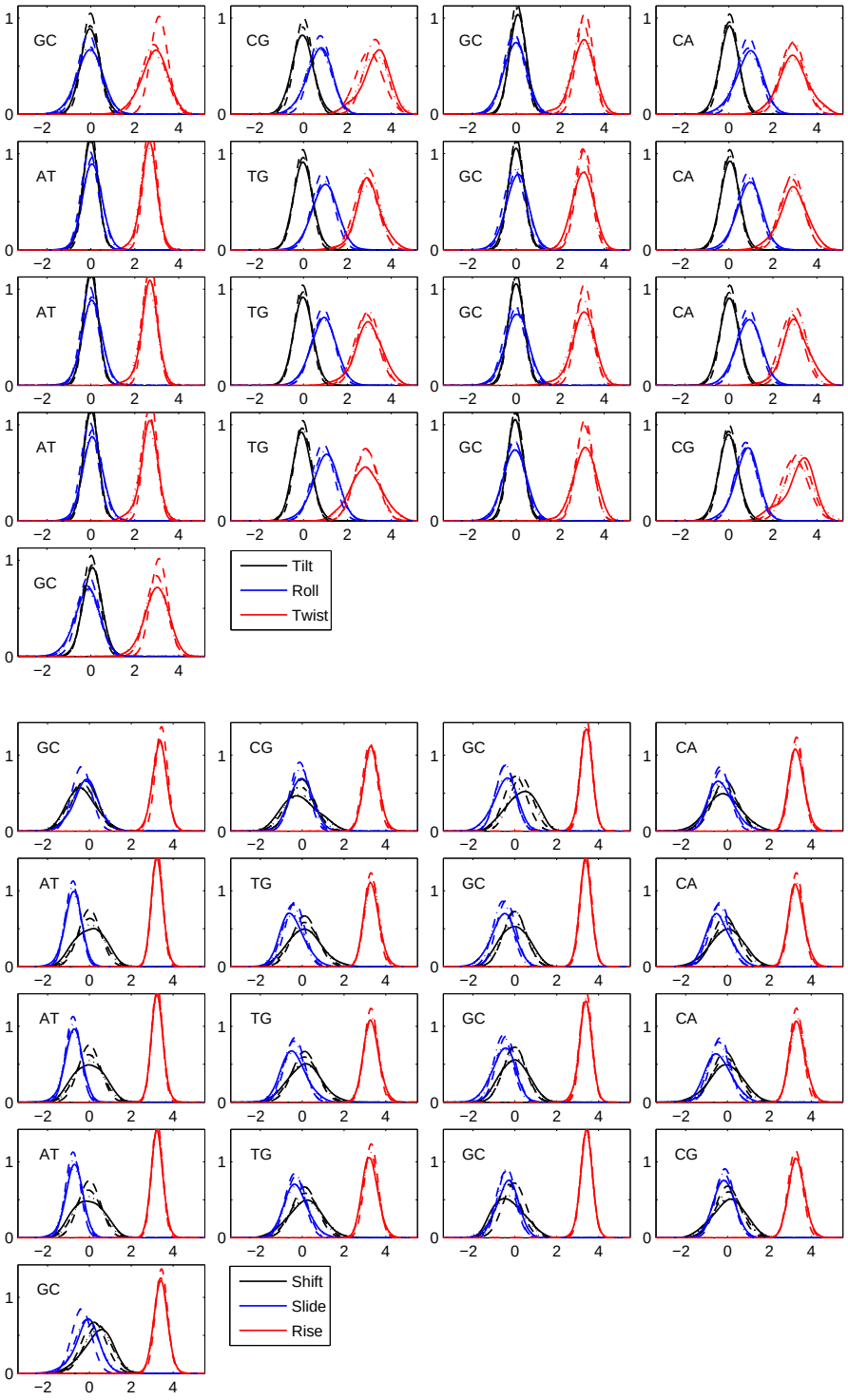


Figure S18: Normalized marginal distributions for inter-basepair coordinates at each junction along oligomer S_3. Junctions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each junction shows the dimer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{3,\mathrm{o}}$, dotted; density $\rho^*_{3,\mathrm{M}}$, dashed-dotted; density $\rho^*_{3,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units.

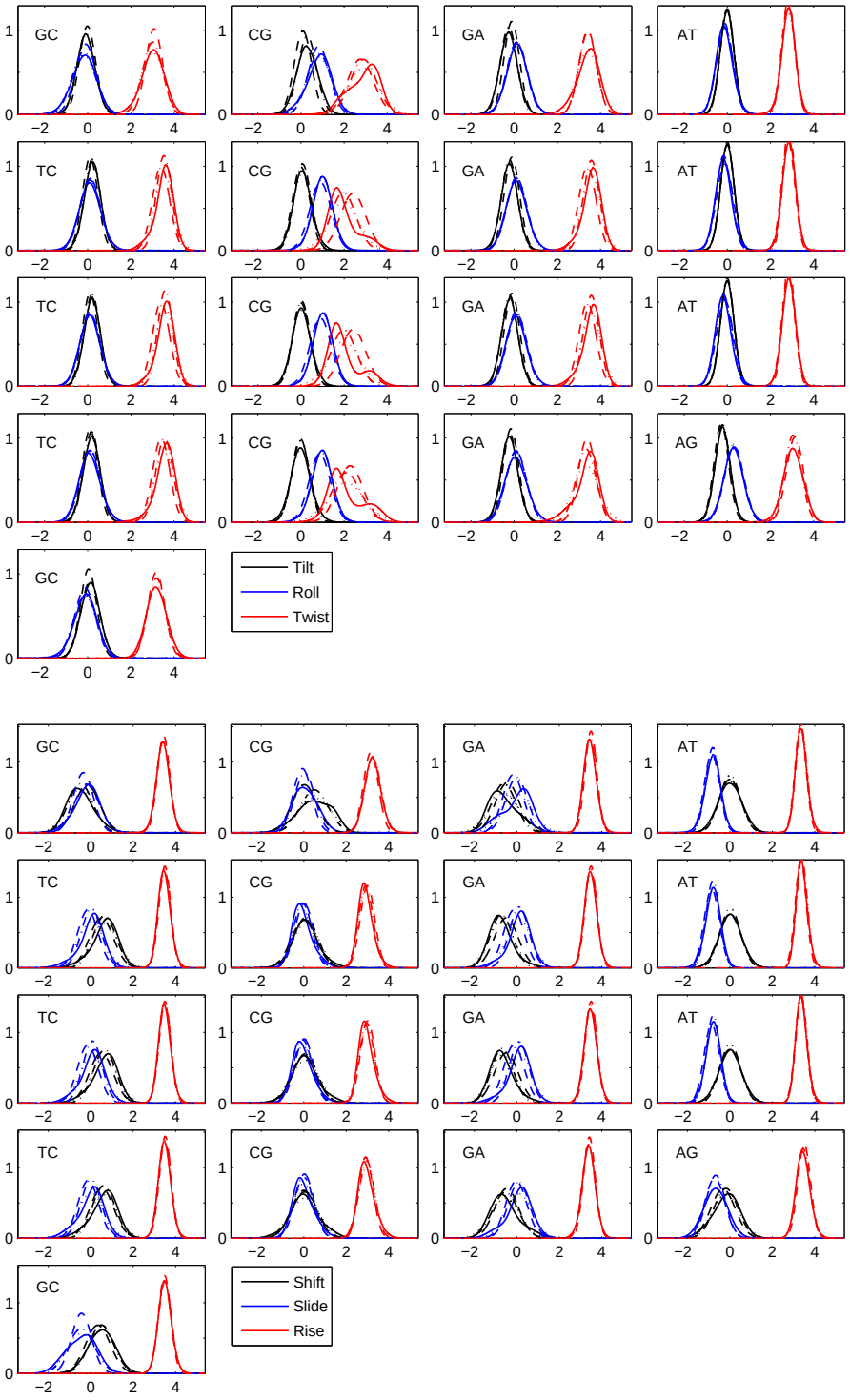


Figure S19: Normalized marginal distributions for inter-basepair coordinates at each junction along oligomer S_8. Junctions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each junction shows the dimer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{8,\mathrm{o}}$, dotted; density $\rho^*_{8,\mathrm{M}}$, dashed-dotted; density $\rho^*_{8,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units.

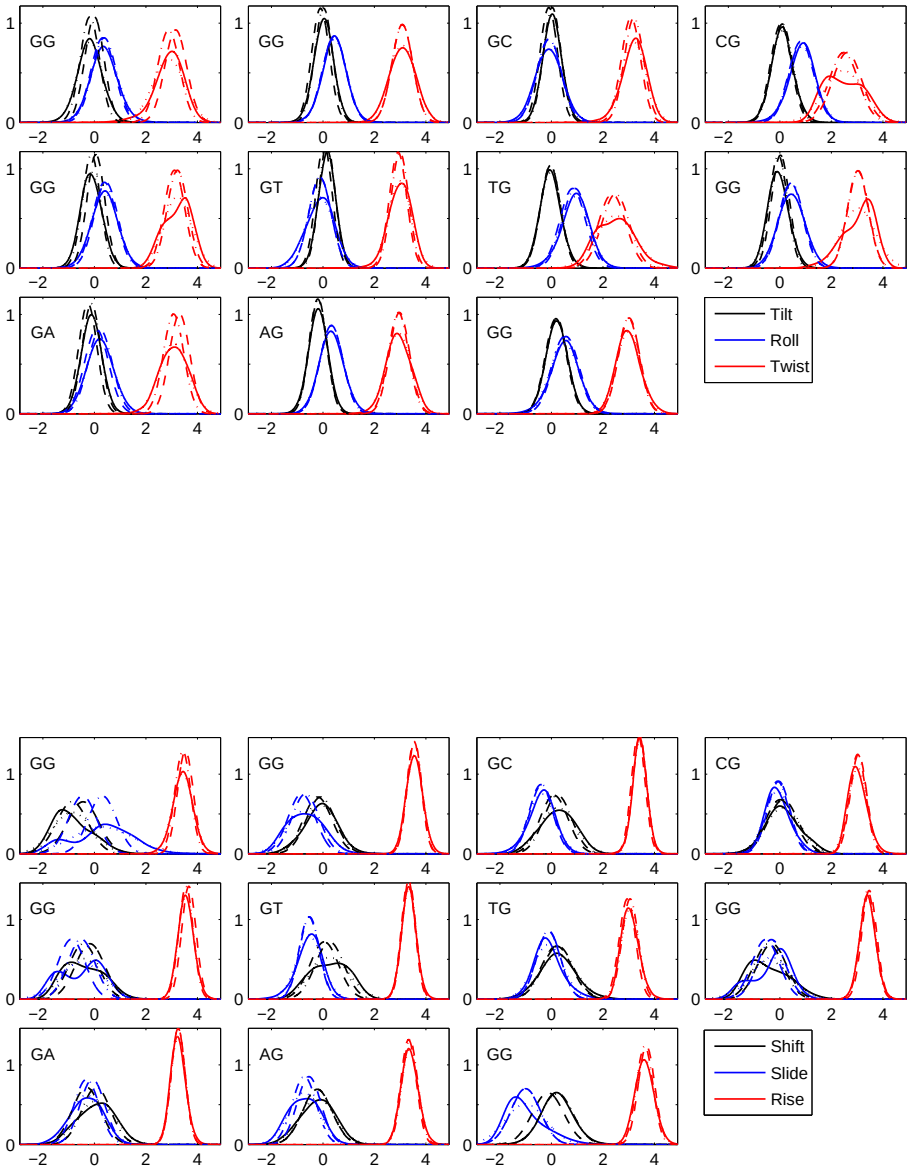


Figure S20: Normalized marginal distributions for inter-basepair coordinates at each junction along oligomer S_{42}. Junctions are ordered left-to-right beginning at top-left in each of the two groups. The panel for each junction shows the dimer on the reference strand and the marginals from each of four sources (MD data, solid; density $\rho_{42,\mathrm{o}}$, dotted; density $\rho^*_{42,\mathrm{M}}$, dashed-dotted; density $\rho^*_{42,\mathrm{m}}$, dashed) for each of three coordinates (black, blue, red) in dimensionless units.