arXiv:2611.32850v1 [cond-mat.soft] 24 Nov 2026

DNA缺陷的几何与拓扑

A. Fassino[1], E.Carlon[1], D.Marenduzzo[2†] 软物质与生物物理，鲁汶大学，*Celestijnenlaan 200D，3001* 鲁汶，比利时[2]苏格兰大学物理联盟，爱丁堡大学物理与天文学学院，爱丁堡，*EH9 3FD*，英国

DNA缺陷经常出现，可能源于损伤、蛋白质结合，或RNA聚合酶等分子酶的作用。在此，我们结合几何学、原子级模拟以及对真实DNA结构的分析，对DNA缺陷及其动力学进行了定量表征。用于从几何上描述缺陷的复合序参量，是一个从骨架指向碱基的极性矢量，以及一个用于在单链中捕捉碱基三维取向的四元数。缺陷分为三类：极性矢量翻转、碱基堆叠缺陷，以及同时包含二者的复合缺陷。所有类型的缺陷都存在于真实DNA结构中，并具有功能相关性。此外，对施加弹性形变的DNA进行偏置分子动力学模拟，形成了翻转缺陷或堆叠缺陷；其丰度和动力学强烈依赖于序列及形变类型。我们的研究结果建立了几何学与生物化学之间的联系，还可用于识别RNA分子和核小体阵列中的缺陷。

在活细胞中，最常见的 DNA 形式是 B型：其几何结构为右手双螺旋，每个螺距约含 10.5 个碱基对，并具有一种特征性的结构不对称性，由此形成大沟和小沟 [1]。偏离这种理想几何结构的程度，可通过相邻碱基对之间的相对平移——滑移、错移和升高——以及相对旋转——滚转、扭转和倾斜——来量化 [1, 2]。对于足够小的变形，其相关弹性代价可用谐和近似描述 [3], ，这与液晶的 Frank弹性自由能十分相似；后者用于量化展曲、弯曲和扭曲形变的能量代价 [4]。这一框架在描述热涨落、序列依赖性曲率和中等程度的超螺旋方面已被证明非常成功 [5]。

然而，尽管教科书将 DNA 描绘成一种非常稳定的分子，实际上它很脆弱，而且相对容易发生幅度过大的变形，无法将其视为 B型双螺旋的小幅弹性扰动 [6]。这类构型更适合解释为螺旋结构的缺陷。常见例子包括 DNA损伤 [7–11], ，例如由氧化损伤或紫外线造成的损伤 [12]。损伤往往会导致 DNA损伤病灶，使双螺旋发生强烈的局部扭曲，并可能引起碱基翻转、局部解堆叠以及沃森-克里克碱基配对破坏。事实上，真核细胞（例如人细胞）会将 DNA 缠绕在蛋白质上，形成染色质纤维，以此保护 DNA；除了促使其压缩外，这种结构也很可能在演化过程中形成，以从力学上保护 DNA，避免缺陷形成 [1]。

DNA 的大幅变形也可能独立于化学损伤而产生。例如，蛋白质结合经常会诱导 DNA 发生严重弯曲或局部解绕 [1]。近期研究刻画了这些高度弯曲（扭结）DNA 构象的自由能景观 [13]。类似地，高度转录的基因组区域会积累扭转应力和超螺旋 [14–16], ，从而驱动局部结构显著偏离经典 B型DNA，并可能成核形成类似缺陷的构象 [17, 18]。尽管这些强烈变形的 DNA 状态具有重要的生物学意义且经常出现，但与小幅弹性扭曲相比，从几何学和理论角度对它们的研究仍相对较少。部分原因在于，后者可以用二次自由能泛函进行自然描述，而大的结构扰动则需要一种能够描述离散几何转变的框架，其中包括碱基翻转、局部链缺陷以及碱基对标架的强烈重新定向。

在本研究中，受软凝聚态物质 [19] 和液晶 [4, 20–22]中拓扑缺陷类比的启发，我们建立了一个用于描述 DNA结构中大尺度局域缺陷的几何框架。为此，我们用局部取向标架及其相关四元数表示每个核苷酸，从而能够在几何和拓扑层面对大幅离散旋转与局部翻转进行刻画。虽然本文聚焦于DNA缺陷，但所引入的形式体系具有相当普适性，也可应用于其他问题，例如RNA或染色质中的核小体构象。

*DNA*缺陷的几何与拓扑 – 我们将单条DNA链建模为一维字符串，以 s为参数，其中碱基的局部取向由两个复-分量 [图 1(a)]。第一个是“骨架到碱基向量”$\mathbf{n}(s)$，它连接 $O5'$ 氧原子与碱基对中的一个氮原子：对于嘧啶，它连接的是 $N1$ 原子；对于嘌呤，则连接$N9$ [23]。$\mathbf{n}(s)$向量近似垂直于链骨架的单位切向量$\mathbf{t}(s)$，因此骨架到碱基向量可用单个角度参数化，并对应于 S^1的一个元素。第二个分量是碱基取向，它被描述为一个矩形“砖块”，可以视为双轴向列相分子或一个三元组，因此等价于一个单位四元数 $q(s)$，也可以表示为 SO(3) 中的三维旋转。[24]我们将二元组（$\mathbf{n}(s)$ $q(s)$)称为描述沿一条 DNA 链的碱基取向的局部序参量。

因此，描述一条链上碱基的局域序参量流形为 $\mathcal{M} = S^1 \times SO(3)$。如果考虑一条链上（缺陷两侧）的初态和末态相同的情形，例如双螺旋转过一整圈之后，或在一个 DNA 环中，那么这些状态由 $\mathcal{M}$中的同一个元素表示。因此，我们可以设想将两端连接起来，使这些构型可由 $\mathcal{M}$的第一同伦群元素进行拓扑分类，即 $\pi_1(\mathcal{M})$ [19]，；我们在 [25] 中证明其等于 $\mathbb{Z} \times \mathbb{Z}_2$。从物理上说，$\pi_1(\mathcal{M})$的元素代表宏观不同构型的类别；若不实际切断链或彻底翻转某个取向，这些构型就无法连续变形为彼此。这些元素可与一对整数 (m, n) 关联。其中，m 表示所考察向量沿链的绕行数 [25]，而n 则表征描述碱基（或砖块）的三元组的拓扑。后一个指标对应于三元组的类自旋扭转奇偶性；狄拉克腰带技巧直观地展示了这一拓扑性质 [26]。

虽然这些全局拓扑不变量能够对序参量空间中的所有环进行分类，但我们在此主要关注局域形变或点状缺陷。对于这些情况，m 实际上只能取 0 或 1 [25]。因此，我们可以将每个碱基步分为四种拓扑状态。第一，$(m, n) = (0, 0)$ 对应于无缺陷的碱基步。第二，$m = 1$对应于“翻转缺陷”，它与骨架到碱基向量的局部反向有关 [图 1(b)(i)]。第三，$n = 1$ 表示“堆叠缺陷”，即碱基三元组发生突变式 π 旋转 [示例如图 1(b)(ii) 所示]。后一种旋转可以绕链切向量的轴，也可以绕轴 $\hat{\mathbf{e}}_3$ （四元数旋转）、绕 $\hat{\mathbf{e}}_1$ （四元数滚转），或绕 $\hat{\mathbf{e}}_2$ （四元数摆动）进行。这三种相应的堆叠缺陷在拓扑上等价，且都具有 $n = 1$；然而，由于与骨架相连，只有 π 四元数滚转在不发生碰撞的情况下才具有空间可行性。实际上，旋转和摆动堆叠缺陷还需要同时出现翻转缺陷，因此对它们而言，$(m, n) = (1, 1)$ 是翻转与堆叠缺陷的组合 [而对于滚转堆叠缺陷，则可以是 $(m, n) = (0, 1)$]。

现在我们讨论如何用一条（$\mathbf{n}(s)$ q

(s))曲线描述理想缺陷。为简单起见，考虑一条沿 z 轴对齐的直线 DNA骨架（如图 1 所示；因此 $s = z$）。对于翻转缺陷 [图 1(b)(i)]，骨架到碱基向量 $\mathbf{n}(s)$呈径向向外方向（例如沿 $+\hat{\mathbf{x}}$），但在缺陷位置 $s = s_0$处，碱基翻转了 π:

$$\mathbf{n}(s) = \hat{\mathbf{x}}, \text{ if } s \neq s_0; \qquad \mathbf{n}(s_0) = -\hat{\mathbf{x}}. \tag{1}$$

一种（滚转）堆叠缺陷 [图 1(b)(ii)]也可以描述为：骨架沿着 z，并且具有一个恒定的

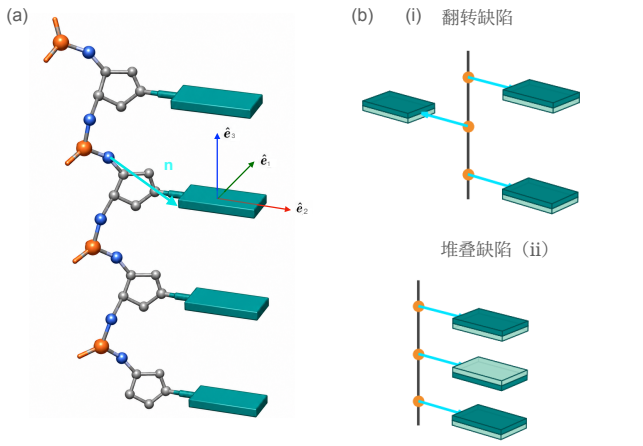


图 1。（a）$\mathbf{n}(s)$ 和 $q(s)$ 的定义。（b）（i）翻转缺陷示例，其中骨架到碱基向量发生翻转。（ii）纯堆叠缺陷示例，其中四元数砖块在 π 的两步中滚转 2π。翻转缺陷和堆叠缺陷也可以组合出现，即（i）和（ii）同时发生。

$\mathbf{n} = \hat{\mathbf{x}}$，四元数字段为：

$$q(s) = 1 \text{ if } s < s_0; \; q(s_0) = \mathbf{j}; \; q(s) = 1 \text{ if } s > s_0. \tag{2}$$

这种纯堆叠示例的一个实例是 Hoogsteen 碱基配对 [27]，，其中嘌呤绕其糖苷键旋转，与相对的嘧啶建立一种替代性的氢键模式。Watson–Crick/Hoogsteen 互变发生在 μs–ms 时间尺度上 [28]。可以使用类似于公式（2）的公式来表征四元数旋转和四元数摆动，将**j**替换为 **i** 或 **k**；不过如前所述，由于碱基的几何形状，这些情况还需要一个翻转缺陷，而这通常会造成空间位阻冲突。除了公式（1、2）中的真正缺陷（它们无法通过连续变形得到无缺陷分子）之外，下面我们还将追踪一些同样具有生物学相关性的强变形，尽管它们并不是真正的拓扑缺陷，例如某个碱基四元数转动 $\pi/2$，随后又松弛恢复。

为确保我们对离散局部转变（翻转和堆叠缺陷）的分类反映的是碱基真实的几何扰动，而非 DNA骨架的曲率，我们还纳入了一种基于平行移动（Bishop）标架的几何校正 [25]。当单链发生显著弯曲或缠绕数变化时，连续核苷酸在全局（实验室）参考系中的取向会发生变化。这种宏观曲率会对原始的碱基步进向量点积和四元数转变引入几何贡献，从而掩盖缺陷。通过将极性骨架到碱基的跟踪向量$\mathbf{n}(s)$以及刚体三联架投影到沿 $O5'$ 骨架中心线

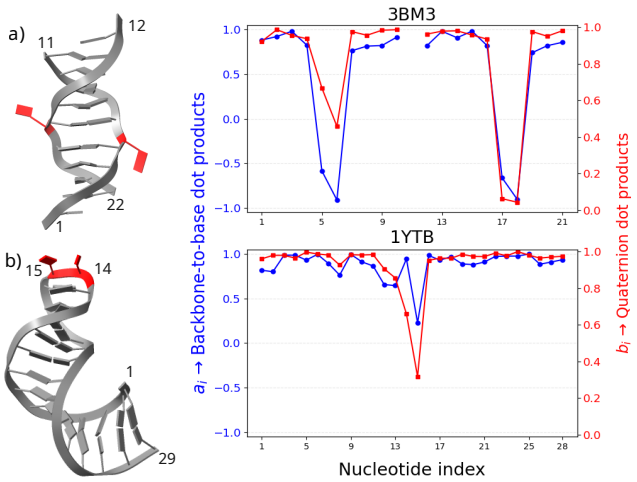


图2。（a）含有两个翻转缺陷的DNA分子示例（PDB条目：3BM3）。其中一个翻转缺陷同时也是堆叠缺陷。（b）堆叠缺陷示例（1YTB）。快照中以红色标出缺陷。

轨迹构建的平行移动 Bishop 标架中，我们随后消除了由三维缠绕引起的空间变化——我们在下文中采用的这一过程详见 [25]。

天然*DNA*缺陷 – 天然DNA结构中的构象缺陷主要通过两种机制产生：内在热涨落，以及外部因素引发的扰动，例如扭转应力（如转录诱导的超螺旋 [14, 16]）或蛋白质结合。在某些情况下，蛋白质缺陷也是促进修复蛋白结合的信号 [29, 30]。而由热涨落产生的缺陷通常是DNA双链体中短暂的局部激发，对应于短暂偏离经典B型几何构型的状态，随后又恢复至天然状态。尽管光谱技术能够提供这些瞬态激发的群体比例和寿命信息，但其详细的原子动力学仍无法直接观测。相比之下，存入蛋白质数据库（PDB） [31] 的X射线晶体学和冷冻电镜结构提供了DNA构象的高分辨率快照，其中也包括局部扭曲的状态。

图2左侧给出了一个由蛋白质结合（II型限制性内切酶）诱导的局部缺陷的蛋白质数据库结构示例。蛋白质-DNA结合会在每条DNA链中产生一个碱基翻转缺陷 [32]。右图绘出了两个序参量的数值：两个连续骨架到碱基的极性矢量 $a_i \equiv \hat{n}_i \cdot \hat{n}_{i+1}$ （蓝色）的点积，以及两个连续碱基四元数 $b_i \equiv |q_i \cdot q_{i+1}|$ （红色）的点积。索引 $1 \le i \le 9$ 标记A链上的碱基，而 $10 \le i \le 21$标记B链上的碱基，如图2(a)左侧所示。我们注意到，碱基 i 处的单个碱基缺陷会导致其相邻的两个序参量 a_{i-1}发生偏离，a_i，以及 b_{i-1}、 b_i。我们的几何分析表明，两条链上的缺陷都对应于骨架到碱基的极性矢量近乎纯粹的反转，如$a_{i-1} \approx a_i \approx -1$所示。然而，两条链的解堆叠程度不同，B链中的缺陷明显比A链中的缺陷具有更强的解堆叠程度（$b_{i-1} \approx b_i \approx 0$），而A链中为（$b_{i-1} \approx b_i \approx 0.5$）。图2(b)的结构显示，单个碱基对发生了较为有限的翻转和较弱的解堆叠（$a_{i-1} \approx 1$、 $a_i \approx 0$ 和 $b_{i-1} \approx 0.6$、$b_i \approx 0.3$）。这些由蛋白质诱导的畸变提供了实验解析的构象缺陷实例，并为表征非典型DNA状态提供了有价值的结构基准。

原子级模拟中的*DNA*缺陷 – 来自蛋白质数据库的实验结构提供了DNA的静态快照，因此对于构象缺陷成核和传播的动态路径只能提供有限的见解。为研究给定DNA序列形成缺陷的倾向及其后续动力学，我们开展了全原子分子动力学模拟（有关所用方案的更多详情，请参见 [25] ）。简而言之，为诱导缺陷形成，我们使用核酸刚性碱基偏置（RBB-NA）算法 [23]施加局部构象偏置。RBB-NA在一类粗粒化刚性碱基坐标 α 上施加谐和偏置势，其形式为

$$V_{\text{bias}}(\alpha) = \frac{K}{2}\,(\alpha - \bar{\alpha})^2\,, \tag{3}$$

其中 K 是约束的刚度， $\bar{\alpha}$ 是我们所考虑坐标的目标值。后者可以是十二个刚性碱基参数 [2], 中的任意一个，例如扭转、滚转、螺旋桨扭转、开口或倾斜。偏置（3）将 α 驱动至规定的 $\bar{\alpha}$。RBB-NA算法此前已用于伞形采样研究，以刻画强烈弯曲和/或扭转的DNA构象的自由能景观 [13]。在此，我们采用相同的方法，通过施加足够强的约束，即取K 和 $\bar{\alpha}$的较大值，来诱导偏离典型B型DNA的受控变化。借助该方法，我们可以研究DNA中缺陷形成和传播的若干关键方面。首先，我们可以量化序列依赖效应，例如识别对缺陷形成更易感或更不易感的序列基序。其次，我们的模拟能够揭示这些扰动沿DNA分子发展、弛豫和传播的空间范围及时间尺度。

图 3 展示了四种不同的 14聚体序列——聚腺嘌呤、PolyAT、PolyCG 和聚胞嘧啶——在中间碱基对 [33] 处受到初始螺旋桨扭转偏置后，于 20 纳秒无偏模拟时间尺度内， a_i 和 b_i 沿整个 DNA分子随时间演化的时空图。结果的主要特征是，序列会显著影响结构缺陷在初始偏置区域

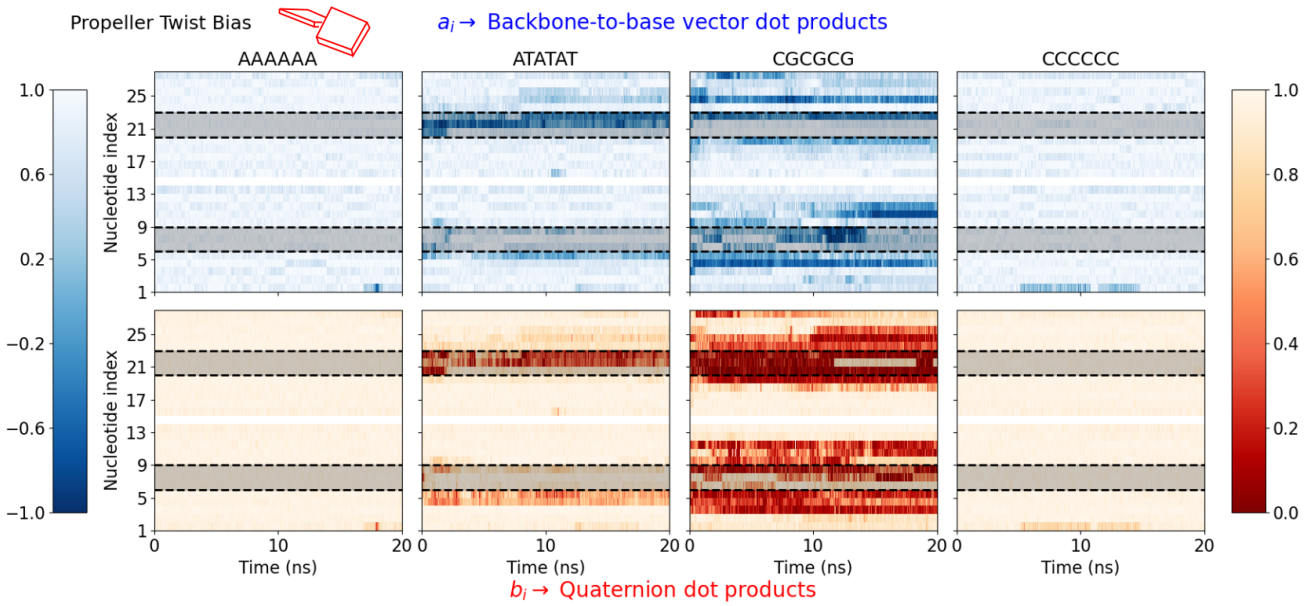


图 3. 运动图展示了骨架到碱基的变形及四元数取向，对应于在 DNA 14聚体中部对螺旋桨扭转施加局部偏置的模拟。偏置施加了 1 ns，$\bar{\alpha} = 0.3$（图中未显示）；随后移除偏置，让系统平衡 20 ns（图中所示）。运动图中的虚线标出了施加偏置的区域。

及其他位置形成的可能性。具体而言，PolyA 和 PolyC 序列几乎没有缺陷。相较之下，PolyAT 和 PolyCG 则发生了严重形变。对于 PolyCG，连续的堆叠缺陷遍布整个 14聚体，而翻转缺陷也很常见，但彼此孤立，通常不连续。对于 PolyAT，翻转与堆叠缺陷共同形成于距离偏置碱基对一定距离的位置，朝向分子的中部。

这些结果与体内实验结果一致：CpG岛中的基因启动子更容易熔解，尤其是在受到聚合酶活性引起的负超螺旋作用时 [16], ，尽管从热力学角度来看，CG碱基对应比AT碱基对更稳定、不易熔解 [1]。这些结果也与单分子实验相符：在受到扭转应力时，富含CG的序列比富含AT的序列更容易熔解 [34]。此外，我们认为，在 PolyAT 中观察到的远距离缺陷形成提供了一个具有功能意义的例子：碱基对处的形变信号（图 3 中的螺旋桨扭转）能够通过 DNA 介导长程通信，进而在相隔数个碱基对结合的蛋白质之间产生有效相互作用，从而促进 DNA 识别与调控中的变构机制 [35]。

图 3 中的时空图表明，翻转缺陷和堆叠缺陷的动力学具有很强的序列依赖性。为量化缺陷形成与退火的动力学，我们将 a_i 和 b_i 的时间依赖轨迹映射到电报模型 [36] 中，该模型采用高斯隐马尔可夫

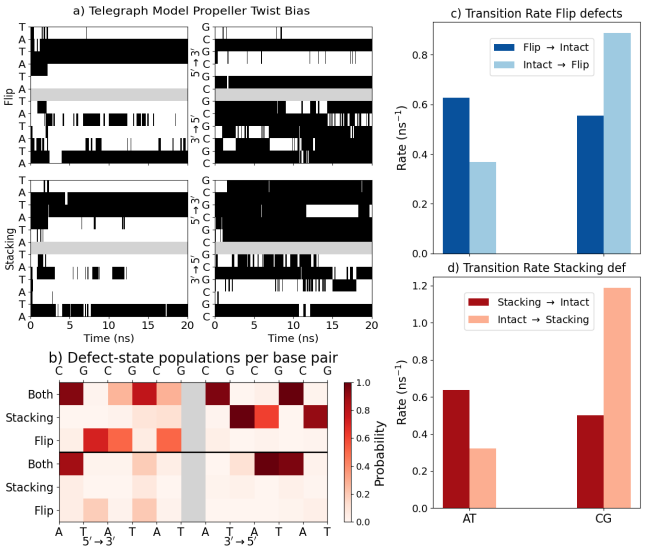


图 4. 基于隐马尔可夫模型分析，在螺旋桨扭转偏置下的缺陷动力学与统计特性。(a) 对 PolyAT 和 PolyCG序列在 20 ns 内的缺陷状态进行跟踪的代表性电报模型轨迹（黑色区间表示缺陷存在）。(b) 每个碱基对中出现特定缺陷状态——纯翻转、纯堆叠或组合状态（“Both”）——的概率。(c、d) 完整碱基对与 (c) 翻转缺陷或 (d) 堆叠缺陷之间的正向和反向转变速率，适用于 AT 和 CG步。

马尔可夫模型 [隐马尔可夫模型；详见图4(a)及文末补充材料]。因此，每个核苷酸都被映射到四种状态之一：无缺陷、翻转缺陷、堆叠缺陷，以及

组合缺陷（翻转加堆叠）。

对隐马尔可夫模型轨迹中占据统计量的分析 [图4(b)] 显示，在螺旋桨扭转偏置下，PolyAT主要呈现复合缺陷，而PolyCG缺陷则表现为复合缺陷或堆叠缺陷，两者通常都呈簇集分布。翻转型和堆叠型缺陷的形成与退火速率量级为 ∼纳秒$^{-1}$ [图4(c,d)], ，且PolyCG中的形成速率显著较高。最后，在螺旋桨扭转偏置模拟中，缺陷退火时间的概率分布 [即图7(a)（补充材料）中的“缺陷存活概率”] 并非指数分布。这表明缺陷修复并不存在单一时间尺度，因为相关缺陷岛经常出现，而缺陷岛的退火时间尺度取决于其大小。

采用倾斜和开口偏置的模拟（见补充材料）证实了螺旋桨扭转模拟中观察到的定性趋势：缺陷形成和退火取决于序列，且为 ∼ 纳秒$^{-1}$。不过，模拟也表明，所施加偏置的几何形态会产生显著影响：例如，与螺旋桨扭转偏置相反，开口偏置会使PolyAT序列出现大量缺陷，而PolyCG序列中的缺陷则很少。

结论 总之，我们借鉴软物质物理学中的缺陷和拓扑学概念，提出了一个用于表征DNA中大型局部结构扰动的几何框架 [19]。通过引入由极性骨架至碱基矢量 $\mathbf{n}(s)$ 和与DNA碱基的砖状几何结构相对应的单位四元数 $q(s)$ 构成的复合序参量，我们证明DNA缺陷可分为三类：翻转缺陷、堆叠缺陷，以及同时包含二者的复合缺陷。

首先，通过分析蛋白质数据库中的数据，我们证明DNA缺陷常见于体内，并且具有功能相关性。例如，限制性酶利用翻转缺陷，在其活性切割位点内分离特定的目标序列；而堆叠缺陷可能充当机械信号，被转录因子读取。我们还通过偏置原子级分子动力学模拟表明，在弹性约束下，缺陷形成与退火具有很强的序列依赖性。值得注意的是，与同聚物链段（聚腺嘌呤和聚胞嘧啶）相比，交替序列（PolyAT和PolyCG）对缺陷形成的易感性显著增强；其中，CpG步在螺旋桨扭转应力下最为脆弱，而这种应力可能出现在高表达基因中。CpG区域的这种机械脆弱性，或可为这些区域中熔化结构和非B型DNA结构形成的富集现象提供生物学解释 [16]。我们的模拟还表明，缺陷形成与退火是一个受到缺陷岛形成强烈影响的时空相关过程。

尽管我们关注的是裸DNA 中的结构缺陷，但这些序参量也可能适用于其他生物系统，例如染色质。在染色质中，DNA 缠绕在组蛋白八聚体周围，形成核小体。由于核小体具有明确的几何形状和取向，四元数场可以描述连续核小体之间的相对取向，而翻转向量则可以量化它们相对于连接DNA的取向。因此，诸如此处所用的复合序参量，可能为识别核小体阵列层级的缺陷提供一个定量框架 [37]。

致谢 – A.F. 感谢欧盟资助的 MeCha-NiSM 博士网络（使用单分子技术对核酸进行机械表征）在“地平线欧洲”2021–2027 框架计划下提供的资金支持，资助协议编号为 101168851。本研究使用的资源和服务由 VSC（佛兰德超级计算中心）提供，其资金来自佛兰德研究基金会（FWO）和佛兰德政府。

[1] C. R. Calladine 和 H. Drew，《理解*DNA*：分子及其作用方式》（学术出版社，1997）。[2] W. K. Olson、M. Bansal、S. K. Burley、R. E. Dicker- son、M. Gerstein、S. C. Harvey、U. Heinemann、X. J. Lu、S. Neidle、Z. Shakked、H. Sklenar、M. Suzuki、C. S. Tung、E. Westhof、C. Wolberger 和 H. M. Berman，《分子生物学杂志》**313**，229（2001）。[3] A. D. Bates 和 A. Maxwell，《*DNA*拓扑学》（牛津大学出版社，2005）。[4] P.-G. De Gennes 和 J. Prost，《液晶物理学》（牛津大学出版社，1993）。[5] J. F. Marko 和 E. D. Siggia，《科学》**265**，506（1994）。[6] T. Lindahl，《自然》**362**，709（1993）。[7] E. C. Friedberg，《自然》**421**，436（2003）。[8] P. R. Cook，《细胞核结构与功能原理》（威利出版社，2001）。[9] E. C. Friedberg、G. C. Walker、W. Siede 和 R. D. Wood，《*DNA*修复与诱变》（美国微生物学会出版社，2005）。[10] P. A. Sarma、C. Abbadie、Y. de Launoit 和 F. Cleri，DNA **4**，530（2024）。[11] R. An、Y. Jia、B. Wan、Y. Zhang、P. Dong、J. Li 和 X. Liang，《公共科学图书馆·综合》**9**，e115950（2014）。[12] A. R. Poetsch，《计算与结构生物技术杂志》**18**，207（2020）。[13] A. Fassino、E. Carlon 和 A. Voorspoels，《化学理论与计算杂志》**22**，981（2026）。[14] C. Naughton、N. Avlonitis、S. Corless、J. G. Prender- gast、I. K. Mati、P. P. Eijk、S. L. Cockroft、M. Bradley、B. Ylstra 和 N. Gilbert，《自然·结构与分子生物学》**20**，387（2013）。[15] N. Gilbert 和 D. Marenduzzo，《细胞生物学观点》**92**，102448（2025）。[16] C. Naughton、A. Bonato、M. Chiang、S. Corless、J. Stocks、G. R. Grimes、D. Halliday、A. Bentivoglio、C. A. Brackley、D. Marenduzzo、等，bioRxiv，2026（2026）。

[17] J. Mitchell、C. Laughton 和 S. A. Harris，《核酸研究》39，3928（2011）。[18] T. Sutthibutpong、C. Matek、C. Benham、G. G. Slade、A. Noy、C. Laughton、J. P. K. Doye、A. A. Louis 和 S. A. Harris，《核酸研究》44，9121（2016）。[19] N. D. Mermin，《现代物理评论》51，591（1979）。[20] O. Lavrentovich 和 M. Kleman，《液晶的手性》，115（2001）。[21] G. P. Alexander、B. G.-g. Chen、E. A. Matsumoto 和 R. D. Kamien，《现代物理评论》84，497（2012）。[22] N. Johnson、L. Head、O. Lavrentovich、A. Morozov、G. Negro、E. Orlandini、C. Smith、G. Vasil 和 D. Marenduzzo，《物理评论X》15，041052（2025）。[23] A. Voorspoels、J. Vreede 和 E. Carlon，《化学理论与计算杂志》19，902（2023）。[24] 这里值得指出的是，由于四元数等价于SU(2)，它们实际上构成物理旋转空间SO(3)的双覆盖。这意味着 q 和 $-q$ 在SO(3)中对应同一个旋转。[25] 关于用于修正可观测量、以考虑中心线曲率的Bishop修正、GROMACS模拟以及其他结果的更多详情，请参阅XXX处的在线补充信息。[26] E. D. Bolker，《美国数学月刊》80，977（1973）。[27] K. Hoogsteen，《晶体学报》16，907（1963）。[28] E. N. Nikolova、E. Kim、A. A. Wise、P. J. O’Brien、I. And- ricioaei 和 H. M. Al-Hashimi，《自然》470，498（2011）。[29] S. Ghediri、P. A. Sarma、V. Saravanan、C. Abbadie、R. Blossey 和 F. Cleri，《生物分子》15，649（2025）。[30] V. Saravanan、N. Raouraoua、G. Brysbaert、S. Giordano、M. F. Lensink、F. Cleri 和 R. Blossey，《科学报告》15，7993（2025）。[31] H. M. Berman、J. Westbrook、Z. Feng、G. Gilliland、T. N. Bhat、H. Weissig、I. N. Shindyalov 和 P. E. Bourne，《核酸研究》28，235（2000）。[32] R. H. Szczepanowski、M. A. Carpenter、H. Czapinska、M. Zaremba、G. Tamulaitis、V. Siksnys、A. S. Bhagwat 和 M. Bochtler，《核酸研究》36，6109（2008）。[33] 该序列为CGCGAAAAAACGCG，因此两侧为CG，中间有6个聚腺嘌呤。其他序列 [25]也是如此。[34] R. Vlijm、J. V. D. Torre 和 C. Dekker，《公共科学图书馆·综合》10，e0141576（2015）。[35] M. Segers、A. Voorspoels、T. Sakaue 和 E. Carlon，《物理评论快报》131，238402（2023）。[36] C. W. Gardiner，《施普林格协同学丛书》（1985）。[37] R. Garces、R. Podgornik 和 V. Lorman，《物理评论快报》114，238102（2015）。[38] R. Sharma、A. S. Patelli、L. De Bruin 和 J. H. Maddocks，《分子生物学杂志》，167978（2023）。[39] A. Pérez、I. Marchán、D. Svozil、J. Sponer、T. E. Cheatham、C. A. Laughton 和 M. Orozco，《生物物理学杂志》92，3817（2007）。[40] I. Ivani、P. D. Dans、A. Noy、A. Pérez、I. Faustino、A. Hospital、J. Walther、P. Andrio、R. Goñi、A. Bal- aceanu、等，《自然·方法》13，55（2016）。[41] W. L. Jorgensen、J. Chandrasekhar、J. D. Madura、R. W. Impey 和 M. L. Klein，《化学物理学杂志》 79，926 （1983）。[42] T. Darden、D. York 和 L. Pedersen，《化学物理学杂志》98，10089（1993）。

结束内容

具有倾斜和开口偏置的原子级模拟中的DNA缺陷

在正文中，我们展示了具有螺旋桨扭转偏置的原子级模拟结果。在本节中，我们将讨论其他类型弹性扰动下的缺陷动力学。首先，图5展示了与图4中相同方案对应的运动图，但这里考察的是倾斜偏置。结果表明，与螺旋桨扭转偏置相比，倾斜扰动导致的缺陷显著更少。在这种情况下，产生缺陷较多的是PolyAT 序列。与螺旋桨扭转偏置的情况类似，还存在远距离作用：缺陷形成于距最初施加扰动的位置数个碱基对之外。

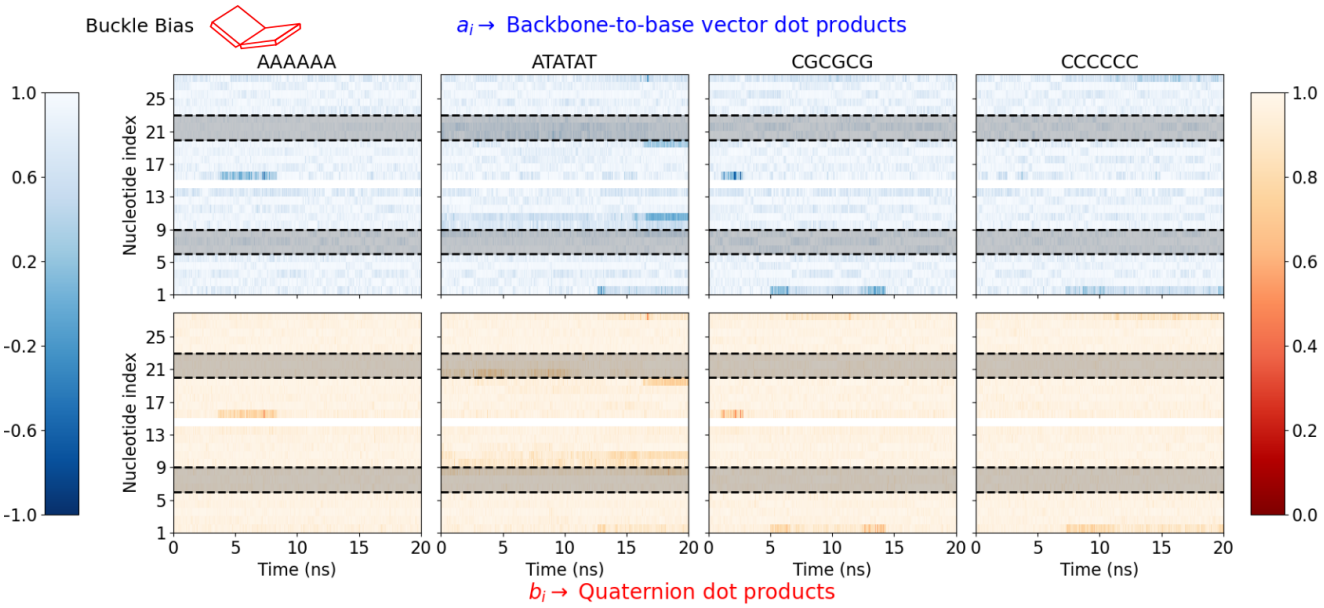


图5。随后进行1纳秒、倾斜偏置等于 $\bar{\alpha} = 0.3$ 弧度的20纳秒模拟所对应的运动图。

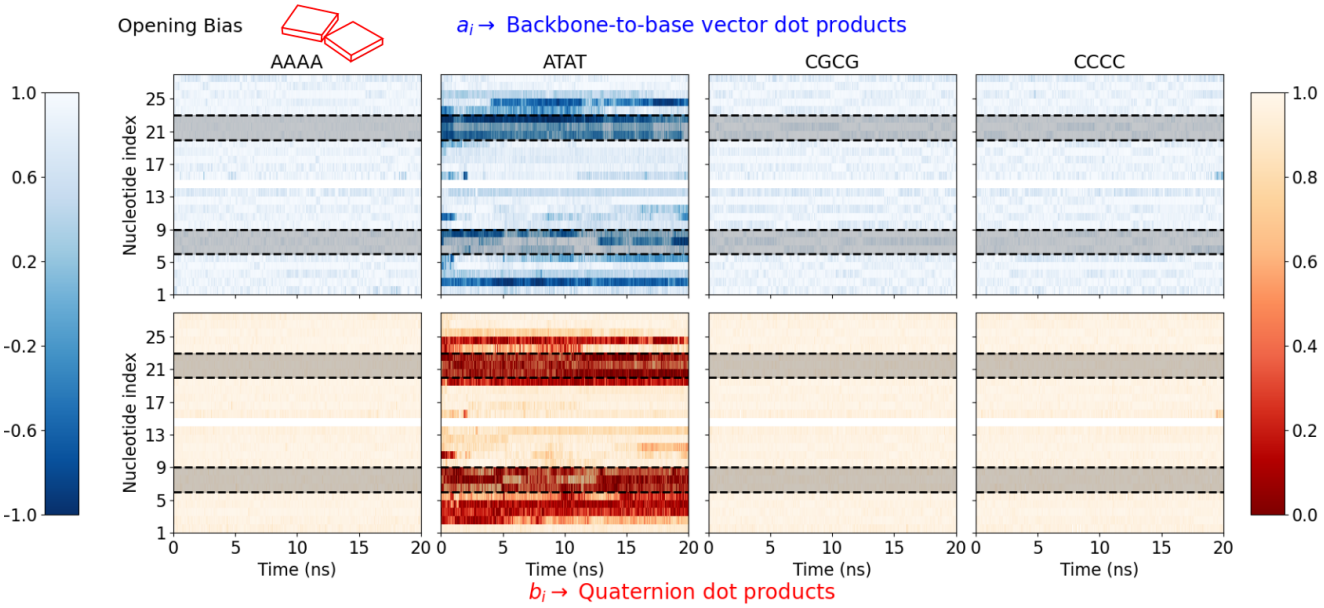


图6。与 随后进行1纳秒、开口偏置等于 $\bar{\alpha} = 0.42$ 弧度的20纳秒模拟所对应的运动图。

图 6 展示的则是分子 中间开口偏置对应的运动图。这种偏置会产生更多

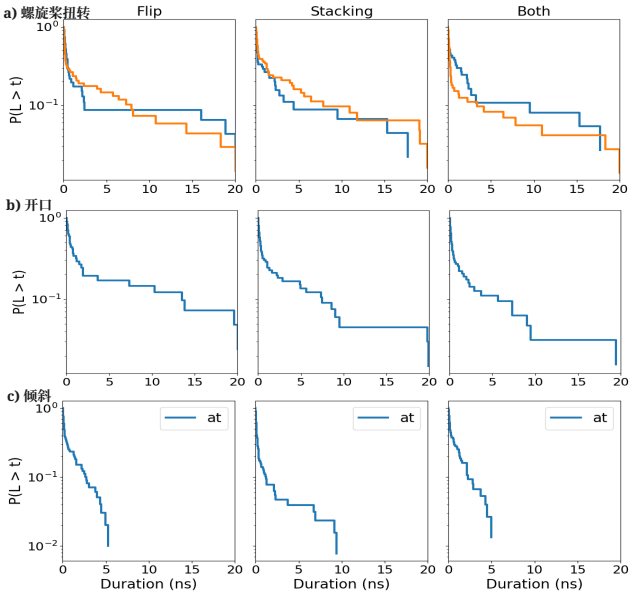


图7。翻转缺陷（左）、堆叠缺陷（中）以及翻转+堆叠缺陷（右）的存活概率。三行分别展示了在螺旋桨扭转（顶行）、开口（中行）和倾斜（底行）上施加偏置后进行模拟所得的结果。

缺陷，与螺旋桨扭转的情况类似。同样，PolyAT链的变形更大，而堆叠缺陷是最常见的变形类型。

正文中螺旋桨扭转偏置的结果表明，PolyAT 比 PolyCG 更具顺应性，因为前者更容易形成孤立缺陷，而后者则倾向于应力积累，从而导致更大尺度的结构缺陷。相反，在施加开口偏置时，形成更多缺陷的是 PolyAT：我们认为，这是因为 AT 之间由两个而非三个氢键保持在一起，而开口与这些氢键的破坏耦合得更强。另一方面，螺旋桨扭转会破坏更多堆叠，而倾斜对氢键和堆叠的影响都较小，因此更容易被 DNA双螺旋吸收。因此，我们的结果表明，缺陷形成取决于序列和所施加的扰动，其关系并非简单直接。

DNA缺陷的隐马尔可夫模型

在图4中，使用隐马尔可夫模型（HMM）算法估算了两种状态的占比。该模型发现，螺旋桨扭转偏置下翻转缺陷的平均值为 $a_i = \cos\theta = 0$ ，而与热涨落相关的平均值为$\cos\theta = 0.8$。根据堆叠序列，某一步的 b_i 特征值约为：AT堆叠为$b_i = 0.36$ ，CG堆叠为0.20，而规范（正常）构型为$0.6 - 0.8$ 。我们使用Python中的高斯隐马尔可夫模型（HMM）识别了这两种状态，该模型通过hmmlearn库中的GaussianHMM类实现。首先，我们测试了具有2、3和4个隐状态的隐马尔可夫模型。在大多数情况下，似然得分表明，两状态模型已足以捕捉底层动力学。特别是，对于所有堆叠缺陷以及PolyAT中的所有翻转缺陷，三状态模型得到的似然与两状态模型基本相同，表明额外的隐状态并无必要。唯一的例外是PolyCG在螺旋桨扭转偏置下的翻转缺陷。在这种情况下，三状态模型的似然得分显著更高，表明存在三个不同的隐状态，其平均值约为$\cos\ \theta = -0.2$、$\cos\ \theta = -0.5$和$\cos\theta = 0.9$，后者对应于平衡构型周围的热涨落。然而，为了保持所有分析的一致性，我们始终采用两状态模型，分别对应于缺陷的存在或不存在。模型分别基于PolyAT和PolyCG的6个中心碱基对的翻转值和堆叠值进行训练，覆盖完整的20纳秒轨迹，并每隔10 ps取一帧。随后，我们根据模型结果为每个观测值分配状态，生成电报图，并进一步确定转移概率矩阵。平均值较低的状态被确定为缺陷态，而平均值较高的状态为无缺陷态。根据自转移概率 P_{ii} ，我们计算了各状态的驻留时间 $\tau_i = -\Delta t/\ln(P_{ii})$，随后确定跃迁速率为 $1/\tau_i$。

存活概率和缺陷寿命

通过隐马尔可夫模型识别出翻转缺陷和堆叠缺陷后，我们便可以量化缺陷形成和退火的时间尺度。除了正文中讨论的速率之外，分析缺陷存活概率，即翻转缺陷、堆叠缺陷和复合缺陷寿命的分布函数，也很有意义。如果缺陷形成和退火是速率恒定的泊松过程，那么这些存活概率应当呈现单指数形式。然而，图 7 显示，对于所有偏置和缺陷类型，曲线都从未呈现单指数形式。这表明，缺陷动力学的弛豫发生在多个时间尺度上，这是由于相关性会形成大小不同的缺陷岛。由于缺陷岛的退火取决于其大小，因此问题中自然会出现多个时间尺度。当缺陷具有持久性时，例如在具有螺旋桨扭转偏置的PolyCG序列中，缺陷的寿命可能延续至我们的模拟时长，这与以下预期一致：它们会持续足够长的时间，从而能够被 DNA 结合蛋白或修复机制扫描到 [30]。

补充材料：DNA缺陷的几何与拓扑

在本补充材料中，我们讨论如何计算第一同伦群 $\pi_1(\mathcal{M})$，如何在缺陷分析流程中校正链曲率，以及我们采用何种协议生成原子级模拟轨迹。

序参量流形的同伦群

描述DNA碱基对的复合序参量（$\mathbf{n}(s)$ $q(s)$）位于流形 $\mathcal{M} = S^1 \times SO(3)$上，该流形是一个笛卡尔积。乘积空间的基本群可分解为各个基本群的直积，即 $\pi_1(S^1 \times SO(3)) \cong \pi_1(S^1) \times \pi_1(SO(3))$ [19]。因此，求解 $\pi_1(\mathcal{M})$的问题归结为计算 $\pi_1(S^1)$和 $\pi_1(SO(3))$。

首先，我们可以通过 S^1中的一个角度来参数化骨架到碱基向量。该向量垂直于包含碱基的链，在这里我们将该角度记为 $\phi(s)$。闭合环路 γ 由整数绕数 [19],进行分类

$$m = \frac{1}{2\pi} \oint_\gamma d\phi, \tag{S1}$$

因此， $\pi_1(S^1) \cong \mathbb{Z}$。

类似地，对于描述碱基三元组的旋转群 $SO(3)$，单位四元数 $q \in S^3$ 对 $SO(3)$ 构成双覆盖（其中q 和 $-q$ 表示同一旋转） [19]。由于 S^3 是单连通的，绕固定轴旋转 2π 一周会在 S^3 中描绘一条从 $+1$ 到 -1 的路径，因此在 $SO(3)$ 中定义了一个不可缩环路；而旋转 4π 两周则会回到 $+1$，并且可以收缩为一点。因此， $\pi_1(SO(3)) \cong \pi_1(S^3/\mathbb{Z}_2) \cong \mathbb{Z}_2$。

结合这两个因子可得

$$\pi_1(\mathcal{M}) \cong \mathbb{Z} \times \mathbb{Z}_2, \tag{S2}$$

其同伦类由 $(m,\ n) \in \mathbb{Z} \times \mathbb{Z}_2$ 标记。尽管理论上 m 是一个整数绕数，但实际上，一个翻转缺陷对应于骨架到碱基向量的两次局部 π 反转，因此，在应用于 DNA缺陷时，m 只能取 0（无翻转缺陷）或 1（有翻转缺陷）。

BISHOP 标架与用于缺陷识别的曲率校正

Frenet 标架与 Bishop 标架之间的关系

Bishop 标架（亦称平行移动标架）沿表示 DNA骨架轨迹的三维空间曲线 $\mathbf{r}(s)$ 提供一个连续移动的参考系。与标准 Frenet- Serret 公式不同，后者在局部曲率为零（$\kappa = 0$）或发生拐点时会出现几何奇点；Bishop 标架通过对法平面施加平行移动条件，将内部扭转降至最低。

设轨迹为 the DNA骨架的轨迹以其弧长 s 为参数，从而得到单位切向量场：

$$\mathbf{T}(s) = \frac{d\mathbf{r}}{ds} \tag{S3}$$

Bishop 标架定义了两个法向量 $\mathbf{N}_1(s)$ 和 $\mathbf{N}_2(s)$，它们张成垂直于$\mathbf{T}(s)$的平面，由此构成右手正交归一基 $\{\mathbf{N}_1,\mathbf{N}_2,\mathbf{T}\}$。该系统的典型特征是，其法向量场不会绕切向量场旋转（相对挠率为零）；从数学上说，基向量沿 DNA链的曲线坐标变化满足以下微分方程：

$$\frac{d\mathbf{N}_1}{ds} = -\kappa_1 \mathbf{T}, \quad \frac{d\mathbf{N}_2}{ds} = -\kappa_2 \mathbf{T}, \quad \frac{d\mathbf{T}}{ds} = \kappa_1 \mathbf{N}_1 + \kappa_2 \mathbf{N}_2 \tag{S4}$$

其中， κ_1 和 κ_2 是法曲率向量场的分量。

为追踪内部分子取向随空间的变化，通过罗德里格旋转公式将基三元组矩阵 $R = [\mathbf{e}_1 \,|\mathbf{e}_2 \,|\mathbf{e}_3] \in SO(3)$ 映射为单位四元数 $\mathbf{q} = (x, y, z, w)$。旋转矩阵的列对应正交单位向量 $\mathbf{e}_1$,$\mathbf{e}_2$, $\mathbf{e}_3$:

$$\mathbf{e}_1 = \begin{bmatrix} 1-2(y^2+z^2) \\ 2(xy+wz) \\ 2(xz-wy) \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 2(xy-wz) \\ 1-2(x^2+z^2) \\ 2(yz+wx) \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 2(xz+wy) \\ 2(yz-wx) \\ 1-2(x^2+y^2) \end{bmatrix} \tag{S5}$$

This 参数化可以以数值上高效的方式跟踪局部取向 y.

四元数与三元组之间的关系

单位四元数和三元组具有相同数量的自由度（3），在物理上也是等价的。给出一些具体公式来实现三元组与四元数表示之间的转换会很方便，因为四元数在数值计算上更易处理，而三元组在物理意义上更直观。

为了表示单位四元数 $\mathbf{q}$ 关于碱基三元组向量 $\{\mathbf{e}_1$,$\mathbf{e}_2$,$\mathbf{e}_3\}$，我们注意到它对应于一次绕单位向量的空间旋转绕轴 $\hat{}\mathbf{u}$ 旋转物理旋转角 θ，使得

$$\mathbf{q} = \cos\left(\frac{\theta}{2}\right) + \sin\left(\frac{\theta}{2}\right)(u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k}), \tag{S6}$$

其中， $\theta \in [0, \pi]$ 表示（实验室）竖直基底轴 $\mathbf{k}_{\text{lab}}$ 与局部基堆叠法向量 $\mathbf{e}_3$之间的倾角。向量 $\hat{}\mathbf{u} = (u_x, u_y, u_z)$ 是垂直于由 $\mathbf{k}_{\text{lab}}$ 和 $\mathbf{e}_3$，张成平面的单位旋转轴，而 $\mathbf{i}$、$\mathbf{j}$ 和 $\mathbf{k}$ 表示四元数代数的虚单位（$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$）。

为了直接从测得的结构数据中提取局部取向坐标，而无需计算中间旋转几何性质，正交归一基三元组向量的分量也可以显式映射到公式（S6）中的系数。设局部法向和面内三元组轴在实验室坐标系中的笛卡尔分量分别为，其中 $\mathbf{e}_3 = (e_{3x}, e_{3y}, e_{3z})$、$\mathbf{e}_1 = (e_{1x}, e_{1y}, e_{1z})$ 和 $\mathbf{e}_2 = (e_{2x}, e_{2y}, e_{2z})$。描述该瞬时局部分子坐标系的唯一单位四元数 $\mathbf{q}$ 随后可表示为

$$\mathbf{q} = \frac{1}{2\sqrt{1+e_{3z}}}\left[(1+e_{3z}) + (-e_{3y})\mathbf{i} + (e_{3x})\mathbf{j} + (e_{1y}-e_{2x})\mathbf{k}\right] \tag{S7}$$

其中， e_{3x}、 e_{3y}和 e_{3z} 表示碱基堆叠法向量的空间投影，用于描述核苷酸碱基平面的倾斜。沿$\mathbf{k}$分量的最后一项涉及标量差值（$e_{1y} - e_{2x}$），它反映了平面内向量$\mathbf{e}_1$ 和$\mathbf{e}_2$ 绕局部堆叠轴发生的局部内部方位旋转，即扭转角。这种代数耦合将物理三元组位置与四元数描述联系起来。

离散化计算

在实际应用中，需要将连续平行移动方程沿连续排列的离散核苷酸位置进行离散化。令 $B_i = [\mathbf{n}_i \,|\,\mathbf{b}_i \,|\,\mathbf{t}_i]$为在碱基 i处追踪 Bishop 标架的 3×3正交矩阵。要将该标架传播到步骤 $i+1$，切向量场从$\mathbf{t}_i$ 演化至$\mathbf{t}_{i+1}$。使$\mathbf{t}_i$ 与$\mathbf{t}_{i+1}$ 对齐所需的最小刚性旋转矩阵 Ω_i 通过罗德里格旋转公式计算:

$$\Omega_i = \mathbf{I} + \sin(\theta_i)\mathbf{K}_i + (1-\cos(\theta_i))\mathbf{K}_i^2 \tag{S8}$$

其中，旋转角由内积 $\theta_i = \arccos(\mathbf{t}_i \cdot \mathbf{t}_{i+1})$定义，而$\mathbf{K}_i$ 表示由单位旋转轴 $\hat{}\mathbf{v}_i = (\mathbf{t}_i \times \mathbf{t}_{i+1})/\|\mathbf{t}_i \times \mathbf{t}_{i+1}\|$生成的反对称叉积矩阵:

$$\mathbf{K}_i = \begin{bmatrix} 0 & -\hat{v}_{z,i} & \hat{v}_{y,i} \\ \hat{v}_{z,i} & 0 & -\hat{v}_{x,i} \\ -\hat{v}_{y,i} & \hat{v}_{x,i} & 0 \end{bmatrix} \tag{S9}$$

离散化平行移动步进更新随后通过线性映射得到解析

$$B_{i+1} = \Omega_i B_i \tag{S10}$$

用于分离翻转和堆叠缺陷的三元组投影

设 $R_{i,\ \mathrm{raw}} = [\mathbf{e}_{1,\ i} \mid \mathbf{e}_{2,\ i} \mid \mathbf{e}_{3,\ i}]$ 为在实验室参考坐标系中计算得到的原始碱基三元组向量矩阵。为将纯粹的局部构象变形与整体结构的弯曲或缠绕数区分开来，将原始碱基轴动态投影到由平行移动的骨架向量生成的移动参考系中：

$$R_{i,\mathrm{corr}} = B_i^T R_{i,\mathrm{raw}} = \begin{bmatrix} \mathbf{n}_i \cdot \mathbf{e}_{1,i} & \mathbf{n}_i \cdot \mathbf{e}_{2,i} & \mathbf{n}_i \cdot \mathbf{e}_{3,i} \\ \mathbf{b}_i \cdot \mathbf{e}_{1,i} & \mathbf{b}_i \cdot \mathbf{e}_{2,i} & \mathbf{b}_i \cdot \mathbf{e}_{3,i} \\ \mathbf{t}_i \cdot \mathbf{e}_{1,i} & \mathbf{t}_i \cdot \mathbf{e}_{2,i} & \mathbf{t}_i \cdot \mathbf{e}_{3,i} \end{bmatrix} \tag{S11}$$

由于变换矩阵 B_i 满足正交条件 $B_i^T = B_i^{-1}$ ，该投影表示刚性空间坐标旋转。R_i的列向量，$_{\mathrm{corr}}$ 给出了变换后的局部碱基三元组方向，记作$\mathbf{e}_{k\,i}^{\mathrm{corr}}$，。

随后计算用于追踪局部翻转或堆叠缺陷且与路径曲率无关的几何量，具体如下：

$$\mathbf{e}_{k,i}^{\mathrm{corr}} \cdot \mathbf{e}_{k,i+1}^{\mathrm{corr}} = \left(B_i^T \mathbf{e}_{k,i}\right) \cdot \left(B_{i+1}^T \mathbf{e}_{k,i+1}\right) \tag{S12}$$

强烈的局部取向变形（例如碱基翻转或尖锐的滚转步缺陷）会触发$\mathbf{e}_k^{\mathrm{corr}}{}_{,\ i}$ ·$\mathbf{e}_k^{\mathrm{corr}}{}_{,\ i+1}$的局部下降，从而指示哪个分子轴是相位奇点的成因。

具有弹性变形和缺陷的DNA的GROMACS模拟

使用带偏置版本的cgDNA+ [38]生成了规则的14聚体B型DNA，其由6个居中的聚腺嘌呤、PolyAT、聚胞嘧啶和PolyCG序列组成，两侧各由4个CG碱基对包围。序列列表见表 SI。偏置势施加于正文中讨论的三个参数：螺旋桨扭转、倾斜和开口。此处，我们对cgDNA+ 施加偏置，以生成局部变形的输入构象，并预期这些构象将接近随后在带偏置的全原子模拟中生成的构象。之后使用x3DNA软件生成结构。在此阶段，根据所选参数 α （螺旋桨扭转、开口或倾斜）施加谐偏置势，具体如下：

$$V_\alpha = \frac{K}{2}(\alpha - \bar{\alpha})^2 \tag{S13}$$

其中， $K = 1000\frac{KJ}{mol}$ 是力常数， $\bar{\alpha}$ 是集体变量的目标值。

Name	序列
序列1 聚腺嘌呤	CGCGAAAAAACGCG
序列2 PolyAT	CGCGATATATGCCG
序列3 PolyCG	CGCGCGCGCGCGCG
序列4 PolyC	CGCGCCCCCCCCGCG

表 SI. 列表　　模拟的序列。偏置势施加于以粗体显示的两个中心核苷酸上　　.

全原子体系准备

随后，按照 [13], 此处简要总结的标准分子动力学（MD）方案对每个系统进行平衡。偏置过程使用 2022.5 版本的 GROMACS 和 2.8.3 版本的 PLUMED 完成。所使用的力场为 Amber99 parmbsc1 [39, 40]。将 14聚体序列置于十二面体盒中，并在聚合物各侧留出 2.0 nm 的额外空间。考虑了周期性边界条件，并使用 TIP-3P 水模型。 [41]。粒子网格 Ewald（PME） [42] 模型调节静电相互作用，非键相互作用的截断距离为 1 nm。用 150 mM NaCl 溶液填充盒子，随后中和体系的总电荷。采用包含 LINCS 约束的蛙跳积分器，并使用时间步长$\Delta t = 2fs$ 进行能量最小化，直至容差达到 1000 kJ/mol。随后进行 100 ps 的 NVT 平衡，将温度固定在 300 K；最后使用 Parrinello-Rahman 压强耦合器进行 100 ps 的 NPT 模拟，使压力达到 1.0 bar。

参数	目标值 $\bar{\alpha}$ [弧度]	力常数 K $[kJ/mol]$
螺旋桨扭转	0.30	300
开口	0.42	300
倾斜	0.3	700

表 SII。RBB-NA 全原子分子动力学模拟中引入的偏置参数列表。

RBB-NA：偏置全原子模拟与系统弛豫

平衡后，使用 RBB-NA 算法 [23]对十二聚体进行增强采样。RBB-NA 从通过 cgDNA+ 程序生成的构型开始运行。对于表 SII 中报告的集体变量选定值，使用 2 fs 的时间步长进行 1 ns 的分子动力学模拟。表中报告的目标值 $\bar{\alpha}$随具体偏置参数而变化，但在 4 种不同序列中保持不变。在所有情况下，我们都选择了较弱的 k 谐和约束。这样做是为了避免过强偏置力可能引起的 LINCS 约束不稳定性，因为这可能导致模拟中断。随后将偏置模拟的终点用作 20 ns 无偏模拟的起始点，时间步长为 $\Delta t = 2$ fs；其结果将在正文中进行研究。