Biophysical Journal, Volume *115*

Supplemental Information

Environmentally Controlled Curvature of Single Collagen Proteins

Nagmeh Rezaei, Aaron Lyons, and Nancy R. Forde

Derivation of the curved Worm-Like Chain (cWLC) model

The standard WLC model assumes that there is an energetic cost to bend an intrinsically straight rod. Specifically, to bend a semi-flexible, intrinsically straight rod with length δl into a circular arc with central angle $\delta\theta$, the energy required is (1)

$$E_{\text{bend}} = \frac{\alpha\delta\theta^2}{2\delta l} = \frac{p\delta\theta^2}{2\delta l}k_{\text{B}}T. \tag{S1}$$

α is the bending rigidity of the chain, related to the persistence length p through the relation $p = \alpha/k_{\text{B}}T$, where $k_{\text{B}}T$ is the product of the Boltzmann constant and the absolute temperature. The central angle of this arc, $\delta\theta$, is equivalent to the angle between the tangents at the beginning and end of the segment. The lowest energy conformation of the segment is a straight line, with the energy increasing harmonically about this configuration. The Boltzmann distribution provides the distribution of angles that this segment adopts in the presence of thermal noise, which is given by

$$P(\delta\theta) = \sqrt{\frac{p}{2\pi\delta l}}\exp\left(-\frac{p(\delta\theta)^2}{2\delta l}\right). \tag{S2}$$

This a normal distribution with mean $\langle\delta\theta\rangle = 0$ and variance $\sigma_{\delta\theta}^2 = \frac{\delta l}{p}$.

If the chain is intrinsically curved, the bending energy is modified to reflect deviations from this bent state.

$$E_{\text{bend}} = \frac{p(\delta\theta - \kappa_o\delta l)^2}{2\delta l}k_{\text{B}}T, \tag{S3}$$

where κ_o is the intrinsic curvature of the chain, defined as $\kappa_o = \langle\frac{d\theta}{dl}\rangle$. The distribution of angles that this intrinsically curved segment adopts in the presence of thermal noise is given by

$$P(\delta\theta) = \sqrt{\frac{p}{2\pi\delta l}}\exp\left(-\frac{p(\delta\theta - \kappa_o\delta l)^2}{2\delta l}\right). \tag{S4}$$

As before, this is a normal distribution with variance $\sigma_{\delta\theta}^2 = \frac{\delta l}{p}$, but now centered about a mean $\langle\delta\theta\rangle = \kappa_o\delta l$.

We now consider a longer chain segment, comprised of N segments each of length δl, such that the total length of this chain $l = N\delta l$. This longer chain is not necessarily a circular arc, but is made up of N short circular arc segments, each with angular difference $\delta\theta_i$ distributed about $\kappa_o\delta l$ as per equation (S4). The angular difference θ between the ends of this larger segment is given by the sum of the N angles adopted by the smaller circular arcs:

$$\theta = \sum_{i=1}^{N}\delta\theta_{\text{i}}. \tag{S5}$$

Since the sum of normal random variables is normally distributed, the full distribution of θ is completely described by its mean and variance, given respectively by

$$\langle\theta\rangle = N\langle\delta\theta\rangle = \kappa_o N\delta l = \kappa_o l \tag{S6}$$

and

$$\sigma_\theta^2 = N\sigma_{\delta\theta}^2 = \frac{N\delta l}{p} = \frac{l}{p}. \tag{S7}$$

The angular distribution of this segment of length l is therefore given by

$$P(\theta) = \sqrt{\frac{p}{2\pi l}}\, e^{-\frac{p(\theta-\kappa_o l)^2}{2l}}. \tag{S8}$$

Note that, because $\langle\theta\rangle \neq 0$, this chain exhibits a net global curvature.

For any arbitrary segment length s, we can now calculate the average tangent vector correlation as

$$\langle\hat{t}(s+s')\cdot\hat{t}(s')\rangle = \langle cos\theta(s)\rangle = \sqrt{\frac{p}{2\pi l}}\int_{-\infty}^{\infty}\exp\left(-\frac{p(\theta-\kappa_o s)^2}{2s}\right)\cos\theta\, d\theta, \tag{S9}$$

where s' defines an arbitrary starting position of the segment along the contour. This simplifies to

$$\langle cos\theta\rangle = \exp\left(-\frac{s}{2p}\right)\cos(\kappa_o s). \tag{S10}$$

However, since $\hat{t}(s+s')\cdot\hat{t}(s') \leq 1$ always, then its average $\langle\hat{t}(s+s')\cdot\hat{t}(s')\rangle \leq 1$. This requires that $s \geq 0$, so we can write

$$\langle cos\theta(s)\rangle = \exp\left(-\frac{|s|}{2p}\right)\cos(\kappa_o|s|) \tag{S11}$$

for all s.

Similarly, we can calculate the mean squared end-to-end distance of the segment as

$$\langle R^2(s)\rangle = \langle\vec{R}(s)\cdot\vec{R}(s)\rangle = \left\langle\left(\int_0^s\hat{t}(s')ds'\right)\left(\int_0^s\hat{t}(s'')ds''\right)\right\rangle = \int_0^s\int_0^s\langle\hat{t}(s')\cdot\hat{t}(s'')\rangle ds'ds'', \tag{S12}$$

where the order of operations was interchanged because averaging and integrating are both linear operations. Using $\langle\hat{t}(s')\cdot\hat{t}(s'')\rangle = \exp\left(-\frac{|s''-s'|}{2p}\right)\cos(\kappa_o|s''-s'|)$ from equation (S11) gives

$$\langle R^2(s)\rangle = \int_0^s\left\{\int_0^{s''}\exp\left[-\frac{(s''-s')}{2p}\right]\cos[\kappa_o(s''-s')]\,ds' + \int_{s''}^{s}\exp\left[-\frac{(s'-s'')}{2p}\right]\cos[\kappa_o(s'-s'')]\,ds'\right\}ds''. \tag{S13}$$

Evaluating this expression yields

$$\langle R^2(s)\rangle = \frac{4sp}{(1+4\kappa_0{}^2p^2)^2}\left\{1-\frac{2p}{s}(1-4\kappa_0{}^2p^2)\left[1-\cos(\kappa_0 s)\exp\left(-\frac{s}{2p}\right)\right] + \frac{4\kappa_0 p^2}{s}\left[\kappa_0 s - 2\sin(\kappa_0 s)\exp\left(-\frac{s}{2p}\right)\right]\right\}. \qquad \text{(S14)}$$

Once again, this expression is valid only for $s \geq 0$, as negative values of s can produce values of $\langle R^2(s)\rangle$ that are negative. Thus,

$$\langle R^2(s)\rangle = \frac{4|s|p}{(1+4\kappa_0{}^2p^2)^2}\left\{1-\frac{2p}{|s|}(1-4\kappa_0{}^2p^2)\left[1-\cos(\kappa_0|s|)\,e^{-\frac{|s|}{2p}}\right] + \frac{4\kappa_0 p^2}{|s|}\left[\kappa_0|s| - 2\sin(\kappa_0|s|)\,e^{-\frac{|s|}{2p}}\right]\right\}. \qquad \text{(S15)}$$

In the case where $\kappa_o = 0$, which corresponds to the standard two-dimensional worm-like chain, the expressions for the tangent vector correlation and mean squared end-to-end distance reduce to

$$\langle cos\theta(s)\rangle = \exp\left(-\frac{|s|}{2p}\right) \qquad \text{(S16)}$$

and

$$\langle R^2(s)\rangle = 4|s|p\left\{1-\frac{2p}{|s|}\left[1-\exp\left(-\frac{|s|}{2p}\right)\right]\right\}. \qquad \text{(S17)}$$

These are identical to previously derived results (2), and to equations (2) and (1) in the main text, respectively.

Because both $\langle R^2(s)\rangle$ (equation S15) and $\langle cos\theta(s)\rangle$ (equation S11) are even functions of curvature κ_o, we report only the magnitude of the curvature from our fits.

SmarTrace Algorithm

From AFM images of biopolymers, we wished to analyze chain configurations and extract mechanical properties. For this purpose, we developed a robust and efficient chain tracing software in MATLAB, dubbed “SmarTrace”, which traces the centerline of each molecule with sub-pixel resolution and provides comprehensive statistical analysis of the chains.

Chain tracing

An overview of the SmarTrace workflow is discussed below. Details are available in reference (3).

1. In a visual user interface (adapted from the EasyWorm package (4) within MATLAB (5)), the user selects a few points on or near the backbone of the chain to be traced.
2. SmarTrace fits an initial spline to these user-defined points and extracts points along the spline separated by one nanometer.
3. The program uses top-hat and median filtering (6) to improve the signal-to-noise ratio of the region surrounding the chain.
4. To detect the best path describing the chain, a search window is defined for each point on the spline curve along the tangential direction of the initial spline. A search grid with sub-pixel resolution determines (interpolated) intensity values of the image for each grid point.
5. For each point on the initial spline, a template pattern is matched with each point in the grid within the search window. The template resembles the intensity pattern of a cross-section of the chain with varying widths.
6. A matching score is calculated for each possible width and location of this pattern. Cross-correlation scores are used to determine the best centerline position and width of the chain.
7. To ensure stable results, a penalty term is added for sudden changes in width and/or direction of the chain.
8. After the scores are finalized, the re-weighted maximum scores are used to extract the width and center of the chain at points spaced approximately 1 nanometer apart. A B-spline is fit to these points, resulting in a piecewise smooth polynomial that represents the polymer chain.

This method is not very sensitive to image quality and can successfully detect the chain backbone in noisy images; even if parts of the chain are slightly faded, the code is still able to trace the entire chain. The results also do not depend on where the user selects the points (validated by having different users trace the same set of experimental chains, and finding the same persistence length), and the initially selected points do not need to be located exactly on the chain centerline. These features, along with an efficient computation algorithm, make the code fast and easy to use for tracing and analyzing images of single polymers, as obtained for example by atomic force microscopy, electron microscopy or fluorescence microscopy.

Statistical analysis of chain properties

After the chains have been traced, statistical approaches are used to analyze their flexibility. First, the traced chains are partitioned into segments of varying lengths. Utilizing a method introduced in reference (7), each molecule is divided into multiple segments, with lengths drawn randomly from a pre-defined set of input values (here, 10 nm, 20 nm, 30 nm, ..., 200 nm). Once the lengths of these segments have been determined, their positions on the chain are shuffled to avoid accumulating shorter lengths towards one end of the chain. This process is repeated 50 times for each chain, allowing different regions of the chain to contribute to the statistics of different segment lengths. Within each draw, the sampled chain segments are nonoverlapping. As rare, longer segments are less useful for mathematical fitting and persistence length calculations, choosing a relatively short maximum segment length provides more samples in each bin. This method also allows for the use of partially traced chains, which is particularly helpful when ends of a molecule are not clear or chains intersect. In the current work, end regions (5 pixels ≈ 19.5 nm) of the chains were excluded from further analysis.

Several statistics are then calculated from the samples at each segment length s: the mean-squared end-to-end distance, $\langle R^2(s)\rangle$, and the mean cosine of the angle between the start and end of the segment, $\langle \cos\theta(s)\rangle$.

Validation of SmarTrace and analysis procedures

Validation with DNA images

DNA molecules of ~1 µm contour length were prepared by digestion of the pBluescript KS(+) plasmid with SspI, then purifying the longer linear, blunt-ended 2828 bp fragment. For imaging, DNA was deposited from a buffer containing 4 mM HEPES, 10 mM NaCl, 2 mM $MgCl_2$ at pH 7.4, then dried before imaging (Figure S1A). Note that Mg^{2+} is required for DNA deposition due to its negatively charged backbone; in contrast, at the pH values we have studied collagen, the protein has positively charged residues and thus does not require bridging cations for deposition.

DNA images were traced and analyzed with SmarTrace. Analysis of the angular distributions showed that they exhibit a kurtosis of 3 at all but the shortest segment length analysed (Figure S1B), consistent with a Gaussian distribution and equilibration. The mean squared end-to-end distance, $\langle R^2(s) \rangle$, is shown in Figure S1C along with a 2D worm-like chain fit (Equation 1). Figure S1D shows the experimental data for tangent vector correlation, $\langle \cos\theta\ (s) \rangle$, and the 2D WLC fit (Equation 2). A persistence length of p = 62 ± 3 nm was obtained from these analyses (Table S1), consistent with previous results (8). Statistical analysis demonstrates that DNA deposited and imaged under these conditions is better described by the standard WLC than by the curved WLC (Table S2), and with a fit to the cWLC returning $\kappa_o = 0$ and p = 62 nm, *i.e.*, reducing to the standard WLC model.

We also traced and analysed images of DNA molecules with 1 µm contour length (N=24), which were provided with the software package Easyworm (4). A persistence length of 52 ± 2 nm was obtained from these analyses, consistent with expectations for DNA (2, 4).

Validation with simulated worm-like chains

To validate the methodology used for chain tracing and analysis within SmarTrace, two-dimensional worm-like chains were simulated and converted into pseudo-AFM images. These images were then traced with SmarTrace to ensure that the algorithm was able to accurately recover the input persistence length and curvature. The workflow for the simulation is as follows:

1. The desired persistence length p, curvature κ_0, contour length L and width of the chains w – as well as the average number of chains per image – are input by the user.
2. For the first chain, an angle θ_0 between 0 and 360°, as well as two values x_0 and y_0 between 0 and 2000 nm are sampled randomly from a uniform distribution. These values represent the starting angle and initial xy-coordinates of the chain, respectively.
3. The curvature of the chain is chosen to be either κ_0 or $-\kappa_0$ with equal chance; this represents the ability of the chain to "lie down" on the mica surface with either a left-handed or right-handed curvature.
4. Using a step size $\delta s = 0.5$ nm, an angle $\delta\theta$ is sampled from a normal distribution with mean $\pm\kappa_o\delta s$ and variance $\frac{\delta s}{p}$.

5. The next point on the chain is placed at $x_1 = x_0 + \delta s \cos(\theta_1)$ and $y_1 = y_0 + \delta s \sin(\theta_1)$, where $\theta_1 = \theta_0 + \delta\theta$.
6. This process is repeated, choosing a random $\delta\theta$ as above, generating $\theta_{i+1} = \theta_i + \delta\theta$ and placing the next point on the chain at $x_{i+1} = x_i + \delta s \cos(\theta_{i+1})$ and $y_{i+1} + \delta s \sin(\theta_{i+1})$.
7. This process is terminated after n steps, where $n\delta s = L$.
8. Steps 2 through 7 are repeated for every chain in the image.
9. A 512-by-512 array of pixels is overlaid on top of the image, discretizing the image into square pixels with a side length of 3.90625 nm, the same as all experimental images used in this work.
10. The intensity of each pixel is populated by considering every point on each chain as an intensity source, which contributes an intensity

$$I(x, y) = I_o \exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2w^2}\right) \tag{S18}$$

to the pixel centered at (x, y). The intensity contributions from every point on every chain are then summed to yield the overall intensity of this pixel.
11. At this point, a 512-by-512 pixel noise matrix is generated and overlaid with the image containing the simulated chains. This noise matrix contains experimentally realistic correlated noise, obtained as described in the following subsection.
12. The matrix containing the intensities from the simulated chains is then scaled and summed with the noise matrix, with the scaling factor being chosen to visually match the signal-to-noise ratio of our experimental AFM images.
13. This noisy image is then converted to an 8-bit grayscale image, allowing it to be traced with SmarTrace.

We generated two sets of images with parameters chosen to emulate experimentally gathered AFM images of collagen: both sets were given a width parameter of 7 nm and a contour length of 300 nm – chosen to replicate AFM images of collagen. For simplicity, background noise was not included in these simulations.

The first set was generated with a persistence length of 85 nm and zero curvature. An example image and standard WLC model fits to the data (Equations 1, 2 from the main text) are shown in Figures S2A-C. Results from fits with both the WLC and cWLC are included in Table S1, in which it is clear that no intrinsic curvature is found in these chains (see also Table S2). The traced data also reproduce the Gaussian properties of the simulated chains, as shown by the kurtosis and normality of the angular distributions at different segment lengths (Figures S2D and S2E). The standard error in the kurtosis (SEK), used to generate the error bars in Figure S2D, is given by (9)

$$\mathrm{SEK} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}, \tag{S19}$$

where n is the number of observations comprising the distribution. The expected distribution shown in Figure S2E is given by Eq. S8 with $p = 85$ nm, $s = 50$ nm and $\kappa_o = 0$ nm^{-1}.

The second set of chains was generated with $p = 50$ nm and an intrinsic curvature of $\kappa_o = 0.02$ nm^{-1}; an image of these chains is shown in Figure S3A. Due to the presence of chains with both positive and negative curvature (see step 3, above), the angular distributions extracted from these curved chains will be a sum of two Gaussian distributions, with mean $\pm\kappa_o s$ and variance $\frac{s}{p}$. These two distributions may not have equal amplitude (for a small number of sampled chains), making the properties of the expected distribution difficult to calculate. However, the expected distribution of the absolute-values of the angles is given by

$$P(|\theta|) = \sqrt{\frac{p}{2\pi s}}\left[\exp\left(-\frac{p(|\theta|-\kappa_o s)^2}{2s}\right) + \exp\left(-\frac{p(|\theta|+\kappa_o s)^2}{2s}\right)\right] \qquad \text{(S20)}$$

regardless of the proportion of the two distributions. Figure S3B shows the expected distribution for $p = 50$ nm, $\kappa_o = 0.02$ nm^{-1} and $s = 50$ nm, plotted with a histogram of the angles extracted from the simulated chains at a 50 nm segment length, demonstrating good agreement between the two distributions. Figures S3C and S3D show fits of the curved WLC expressions (Equations 3, 4) to the traced data, both of which yield persistence lengths and curvatures close to the input parameters. Fitting with the standard WLC model results in an underestimate of persistence length (Table S1), and is a poorer model for these chains (Table S2).

Simulation of AFM Image Noise

In order to model to noise present in the background of our AFM images, we characterized the intensity fluctuations in typical images, then used this information to produce simulated images of chains with realistic noise.

We wished to simulate a row of the image by the vector $\vec{X} = (X_1, \dots, X_n)^T$, where each element X_i of the vector is a normally distributed random variable with mean μ_i and variance σ_i^2. Here, X_1 refers to the intensity of the leftmost pixel in the row, X_2 to the intensity of the pixel second from the left, and so on. We allowed correlation between nearby pixels within a row, *i.e.*, we allowed for $\langle X_i X_j \rangle \neq 0$. This is similar to experimental observations of correlated noise in the scanning direction. We calculated the intensity of each row by

$$\vec{X} = \boldsymbol{L}\vec{Z} + \vec{\mu}, \qquad \text{(S21)}$$

where the elements μ_i represent the average intensity of the pixels in the i^{th} column; $\vec{Z} = (Z_1, \dots, Z_n)$, where the Z_i are independently distributed normal random variables with zero mean and unit variance ($\langle Z_i Z_j \rangle = \delta_{ij}$); and $\boldsymbol{L}$ is an *n*-by-*n* matrix determined from experimental images that introduces correlations in the simulated noise.

To construct $\boldsymbol{L}$ and $\vec{\mu}$, we used three scarcely populated 2000 nm-by-2000 nm AFM images (containing fewer than 6 chains per image) and assumed the intensities to be uncorrelated between rows. μ_i was given by the average intensity of the *n* pixels in the i^{th} column. We also determined the correlation matrix $\boldsymbol{C}$, in which the correlation matrix elements $\langle c_i c_j \rangle$ represent the covariance between the intensities of columns i and j, averaged over all rows, while diagonal elements comprise the variance of intensities within each column ($c_{ii} = \langle c_i c_i \rangle = \sigma_i^2$). Because $\boldsymbol{C}$ is Hermitian and positive-definite, $\boldsymbol{L}$ can be found from $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^T$. Practically, we determine $\boldsymbol{L}$ by Cholesky decomposition of $\boldsymbol{C}$, implemented using the MATLAB function `chol(C)` (5).

Realizations of $\vec{X}$ are then generated by simulating a vector $\vec{Z}$ of normal variables with zero mean and unit variance, matrix multiplying by $\boldsymbol{L}$, and adding $\vec{\mu}$. These values are then used to populate the first row of pixel intensities in the noise matrix, after which new realizations of $\vec{X}$ are simulated to populate the subsequent rows.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) (10) was used to assess whether the standard or curved WLC better describes the data:

$$BIC = -2\ln \hat{L} + k \ln N. \tag{S22}$$

$\hat{L}$ represents the maximum likelihood, *k* the number of parameters in the model (*k* = 1 for the WLC and *k* = 2 for the cWLC), and *N* the number of data points to be fit (*N* = 20 for each data set in our study).

To determine the BIC, we need to obtain the maximum likelihood, by determining the fit parameters θ (e.g. p for the WLC or p and κ_o for the cWLC) that maximize the likelihood function $L(\theta; y_i, \sigma_i^2, f)$ for each model description *f*, given our data. We assume that our observed data are independent and sampled from normal distribution with means y_i and variances σ_i^2. In this case, the likelihood function is given by

$$L(\theta; y_i, \sigma_i^2, f) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-\left(y_i - f(s_i,\theta)\right)^2}{2\sigma_i^2}\right). \tag{S23}$$

This can be rewritten as

$$L(\theta; y_i, \sigma_i^2, f) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{\left(y_i - f(s_i,\theta)\right)^2}{\sigma_i^2}\right). \tag{S24}$$

The sum within the exponential is the chi-squared statistic

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{\left(y_i - f(s_i,\theta)\right)^2}{\sigma_i^2}, \tag{S25}$$

allowing us to further simplify the likelihood function to

$$L(\theta; y_i, \sigma_i^2, f) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma_i} \exp\left(-\frac{\chi^2(\theta)}{2}\right). \tag{S26}$$

Since χ^2 is strictly positive, and the variable parameters θ appear only within $\chi^2(\theta)$, we simply need to choose the parameters such that the chi-squared value is minimized, as this will maximize the value of L. This is precisely what has been done in the fitting, where θ have been determined to minimize $\chi^2(\theta)$. The maximum value of the likelihood function is then just

$$\hat{L}(\theta; y_i, \sigma_i^2, f) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma_i} \exp\left(-\frac{\chi^2_{min}}{2}\right). \tag{S27}$$

Now, the BIC can be determined from equation (S21):

$$BIC = \chi^2_{min} + n\ln(2\pi) + 2\sum_{i=1}^{n}\ln(\sigma_i) + k\ln N. \quad \text{(S28)}$$

Since the second and third terms are independent of the model tested, for comparison purposes we need to calculate only

$$BIC = \chi^2_{min} + k\ln N. \quad \text{(S29)}$$

The model with the lower BIC is more likely; the probability that the other model describes the data is given by

$$P_{other} = \exp\left[-\frac{1}{2}\left(\mathrm{BIC}_{\mathrm{larger}} - \mathrm{BIC}_{\mathrm{smaller}}\right)\right]. \quad \text{(S30)}$$

The results of this analysis are tabulated in Table S2 for experimental data on collagen and DNA and for simulated chains. Because often the best fit with the cWLC model results in zero curvature, χ^2_{min} is identical between the WLC and cWLC models. In this case, the values of BIC differ by ~3 due to the difference in number of model parameters *k*, and thus by equation (S30) have a probability of ~0.2 that the cWLC model better describes the data.

Table S1 - Comprehensive list of fitting parameters obtained for different collagen types and co-solute conditions. Persistence lengths, p, are given in nm, while curvature κ_0 is in units of nm^{-1}. Reported errors, Δ, represent 95% confidence intervals in the fitting parameters. χ^2_{red} values are determined by dividing the minimized χ^2_{min} values by the number of degrees of freedom, N-k, where N is the number of points in each data set (N = 20) and k is the number of model parameters (k = 1 for the WLC and k = 2 for the cWLC).

Collagen Type	Solution	Length Traced (μm)	Standard Model Fits						Curved Model Fits										
			$\langle R^2 \rangle$			$\langle \cos\theta \rangle$			$\langle R^2 \rangle$					$\langle \cos\theta \rangle$				
			p	Δp	χ^2_{red}	p	Δp	χ^2_{red}	p	Δp	κ_o	$\Delta\kappa_o$	χ^2_{red}	p	Δp	κ_o	$\Delta\kappa_o$	χ^2_{red}
Rat I	Water	14.7	64	13	*111*	53	11	*55*	117	10	0.0165	0.0007	*3.1*	87	6	0.0125	0.0005	*1.6*
Rat I	100 μM KCl	6.3	41	5	*19*	38	7	*19*	56	7	0.0150	0.0023	*6.1*	59	12	0.0134	0.0022	*7.2*
Rat I	1 mM KCl	12.4	78	12	*42*	61	10	*30*	119	9	0.0125	0.0007	*2.2*	90	6	0.0105	0.0006	*1.7*
Rat I	10 mM KCl	17.2	97	5	*5.6*	85	4	*1.9*	95	10	0.0	0.3	*6.3*	83	6	0.0	0.4	*2.2*
Rat I	100 mM KCl	19.9	123	13	*27*	118	18	*23*	121	29	0.0	4.2	*29*	118	36	0.0	100.8	*25*
Rat I	1 mM HCl	20.4	43	6	*94*	36	5	*38*	65	4	0.0175	0.0008	*3.8*	49	3	0.0145	0.0011	*3.6*
Rat I	10 mM KCl + 1 mM HCl	33.6	83	11	*70*	66	10	*57*	135	10	0.0115	0.0006	*3.1*	99	4	0.0094	0.0003	*1.2*
Rat I	100 mM KCl + 1 mM HCl	20.9	106	6	*13*	101	8	*12*	102	11	0.0	0.4	*16*	100	13	0.0	6.5	*13*
Human I	100 mM KCl + 1 mM HCl	20.7	89	4	*7.4*	84	6	*8.8*	87	7	0.0	0.2	*8.0*	88	12	0.0	0.3	*10*
Human II	100 mM KCl + 1 mM HCl	12	100	4	*3.0*	96	5	*2.2*	106	6	0.0047	0.0016	*2.1*	95	8	0.0	0.1	*2.4*
Human III	100 mM KCl + 1 mM HCl	16.7	85	2	*2.1*	83	7	*8.5*	85	4	0.0014	0.0054	*2.2*	81	11	0.0	0.3	*9.1*
Rat I	20 mM Acetic Acid	24.7	35	5	*101*	31	4	*39*	55	3	0.0195	0.0008	*3.1*	43	2	0.0161	0.0008	*2.0*
Human I	20 mM Acetic Acid	50.4	42	6	*186*	34	6	*132*	66	3	0.0180	0.0006	*4.1*	52	2	0.0166	0.0005	*2.1*
Human II	20 mM Acetic Acid	13.2	64	14	*120*	50	12	*78*	109	9	0.0180	0.0008	*3.9*	85	5	0.0150	0.0005	*1.3*
Human III	20 mM Acetic Acid	40.4	50	9	*229*	39	7	*130*	79	6	0.0181	0.0008	*8.3*	59	2	0.0160	0.0005	*1.9*
Simulated WLC	*p*=85 nm	24.5	83	5	*17*	87	2	*1.5*	84	9	0.0000	0.3635	*18*	87	4	0.0010	0.0034	*1.5*
Simulated cWLC	*p*=50 nm, κ_o=0.02 nm^{-1}	17.7	41	5	*68*	36	9	*118*	58	2	0.0178	0.0006	*1.2*	65	4	0.0192	0.0008	*2.6*
2828 bp dsDNA	See supporting text	19.6	61	1	*1.7*	62	2	*1.6*	61	3	0.0000	1.6412	*1.8*	62	4	0.0	0.6	*1.7*

Table S2. Model selection using the Bayesian Information Criterion (BIC). BIC was calculated for the straight (WLC) and curved (cWLC) models using equation (S29).

		$\langle R^2 \rangle$				$\langle \cos\theta \rangle$			
Sample	Condition	BIC straight	BIC curved	More probable?	P other?	BIC straight	BIC curved	More probable?	P other?
Rat Type-I	Water	2112	62	curved	< 1E-7	1040	36	curved	< 1E-7
Rat Type-I	100 µM KCl	366	116	curved	< 1E-7	358	136	curved	< 1E-7
Rat Type-I	1 mM KCl	804	46	curved	< 1E-7	579	37	curved	< 1E-7
Rat Type-I	10 mM KCl	109	119	straight	7E-3	39	45	straight	5E-2
Rat Type-I	100 mM KCl	516	523	straight	5E-2	444	447	straight	2E-1
Rat Type-I	1 mM HCl	1798	75	curved	< 1E-7	730	70	curved	< 1E-7
Rat Type-I	10 mM KCl + 1 mM HCl	1334	62	curved	< 1E-7	1082	27	curved	< 1E-7
Rat Type-I	100 mM KCl + 1 mM HCl	259	285	straight	2E-6	237	240	straight	2E-1
Human Type-I	100 mM KCl + 1 mM HCl	144	150	straight	5E-2	171	186	straight	6E-4
Human Type-II	100 mM KCl + 1 mM HCl	60	44	curved	3E-4	45	48	straight	2E-1
Human Type-III	100 mM KCl + 1 mM HCl	43	46	straight	2E-1	165	169	straight	1E-1
Rat Type-I	20 mM Acetic Acid	1916	61	curved	< 1E-7	752	41	curved	< 1E-7
Human Type-I	20 mM Acetic Acid	3533	79	curved	< 1E-7	2519	44	curved	< 1E-7
Human Type-II	20 mM Acetic Acid	2283	76	curved	< 1E-7	1480	29	curved	< 1E-7
Human Type-III	20 mM Acetic Acid	4348	156	curved	< 1E-7	2477	41	curved	< 1E-7
Simulated WLC	p = 85 nm	332	337	straight	8E-2	31	34	straight	2E-1
Simulated cWLC	p = 50 nm κ_o = 0.02 nm^{-1}	1304	27	curved	< 1E-7	2253	52	curved	< 1E-7
2828 bp dsDNA	4 mM HEPES, 10 mM NaCl, 2 mM $MgCl_2$ (pH 7.4)	36	39	straight	2E-1	33	36	straight	2E-1

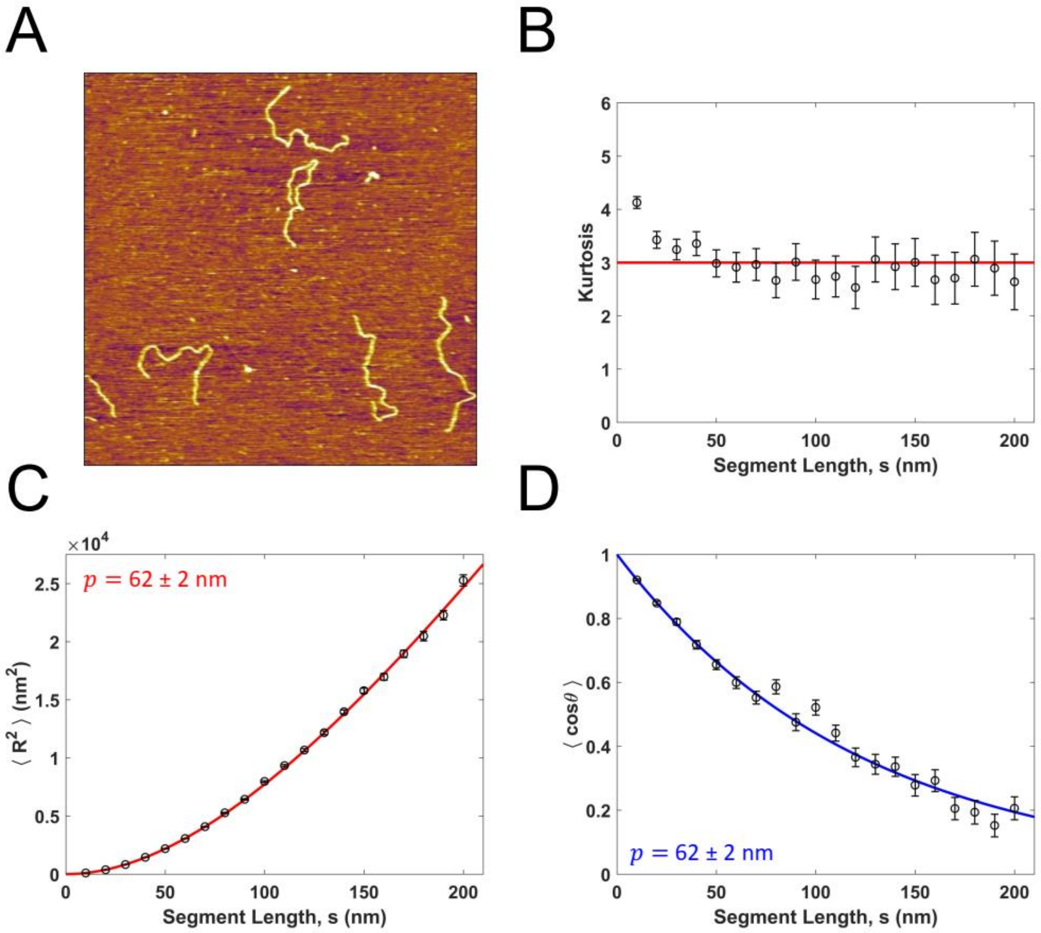

Figure S1. Quantifying the flexibility of dsDNA. (A) Example AFM image of 2828 bp linear dsDNA. (B) Kurtosis of angle distributions is close to 3 for all segment lengths, indicating that angles are normally distributed. (C) Mean squared end-to-end distance versus length along DNA molecules. Symbols represent the mean for each segment length while error bars show the standard error of the mean. The persistence length from the fit (red line; Equation 1) is 62 ± 2 nm. (D) Average tangent-tangent correlation vs. length along DNA molecules. The persistence length from the fit (red line; Equation 2) is 62 ± 2 nm. These values for persistence length are internally consistent and are in agreement with previous measurements of DNA deposited onto mica and imaged in air (8).

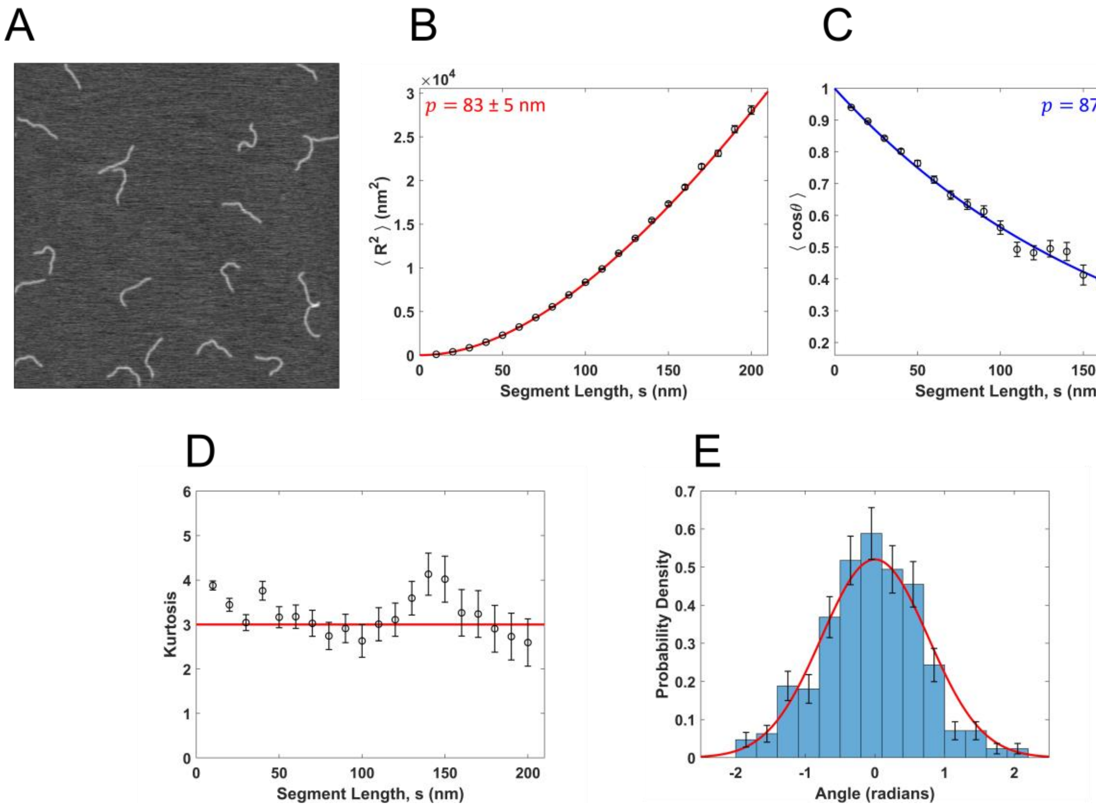

Figure S2. Test of SmarTrace analysis procedures using simulated worm-like chains. (A) Example image of simulated worm-like chains with a persistence length of $p = 85$ nm. After being traced, the data were fit with the standard WLC expressions for mean squared end-to-end distance (Eq. 1) and mean tangent vector correlation (Eq. 2), as shown in (B) and (C), respectively. Both fits yield persistence lengths consistent with the value input into the simulations. (D) Kurtosis of the angular distributions extracted from the simulated chains at different segment lengths. A kurtosis near the expected value of 3 is achieved at all length scales. Errors in the kurtosis are determined using Eq. S19. (E) Individual angular distributions are also well described as Gaussian, as shown by good agreement between the expected and extracted angular distributions at a 50 nm segment length.

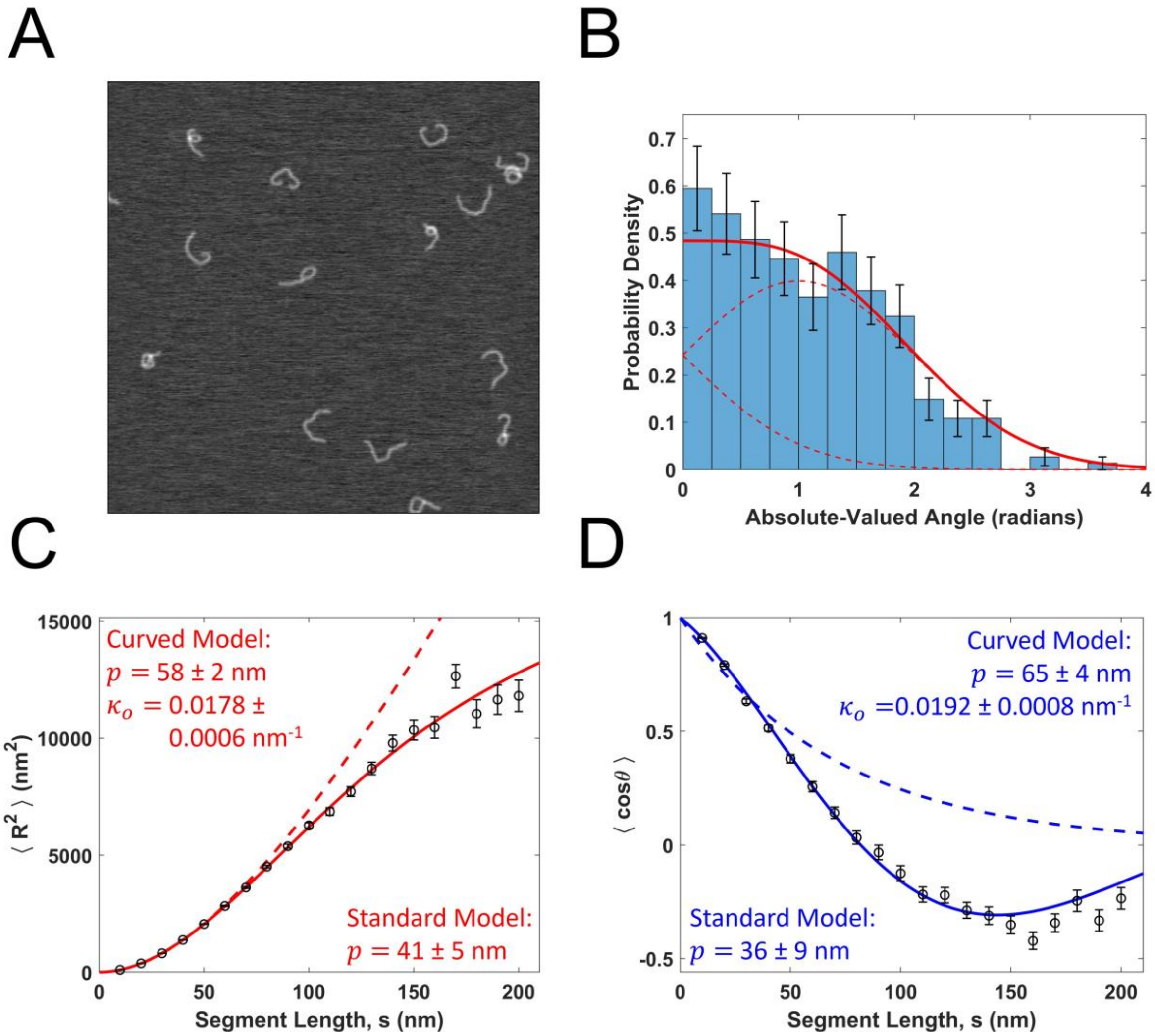

Figure S3. Analysis of simulated curved worm-like chains. (A) Example image of simulated curved worm-like chains with a persistence length of $p = 50$ nm and curvature of $\kappa_o = 0.02$ nm^{-1}. (B) Histogram of the absolute-valued angles extracted from the traces of these simulated chains at a segment length of $s = 50$ nm, as well as the expected probability distribution (solid red line, Eq. S20) and the two Gaussian components that comprise it. (C, D) Fits of data from these simulated chains to the curved worm-like chain expectations for mean squared end-to-end distance (Eq. 3) and mean tangent vector correlation (Eq. 4), respectively. Both fits yield parameters close to those input in the chain simulations.

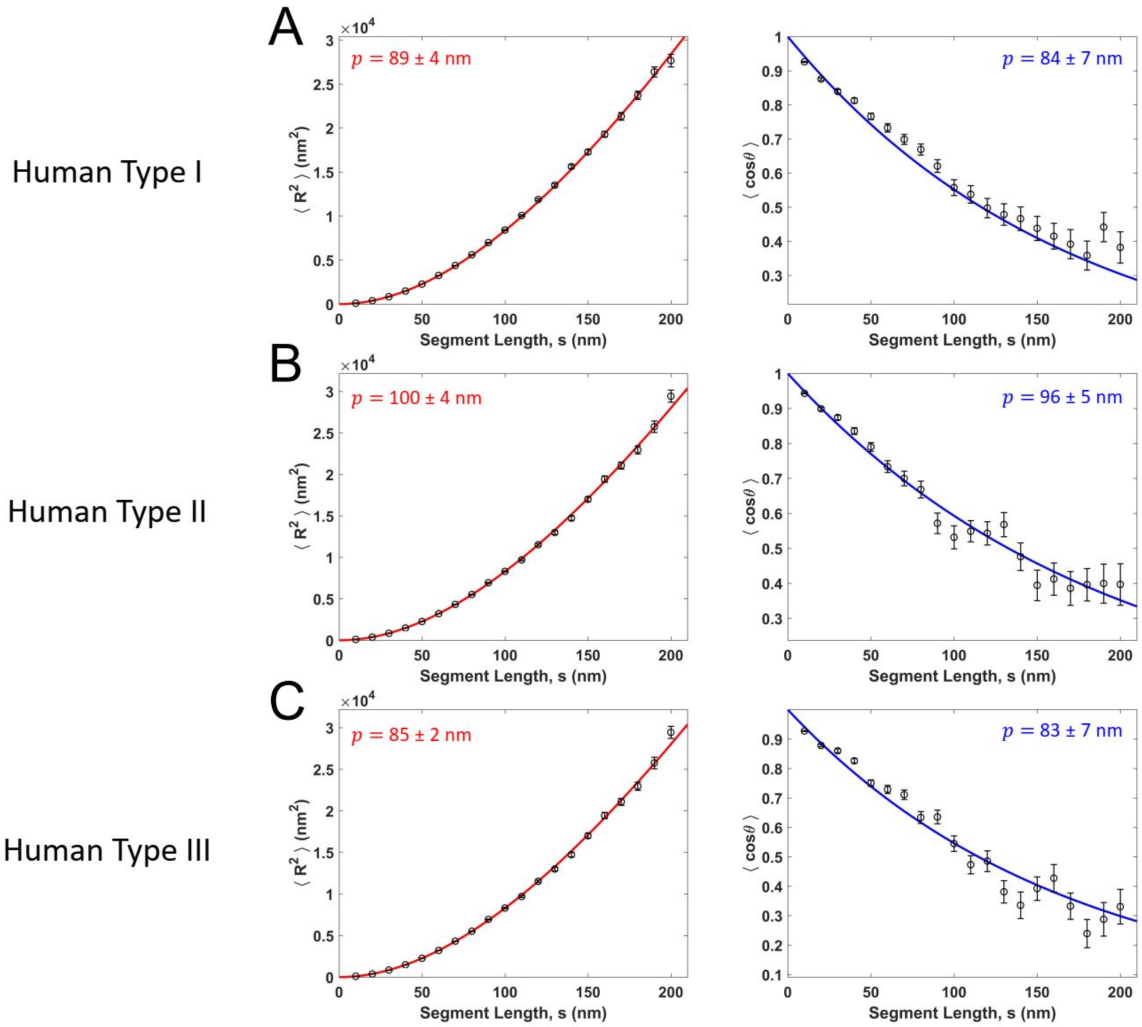

Figure S4. Standard WLC model fits to human collagens (A) type I, (B) type II, and (C) type III, all deposited from 100 mM KCl with 1 mM HCl. $\langle R^2 \rangle$ fits (Eq. 1) are shown in the left column in red, and $\langle \cos\theta \rangle$ fits (Eq. 2) are shown in the right column in blue.

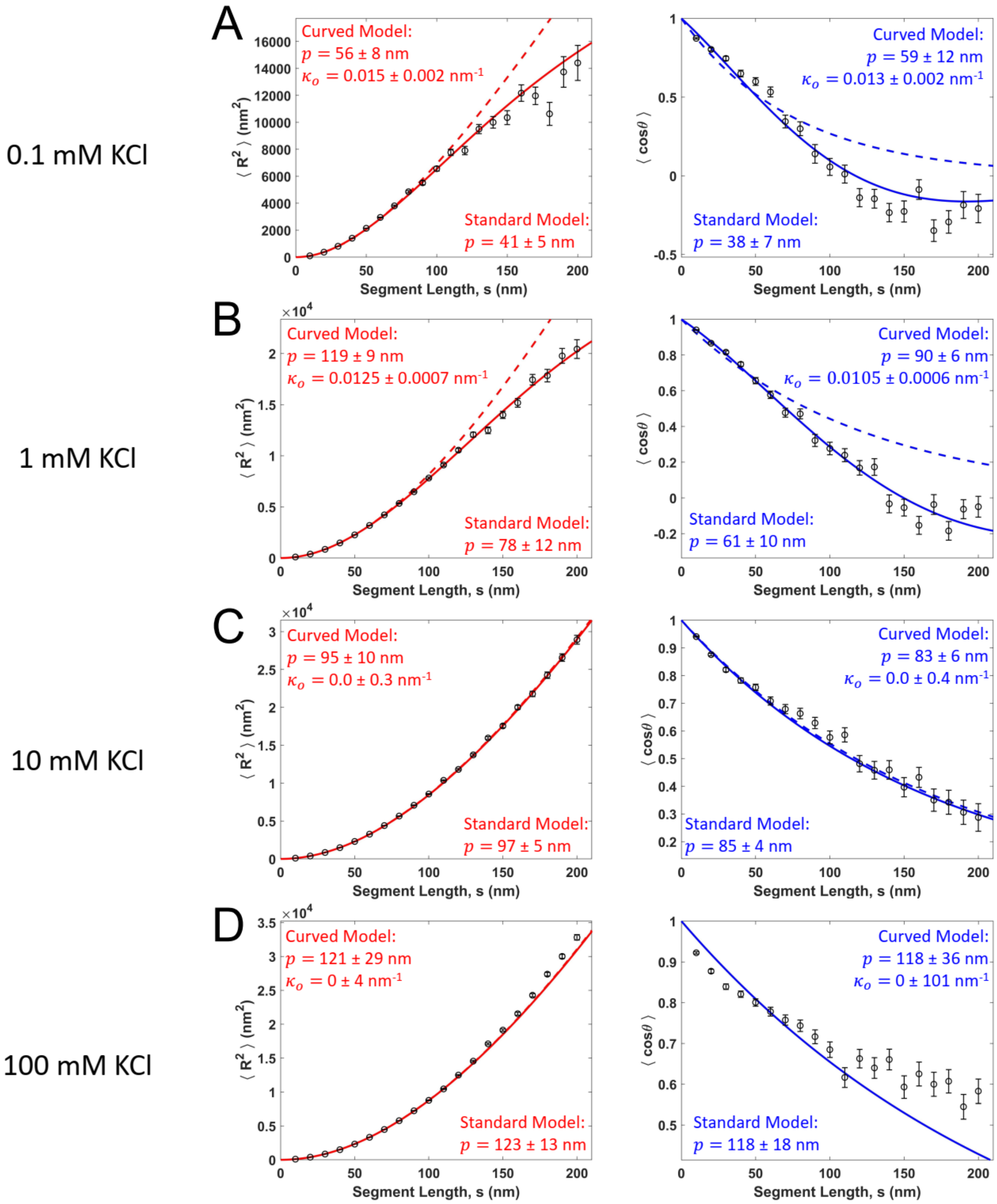

Figure S5. Standard (dashed lines) and curved (solid lines) WLC model fits to rat tail collagen type I data deposited from (A) 0.1 mM KCl, (B) 1 mM KCl, (C) 10 mM KCl and (D) 100 mM KCl. $\langle R^2\rangle$ fits are shown in the left column in red, and $\langle \cos\theta\rangle$ fits are shown in the right column in blue. As the KCl concentration is increased, the standard and curved fits converge to the same result.

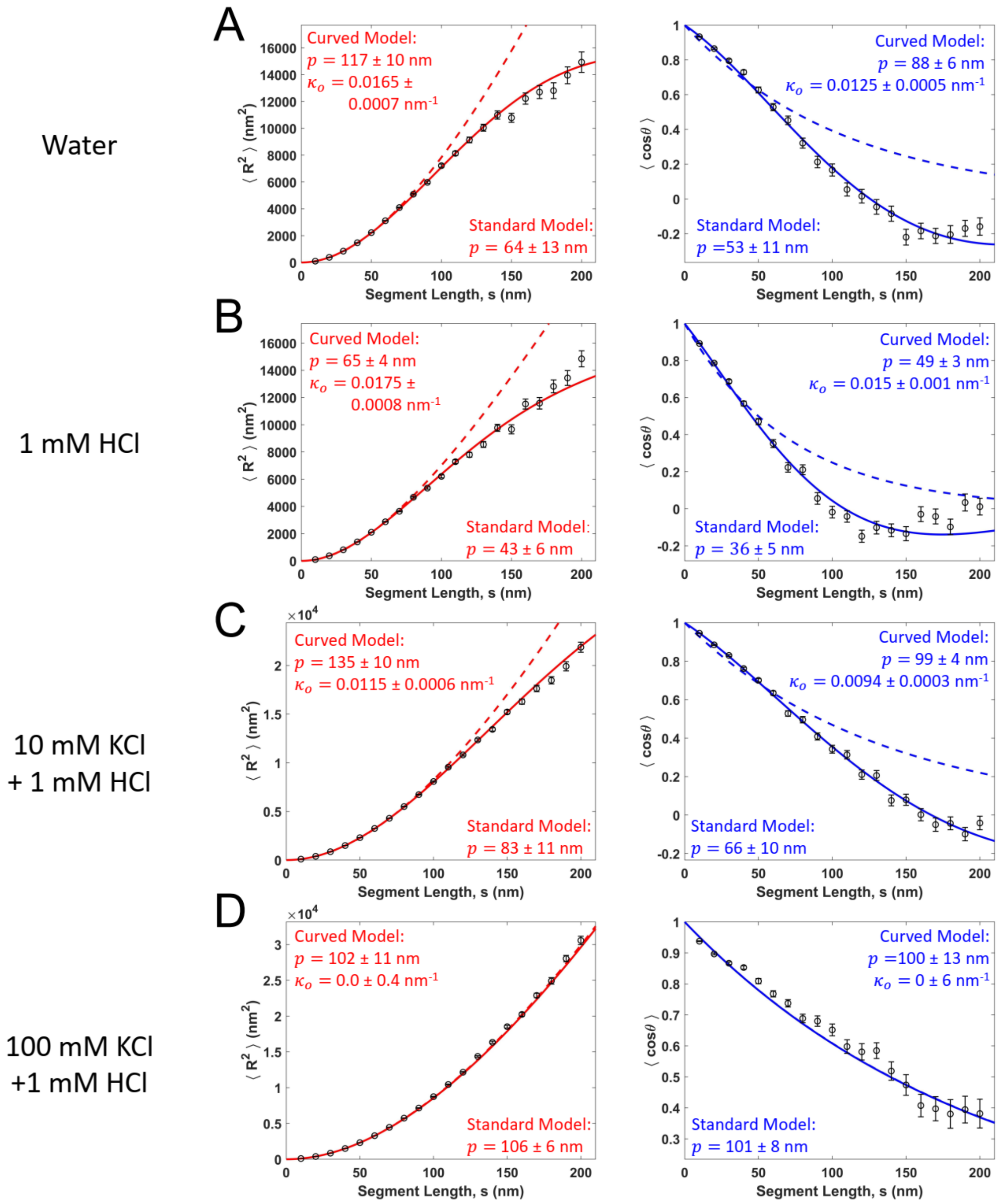

Figure S6. Standard (dashed lines) and curved (solid lines) WLC model fits to rat tail collagen type I data from (A) water, (B) 10 mM KCl + 1 mM HCl and (C) 100 mM KCl + 1 mM HCl. $\langle R^2 \rangle$ fits are shown in the left column in red, and $\langle \cos\theta \rangle$ fits are shown in the right column in blue.

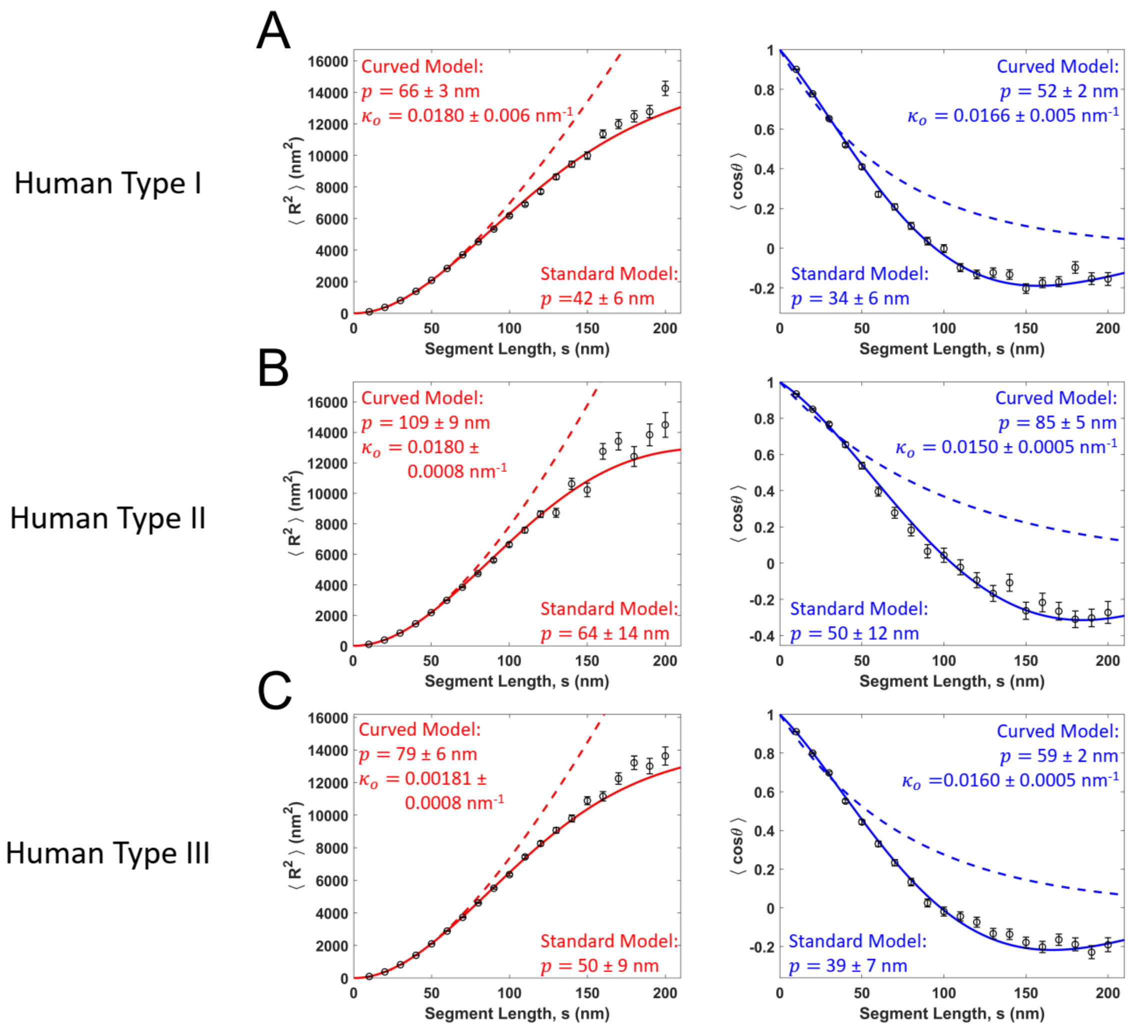

Figure S7. Standard (dashed lines) and curved (solid lines) WLC model fits to (A) type I, (B) type II and (C) type III human collagens deposited from 20 mM acetic acid. $\langle R^2 \rangle$ fits are shown in the left column in red, and $\langle \cos\theta \rangle$ fits are shown in the right column in blue.

Supporting References

1. Landau, L. D., L. P. Pitaevskii, A. M. Kosevich, and E. M. Lifshitz. 1986. Theory of Elasticity. Butterworth-Heinemann.
2. Rivetti, C., M. Guthold, and C. Bustamante. 1996. Scanning Force Microscopy of DNA Deposited onto Mica: Equilibration versus Kinetic Trapping Studied by Statistical Polymer Chain Analysis. Journal of Molecular Biology 264:919-932.
3. Rezaei, N. 2016. Mechanical Studies of Single Collagen Molecules Using Imaging and Force Spectroscopy. Simon Fraser University, Burnaby, Canada.
4. Lamour, G., J. Kirkegaard, H. Li, T. Knowles, and J. Gsponer. 2014. Easyworm: an open-source software tool to determine the mechanical properties of worm-like chains. Source Code for Biology and Medicine 9:16.
5. MATLAB and Statistics Toolbox Release 2017b. The MathWorks, Inc., Natick, Massachusetts, United States.
6. Sonka, M., V. Hlavac, and R. Boyle. 2014. Image Processing, Analysis and Machine Vision. Cengage Learning.
7. Faas, F. G. A., B. Rieger, L. J. van Vliet, and D. I. Cherny. 2009. DNA Deformations near Charged Surfaces: Electron and Atomic Force Microscopy Views. Biophysical Journal 97:1148-1157.
8. Murugesapillai, D., S. Bouaziz, L. J. Maher, N. E. Israeloff, C. E. Cameron, and M. C. Williams. 2017. Accurate nanoscale flexibility measurement of DNA and DNA-protein complexes by atomic force microscopy in liquid. Nanoscale 9:11327-11337.
9. Cramer, D. 1997. Basic statistics for social research : step-by-step calculations and computer techniques using Minitab. Routledge, London; New York.
10. Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer.