Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics

Supporting Information for:
Do monovalent mobile ions affect DNA's flexibility at high salt content?

Alexey Savelyev[1†,‡]

[†]*Institute of Applied Physics, National Academy of Sciences of Ukraine, Petropavlivska st., 58, Sumy, Ukraine, 40030*
[‡]*Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3290*

[1]Correspondence should be addressed to: alexsav.science@gmail.com

1 Preparation and parametrization of a coarse-grained DNA model

1.1 Building a model

We built our coarse-grained (CG) model of DNA by representing each DNA base-pair by two beads of the same type, where each bead is placed in the geometric center of the corresponding atomistic base-pair nucleotide (see Fig. SI **1**). The Biochemical Algorithms Library [1] was used for this purpose.

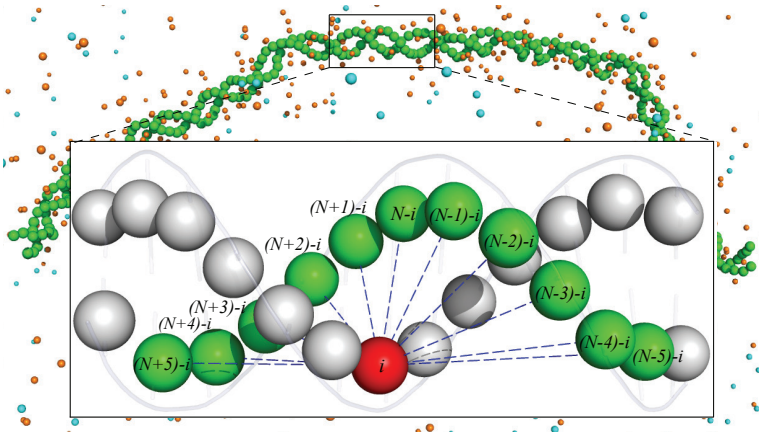

Fig. SI 1: Our recently developed chemically accurate coarse-grained model of the double-stranded DNA with explicit mobile ions [2] was extensively used in this study. Each DNA base-pair is represented by two beads, each placed in the geometric center of the corresponding atomistic nucleotide. Blue dashed lines indicate effective interactions which represent a superposition of stacking and base pairing among two polynucleotides [reffered to as "fan" interactions in Eq. (SI.1)].

As elaborated below, CG DNA model was systematically derived from the all-atom (AA) MD simulations of the corresponding atomistic systems (16-base -pair DNA oligomer in explicit solvent and salt buffer, Ref. [3]), using our recently developed Molecular Renormalization Coarse-Graining (MRG-CG) optimization technique [4, 5]. In so doing, we reproduced in CG models a number of relevant physical observables measured (in the context of MD simulations) from exact AA systems, to ensure the high fidelity of the local DNA interactions, as well as interactions among ions, and DNA and ions. The main motivation for coarse-graining is the significant reduction of the total number of degrees of freedom in original AA system by representing the atomistic nucleotide (of 25 atoms) by a single bead (thus, 2 beads per DNA base pair) and integrating out (removing) the water, while preserving all important aspects of DNA conformational behavior and system's electrostatics. In our representation, CG system was comprised of a 2-bead DNA segment and explicit mobile ions (Na^+ and Cl^-). Such a reduction, or renormalization, allowed us to increase the size of the DNA oligomer up to 150 base pairs and address the study of a large-scale conformational dynamics of DNA at a reasonable computational cost. All types of the effective interactions in CG systems (polymeric DNA, inter-ionic and ion-DNA interactions) have been derived with

MRG-CG technique from a much smaller, however, detailed AA system, a 16-base-pair oligomer. This procedure is outlined below. In this way, a larger simplified CG 150-base-pair DNA system was extrapolated from the smaller AA 16-base-pair system by a simple replication of effective DNA base pairs (beads) in space according to the DNA geometry and prescribing them effective interactions extracted from the underlying atomistic model. Next, having at hand all necessary effective potentials, one can construct CG DNA segment of an arbitrary length. For our purposes, we have chosen 150-base-pair DNA segment whose length roughly equals the DNA's persistence length. It has to be noted that our CG model is averaged over DNA's sequence, and hence, further development is needed if sequence specific effects need to be investigated. In summary, to computationally estimate DNA persistence length we followed the route: Simulating small (16 b.p.) AA system $->$ Deriving the effective potentials for CG system (with MRG-CG technique) $->$ Building and simulating a CG system of necessary length (150 b.p.) $->$ Measuring large-scale characteristics (persistence length).

We used the following effective Hamiltonian to describe DNA chain interactions,

$$\mathcal{H} = \mathcal{U}_{\rm bond} + \mathcal{U}_{\rm ang} + \mathcal{U}_{\rm fan} + \mathcal{U}_{\rm el}. \tag{SI.1}$$

In this expression, the first two terms indicate bond and bending angle potential energies, respectively. While these contributions reflect connectivity of each DNA strand and represent *intra*-strand interactions, a non-standard third term (reffered to as *fan* interactions) is responsible for maintenance of the DNA double-strand formed by two polynucleotides. As shown in the Fig. SI **1**, these *inter*-strand interactions (blue dashed lines) represent a superposition of base-pairing and stacking forces. The last term in Eq.(SI.1) corresponds to the Coulomb electrostatic interactions.

To capture a non-symmetric shape of DNA structural fluctuations (anharmonicities), we have chosen the following polynomial forms for individual energetic contributions,

$$U_{\substack{\rm bond,\\ \rm fan}} = \sum_{\alpha=2}^{4} K_\alpha (l - l_0)^\alpha, \quad U_{\rm ang} = \sum_{\alpha=2}^{4} K_\alpha (\theta - \theta_0)^\alpha, \tag{SI.2}$$

where l and l_0 in the first formula are fluctuating and equilibrium interparticle separations for bond and fan interactions, respectively. θ and θ_0 play analogous roles for the angular potential in the second expression. As customary, equilibrium values l_0 and θ_0, as well as the initial set of coefficients $\{K_\alpha^{(0)}\}$, can be obtained by fitting these polynomials to the corresponding PMFs, extracted from AA MD simulations [6]. To obtain these, we analyzed the dynamics of the atomistic 32-base-pair DNA oligomer solvated in explicit water with added physiological NaCl salt buffer (see previous section).

Inter-ionic interaction potentials were taken from our prior work on coarse-graining a bulk NaCl solution [5]. Particularly, this part of CG Hamiltonian has the following functional form,

$$\mathcal{H} = \sum_{i>j} \left[\frac{A}{r_{ij}^{12}} + \sum_{k=1}^{5} B^{(k)} {\rm e}^{-C^{(k)}\left[r_{ij} - R^{(k)}\right]^2} + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right], \tag{SI.3}$$

defined by the set of parameters, $\{A, B^{(k)}, C^{(k)}\}$, and the positions of Gaussian peaks and minima, $\{R^{(k)}\}$. Five Gaussian functions were introduced to account for short-range hydration effects and

to accurately reproduce atomistic behavior of ions, while $1/r^{12}$ potential is taken to account for the core-core inter-ionic repulsion. The last term stands for Coulomb interactions.

Finally, for the normally charged DNA system, functional forms for interaction potentials among beads of DNA and the ions were derived from a separate series of AA MD simulations of a system comprised of *unconnected* DNA backbone "monomers" (sodium dimethylphosphate) and NaCl salt buffer [7]. This has been done in an attempt to single out a "typical" DNA-bead–ion interaction by suppressing correlation effects caused by DNA connectivity (effects from neighboring DNA beads). These correlation effects were later accounted for by adjusting Hamiltonian parameters with MRG-CG technique (see below). Functional form for these type of effective interactions is similar to inter-ionic potentials, however, with softer excluded volume interactions and lesser number of Gaussian functions to describe hydration effects,

$$\mathcal{H} = \sum_{i>j} \left[\frac{A}{r_{ij}^6} + \sum_{k=1}^{3} B^{(k)} \mathrm{e}^{-C^{(k)}\left[r_{ij}-R^{(k)}\right]^2} + \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r_{ij}} \right]. \tag{SI.4}$$

As in the case of inter-ionic interactions, three Gaussian functions were introduced to account for short-range hydration effects, while softer $1/r^6$ potential mimics the core-core repulsion between DNA bead and ions.

1.2 Optimizing force field parameters using MRG-CG technique

Our optimization scheme which we call Molecular Renormalization Group Coarse-Graining (MRG-CG) technique relies on the RG Monte Carlo method by Swendsen to compute critical exponents in three-dimensional Ising model [8]. The MRG-CG scheme is based on representing an effective Hamiltonian as a linear combination of N relevant dynamical observables, $\mathcal{H} = \sum_{\alpha=1}^{N} = K_\alpha S_\alpha$, whose (various order) correlation functions, $\langle S_\alpha..S_\beta \rangle$, need to be reproduced in CG system. Hence, a "conjugate field", K_α, is prescribed to each observable, playing a role of a Hamiltonian force constant, whose numerical value has to be adjusted appropriately to generate the desired system dynamics. Because of Hamiltonian linearity, it is possible to establish a mathematical connection between these conjugate fields and expectation values of dynamical observables in terms of the covariance matrix of *all* observables,

$$\Delta\langle S_\alpha \rangle = -1/(k_\mathrm{B}T) \sum_\gamma [\langle S_\alpha S_\gamma \rangle - \langle S_\alpha \rangle \langle S_\gamma \rangle] \Delta K_\gamma, \tag{SI.5}$$

where $\Delta\langle S_\alpha \rangle \equiv \langle S_\alpha \rangle_{CG} - \langle S_\alpha \rangle_{AA}$ is the difference between the expectation values of an observable, S_α, averaged over CG and AA systems, and the ΔK_γ's are corrections to trial CG Hamiltonian parameters, $\{K_\alpha^{(0)}\}$. A set of linear equations, Eq.(SI.5), is solved at each CG iteration until the convergence is reached for all observables, $\Delta\langle S_\alpha \rangle \approx 0$, $\alpha = 1..N$. In this way, the process of parameter adjustment explicitly accounts for cross-correlations among various CG degrees of freedom – a key ingredient which is responsible for high fidelity of the local CG dynamics. For example, as we iteratively adjust Hamiltonian parameters for the DNA bending angle potential, we use the information of what impact of that adjustment would have on all other CG structural degrees of freedom (for example, bond or stacking dynamical variables).

In a recent work [4], we interpreted the MRG-CG optimization technique in light of Field Theory [9]. Namely, Hamiltonian linearity allows us to interpret the CG partition function,

$$\mathcal{Z}(\{K\}) \propto \sum \exp\left[-1/(k_{\mathrm{B}}T)\sum_{\alpha=1}^{N} K_\alpha S_\alpha\right],$$

as a generating functional, whose differentiation with respect to "conjugate fields" yields the corresponding auto- and cross-correlation functions of physical observables,

$$\langle S_1 \cdots S_n \rangle \propto \frac{\delta^n \ln \mathcal{Z}}{\delta K_1 \cdots \delta K_n}. \tag{SI.6}$$

Since the optimization is aimed at matching these various order correlation functions in AA and CG systems, the whole procedure is reminiscent of the central idea of RG theory. Indeed, matching the correlation functions of relevant physical observables ensures a significant equivalence of (restricted) AA and CG partition functions, by matching various order derivatives of the free energy. Additionally, an association with RG theory is strengthened by thinking of the parameter adjustment as a "flow" in space of Hamiltonians, spanned by a set of "conjugate fields", $\{K_\alpha\}$, coupled to the corresponding observables.

1.3 Generalizing MRG-CG scheme.

It follows from the last equation that the MRG-CG method can be straightforwardly generalized by demanding to reproduce not only average values (which is a requirement of the present model) but also higher-order correlation functions of observables. Particularly, if AA and CG partition functions generate identical sets of correlation functions of the order n and less, then the value of n may be seen as a quantitative measure of similarity between two systems. For example, in the case of $n = 2$ a set of linear equations (SI.5) will be supplemented by $N(N-1)/2$ additional (and still linear) equations aimed at matching second-order correlators $\Delta\langle S_\alpha S_\beta\rangle \approx 0$. The outlined scheme may be used to extend the current 'homopolymeric' (averaged over sequence) DNA model by introducing all four types of DNA nucleotides to study sequence-dependent effects. This may be straightforwardly acheived with MRG-CG technique by reproducing not only expectation values, but also higher-order correlation functions of various structural observables associated with different monomeric types, to account for finer details of the underlying atomistic system on top of the mean fieldish picture.

1.4 CG Hamiltonian as a linear combination of dynamical observables.

It follows from the structure of our CG Hamiltonian that behavior of the system is described by a small number of observables, which may also be seen as structure-based collective order parameters. For example, according to polynomials, Eq. (SI.2), DNA bond potential energy is described by three collective observables, $S_1^{bond} = \sum_{\text{all bonds}}(l-l_0)^2$, $S_2^{bond} = \sum_{\text{all bonds}}(l-l_0)^3$ and $S_3^{bond} = \sum_{\text{all bonds}}(l-l_0)^4$. Analogously, collective modes characterizing ion-DNA interactions (ionic "shells" around DNA bead) are $S_\alpha^{\text{Gauss}} = \sum_{\text{all pairs}}\left[\mathrm{e}^{-C_\alpha(r-R_\alpha)^2}\right]$, $\alpha = 1...3$, while

the corresponding parameters $\{K_\alpha\}$ are given by the set of constants $\{B^{(k)}\}$, see Eq.(SI.4). As a result, DNA behavior is associated with $N_{\mathbf{DNA}} = 39$ of structural observables (bond, angle and fan interactions) coupled to the corresponding "conjugate fields", $\{K_\alpha\}$, while dynamics of the ionic atmosphere around DNA is described by total of $N_{\mathbf{ions}} = 2 \times 4 = 8$ observables (4 CG degree of freedom per interactions of DNA with Na^+ and Cl^- , respectively).

2 MD simulation protocol

We used the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [10] to carry out all MD simulations. We have built 3 CG systems comprised of the 150-base-pair DNA segment immersed in a NaCl salt buffer of different concentration: 0.1M, 0.5M and 1M, respectively. Initially the systems were minimized according to the standard steepest descent algorithm. Then they were heated up to 300 K during the 20 ns and subsequently equilibrated for another 20 ns in a large periodic box having dimensions $\sim 600 \times 600 \times 600$ Å. Specific linear size of simulation box was chosen to be comparable with the lenghth of a straight 150-base-pair DNA oligomer to avoid the overlapping between neighboring periodic images. Although DNA is bent most of the time, such overlapping is possible since the molecule visits all conformations in the course of MD simulation, including a nearly straight conformation. We used the canonical NVT integration scheme (Nosé-Hoover temperature thermostat) to update particles positions and velocities at each timestep [11]. The particle mesh Ewald method [12] was used to treat long-range (Coulomb) interactions. To determine the biggest timestep we can afford to simulate CG systems with no instabilities, we used the criteria of the total energy conservation, the latter being the energy of the CG system complemented by the contribution from the Nosé-Hoover Hamiltonian [13]. It appeared that it was safe to use the timesteps of up to 5 fs, so we used this upper limit in our MD simulations. Depending on system's size, the production run used for analysis varied from ~ 0.6 μs to ~ 1 μs, to ensure equilibration of ions in a large simulation box. The system's equilibration was judged upon convergence of various structural and conformational characteristics (bond, angle distributions for DNA, radial distribution functions for ions, etc.) computed from different parts of the MD trajectory. Additionally, equilibration of the DNA chain was assured by estimating correlation times for various DNA conformational modes, as elaborated below.

Note, it would also be beneficial to study DNA flexibility at even higher ionic concentrations, $c > 1$ M, and compare our computational results to experimental data. For example, there exist few experimental points for DNA persistence length measured at NaCl concentrations in a range of [1-4] M (see Fig. 2a of the main text). However, as mentioned in the manuscript, simulating a 150-base-pair DNA segment in 1M of NaCl already poses quite a challenge because of a large number of explicit mobile ions and long equilibration times required. For example, the number of Na+ and Cl- ions corresponding to 2M, 3M and 4M in a simulation box having dimensions $\sim 600 \times 600 \times 600$ Å would be, respectively, 400 000, 600 000 and 800 000. Our experience tells that simulating such systems for a few hundreds of nanoseconds demands [300 000 - 600 000] CPU hours using 2.3 GHz Intel processors, which is equivalent to [3 - 6] months of physical time carrying out calculations in parallel regime on 128 CPUs. Therefore, DNA flexibility in this work was studied at ionic concentrations not exceeding 1 M. Nevertheless, a majority of experimental

data, to which our computational results were compared, fall in a range [0.1 - 1] M of NaCl concentrations.

3 Estimating correlation times for DNA conformational modes

In the main text, Eq.(2), we introduced the following temporal correlation functions,

$$C^i(t) = \left\langle \cos[\Delta\alpha^i(t)] \right\rangle, \quad \Delta\alpha^i \equiv \alpha^i(0) - \alpha^i(t), \tag{SI.7}$$

defining how orientational correlation among two tangent vectors separated by i DNA segments along the chain decays with time [$\alpha^i(0)$ and $\alpha^i(t)$ being the angles between the vectors measured at times 0 ant t, respectively]. These functions are shown in the Fig. SI **2** for simulated systems

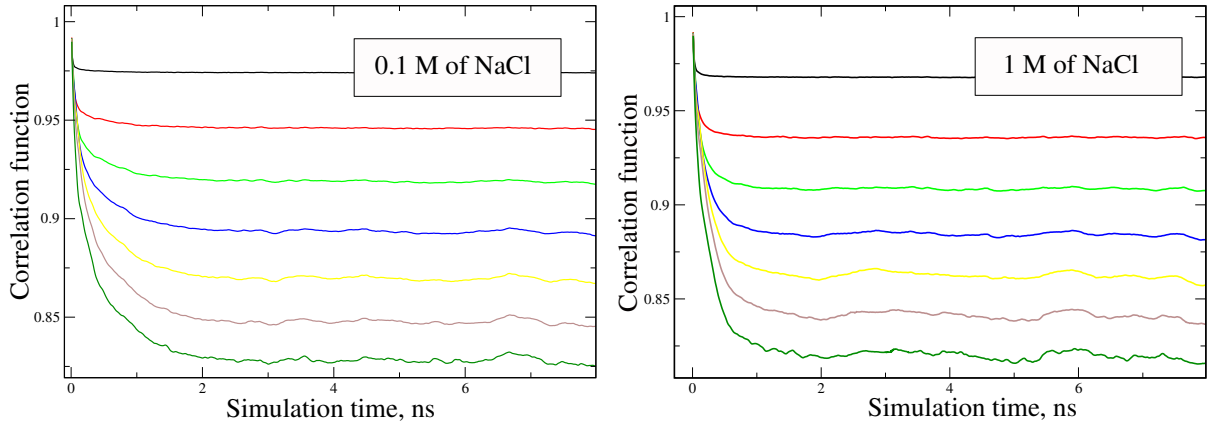

Fig. SI 2: Correlation functions, Eq.(SI.7), computed for tangent vectors separated by (top-down): 0, 2, 4, 6, 8, 10 and 12 DNA segments along the chain, for DNA oligomer in 0.1M and 1M NaCl salt buffers.

of DNA segment immersed in a 0.1M and 1M NaCl salt buffer. Correlation times, τ^i, can be readily estimated from the visual analysis of correlation functions and turned out to be in a range of $\sim [1-2]$ ns for all DNA modes. As mentioned in the main text, we followed the empirical rule that the longevity of MD trajectory needs to be no less than $500\tau_{\max} \simeq [0.5-1]$ μs to ensure a good statistics while calculating DNA persistence length via Eq.(1) (see the main text).

References

[1] Kohlbacher, O., and H. P. Lenhof. 2000. BALL–rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. *Bioinformatics* 16:815–824.

[2] Savelyev, A., and G. A. Papoian. 2010. Chemically accurate coarse graining of double-stranded dna. *Proc Natl Acad Sci U S A* 107:20340–20345.

[3] Savelyev, A., and G. A. Papoian. 2006. Electrostatic, steric, and hydration interactions favor na(+) condensation around dna compared with k(+). *J Am Chem Soc* 128:14506–14518.

[4] Savelyev, A., and G. A. Papoian. 2009. Molecular renormalization group coarse-graining of polymer chains: application to double-stranded dna. *Biophys J* 96:4044–4052.

[5] Savelyev, A., and G. A. Papoian. 2009. Molecular renormalization group coarse-graining of electrolyte solutions: application to aqueous nacl and kcl. *J Phys Chem B* 113:7785–7793.

[6] Nielsen, S. O., C. F. Lopez, G. Srinivas, and M. Klein. 2004. Coarse grain models and the computer simulation of soft material. *J Phys: Condens Matter* 16:R481–R512.

[7] Savelyev, A., and G. A. Papoian. 2008. Polyionic charge density plays a key role in differential recognition of mobile ions by biopolymers. *J Phys Chem B* 112:9135–9145.

[8] Swendsen, R. H. 1979. Monte carlo renormalization group. *Phys Rev Lett* 42:859–861.

[9] Zinn-Justin, J. 2002. Quantum Field Theory and Critical Phenomena. Clarendon press, Oxford.

[10] Plimpton, S. 1995. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics* 117:1–19.

[11] Hoover, W. G. 1985. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* 31:1695–1697.

[12] Darden, T., D. York, and L. Pedersen. 1993. Sequence-specific binding of counterions to B-DNA. *J Chem Phys* 98:10089–10092.

[13] Knotts, T. A., N. Rathore, D. Schwartz, and J. J. de Pablo. 2007. A coarse grain model fro dna. *J Chem Phys* 126:084901.