SUPPLEMENTARY DATA

DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale

Tianyin Zhou[1], Lin Yang[1], Yan Lu[1], Iris Dror[1], Ana Carolina Dantas Machado[1], Tahereh Ghane[1], Rosa Di Felice[1,3], and Remo Rohs[1,2,3,4,5,*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences
[2]Department of Chemistry
[3]Department of Physics and Astronomy
[4]Department of Computer Science
[5]Norris Comprehensive Cancer Center
University of Southern California, Los Angeles, CA 90089, USA

*To whom correspondence may be addressed:

Remo Rohs, Ph.D.
Molecular and Computational Biology Program
1050 Childs Way RRI 404C
University of Southern California
Los Angeles, CA 90089
United States
Tel: +1-213-740-0552
Fax: +1-213-821-4257
Email: rohs@usc.edu

SUPPLEMENTARY MATERIALS AND METHODS

Sequences and PDB IDs of Fis-DNA binding sites used in Figure 1 include (B) binding affinities for the seven sequences F1, F24, F25, F26, F27, F28, and F29 reported in (6), which only differ in their 5 central base pairs (underlined in Supplementary Table S2), and (C-D) comparisons with crystal structures 3iv5 (F1) and 3jrc (F29) of high- and low-affinity DNA targets (6). The DNA sequences of all seven Fis binding sites are listed in Supplementary Table S2.

PDB IDs of structures of the Dickerson dodecamer used in Figure 2 include data from X-ray crystallography (1bna, 2bna, 355d, 428d, 455d, 1dou, 1fq2, and 1jgr) and NMR spectroscopy (1duf and 1naj). The two PDB entries for NMR data of the Dickerson dodecamer each contain 5 structures.

PDB IDs of structures of protein-DNA complexes used in Figure 3 include (A) 1oct (OCT1-POU), (B) 1ig7 (Msx-1), (C) 1akh (MATa1-MATα2), (D) 1hf0 (OCT1-PORE), (E) 3fdq (MogR), and (F) 2or1 (Phage 434 repressor) (1).

PDB IDs of structures of protein-DNA complexes used in Supplementary Figures S3-S7 include (S3) 2fio (Phage Φ29 regulator p4), (S4) 1bi8 (Ubx-Exd), (S5) 1z9c (OhrR regulator), (S6) 1rep (RepE initiator), and (S7) 2fo1 (CSL-Notch-Mastermind).

Sequence data for the *in vivo* nucleosome binding sites used in Figure 4 for the (A) *S. cerevisiae* (22) and (B) *D. melanogaster* (23) genomes have been aligned as described in our recent study (19).

Monte Carlo (MC) simulations. The molecular model for the MC simulations uses a random sampling of DNA conformations based on collective and internal variables (14,15). A basic assumption of this model is that bond lengths and the aromatic ring systems of the bases are fixed. This approximation enables a significant reduction in the degrees of freedom (16). The MC implementation that we applied in this work uses 6 collective variables (3 rigid-body translations and 3 rigid-body rotations) and 6 internal variables (glycosidic torsion angle, sugar phase and amplitude, and two endocyclic torsion and one bond angle) as independent MC variables. The remaining degrees of freedom are sampled as dependent variables of the analytic chain closure (16) following each move. The analytic chain closure enables reversible α/γ flips and BI/BII transitions. In addition to varying the position and orientation of nucleotides as rigid bodies and the sampling of the phosphodiester backbone, the thymine methyl group is rotated as an additional MC variable. The MC method uses an implicit solvent description based on a sigmoidal distance-dependent dielectric function (17) and explicit sodium counter ions to neutralize the system. These ions are moved in terms of their 3 Cartesian coordinates as additional MC variables (15). Random conformational transitions are accepted with a probability based on the Metropolis-Boltzmann criterion with associated Jacobians derived in earlier work (16).

We calculated the total system energy with the AMBER94 force field (29) implemented as previously described (14). During the MC sampling, we dynamically adjusted the maximum move sizes for each MC variable to achieve approximately 50% acceptance rates for each move. The starting configurations for each MC simulation were built using canonical B-DNA with identical structural features for each dinucleotide. We performed 2,121 MC simulations (see Supplementary Table S1 for list of sequences) over 2 million MC cycles of which we considered the initial 500,000 MC cycles as equilibration period (14). The sampling consists of randomly varying the full set of these degrees of freedom of a DNA duplex of finite length including the ion coordinates in each MC cycle. We recorded snapshots every 10th MC cycle along the MC trajectory to generate an ensemble of conformations for each duplex (3,14), analyzed these 150,000 snapshots with Curves (20), and calculated \ensemble averages of structural parameters over the equilibrated part of the MC simulations (3,14,15).

Molecular Dynamics (MD) simulations. Classical MD simulations with explicit solvent molecules and counter ions were carried out for the Dickerson dodecamer with the NAMD code and the parmbsc0 AMBER force field (28), a methodology extensively tested for nucleic acids (26,30).

The starting configuration was taken from the co-crystal structure with PDB ID 1bna. The simulation system was prepared by immersing the DNA oligomer in explicit solvent. Specifically, a cubic box was generated, by adding a buffer layer (~16 Å in each spatial direction) of TIP3P water molecules with standard hydration rules (30). Sodium counter ions were added to neutralize the system. The ions were placed according to the electronegativity map.

The simulation protocol included an equilibration phase (30) composed of various steps of conjugate-gradient optimization of atomic coordinates and restrained finite-temperature dynamics during which the restraints were gradually weakened and eventually released.

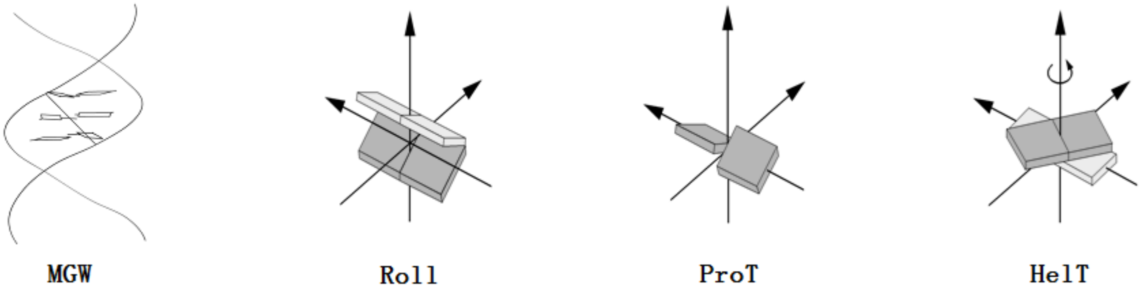
1. We initially performed solvent equilibration only: all the water molecules were first relaxed and then subjected to 20 ps dynamics, while the nucleic acid molecule and the counter ions were kept fixed with a SHAKE tolerance of 10^{-8}. During this solvent equilibration, the temperature of the solvent molecules was slowly raised to 100 K by coupling to the heat bath and the pressure was kept at 1.01325 bar.
2. We performed a full-atom minimization of this partially equilibrated system.
3. The quenched system was heated slowly from 0 to 300 K by coupling it to a heat bath whose temperature was raised at the rate of 50 K every 10 ps.
4. The system was equilibrated for another 100 ps at a temperature of 300 K and pressure of 1.01325 bar.
5. Finally, we ran the production MD simulation of the dodecamer at the same temperature and pressure in the NPT ensemble (P = 1.01325 bar, T = 300 K) using a time step of 2 fs, for a total duration of 100 ns during which we recorded the system coordinates every 1 ps. Periodic boundary conditions and the Particle-Mesh-Ewald algorithm were used.

At the end of the equilibration the trajectory was stable. We monitored the stability of the trajectory during the finite-temperature dynamics by computing the root mean square deviation at every step. Helical parameters (MGW, Roll, ProT, and HelT) for each 1 ps snapshot were computed with Curves (20) and analyzed over the final 90 ns of the 100-ns simulation.

Comparison of CPU time between MC and MD simulations. The MD simulation of the Dickerson dodecamer (12 base-pair duplex DNA) requires 0.2 days/ns (20 days for 100 ns) on 24 processors using 2 nodes Quad-Core Intel Xeon E5462 Core @ 2.8 GHz – 16 Gb RAM – on a cluster that contains 12 such nodes with Myrinet connection. The equivalent time on a single processor of the same cluster is 2.5 days/ns (~8.5 months for 100 ns). In comparison, the MC simulation of the Dickerson dodecamer over 2 million MC cycles on a computing cluster of comparable size requires 2.7 days using 24 processors or 6.5 days using a single processor. Thus, the MC simulation requires approximately 13.5% of the CPU time necessary for the MD simulation if 24 processors are used and only 2.6% of the CPU time necessary for the MD simulation if a single processor is used. This benchmark explains why we used MC simulations to generate the large number of 2,121 MC trajectories to reach the coverage of on average 44 occurrences of each pentanucleotide in our HT method, which speeds up the prediction to instant feedback level.

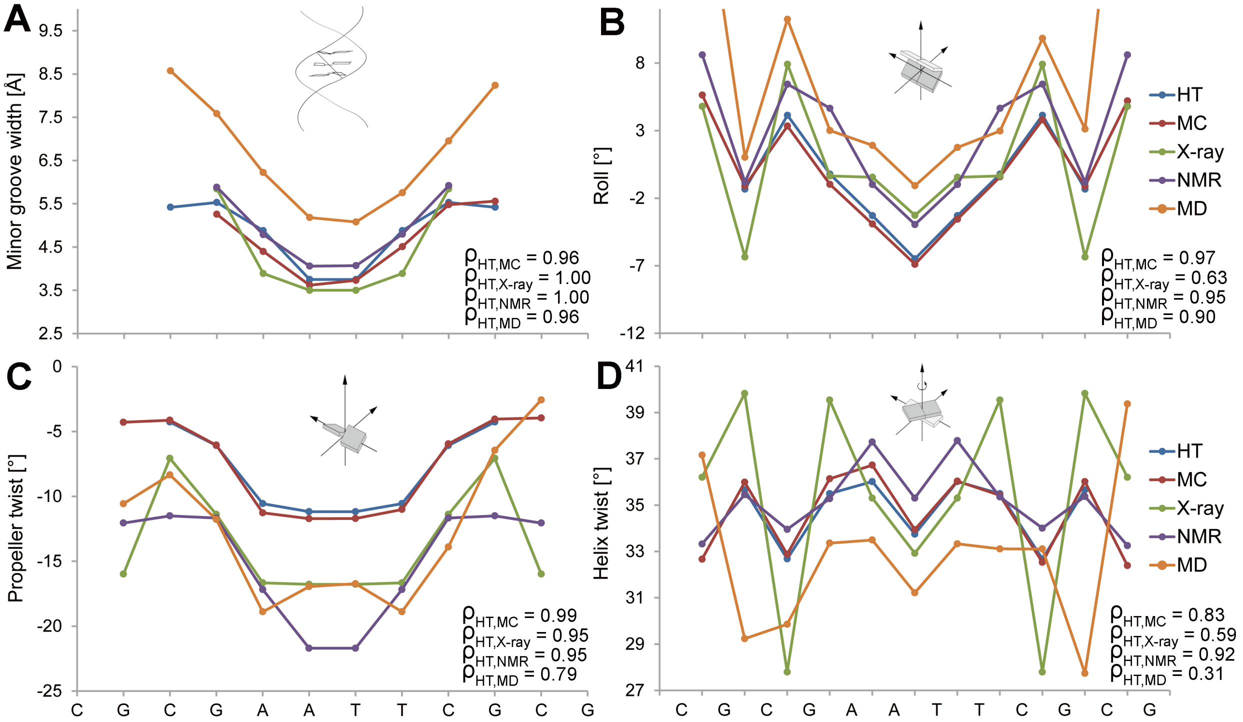
SUPPLEMENTARY FIGURES

Supplementary Figure S1. Schematic representations of predicted structural features.



The predicted structural features include minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT) as defined by Curves (14).

Supplementary Figure S2. Validation of HT prediction of structural parameters for the Dickerson dodecamer of the palindromic sequence CGCGAATTCGCG.



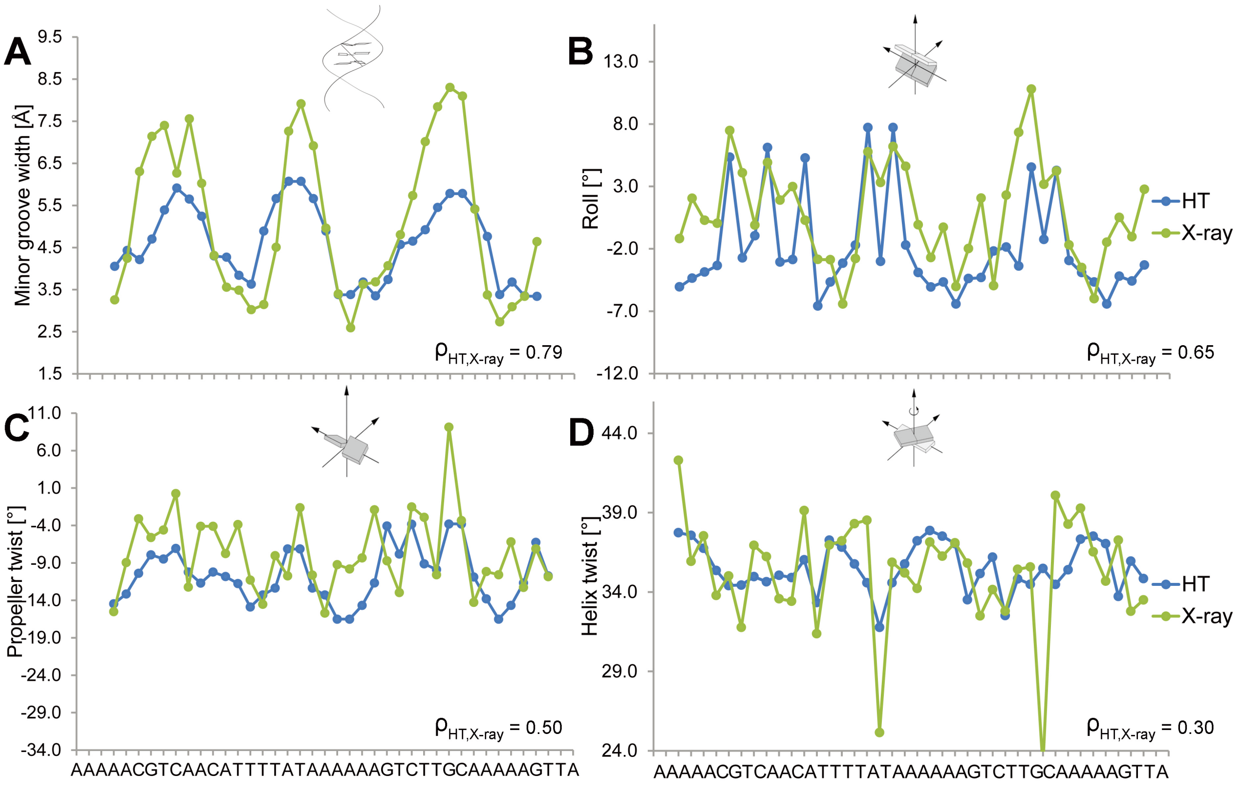
The structural features (A) MGW, (B) Roll, (C) ProT, and (D) HelT of the palindromic DD dodecamer are predicted using the HT (blue) and MC (red) approaches and compared with results from a 100 ns MD simulation (orange) and the symmetrized average profiles derived from 8 crystal structures (green) without chemical modifications and the average profiles derived from 10 NMR structures (purple) using RDC data in the refinement (PDB IDs in Supplementary Table S3).

Spearman's rank correlation coefficients (ρ) demonstrate the statistical similarity between the predicted and experimental structural feature profiles. The agreement between HT predictions, MC and MD simulations, and X-ray and NMR data is apparent. Noteworthy is the known systematic overestimation of MGW and underestimation of HelT in MD simulations (13,26,27). Roll is slightly overestimated in MD simulations and slightly underestimated in MC simulations (see also Supplementary Figure S8). Whereas ProT is, at least in this example, overestimated in MC simulations, the pattern of a single minimum is correctly predicted, leading to a higher Spearman's rank correlation than for MD simulations, which predict more accurate absolute values but a pattern with a double minimum. It should also be noted that HelT assumes more extreme values in co-crystal structures, likely due to crystal packing, in comparison to solution-state NMR structures. Since the HT method does not predicts such extreme HelT values of < 30° and > 39°, the Spearman's rank correlations are always lower for HelT compared to the other structural features (see Supplementary Table S4).

Whereas structural features derived from X-ray and NMR data were symmetrized for the palindromic Dickerson dodecamer, results from HT predictions, MC and MD simulations shown here were not symmetrized. The HT predictions are symmetric due to the underlying sliding pentamer model. Deviations from the palindromic symmetry in predictions of structural features by MC and MD simulations are a measure of equilibration of the conformational sampling. The MC and MD predictions suggest that the 2 million-cycle MC simulation is better equilibrated than the 100-ns MD simulation.

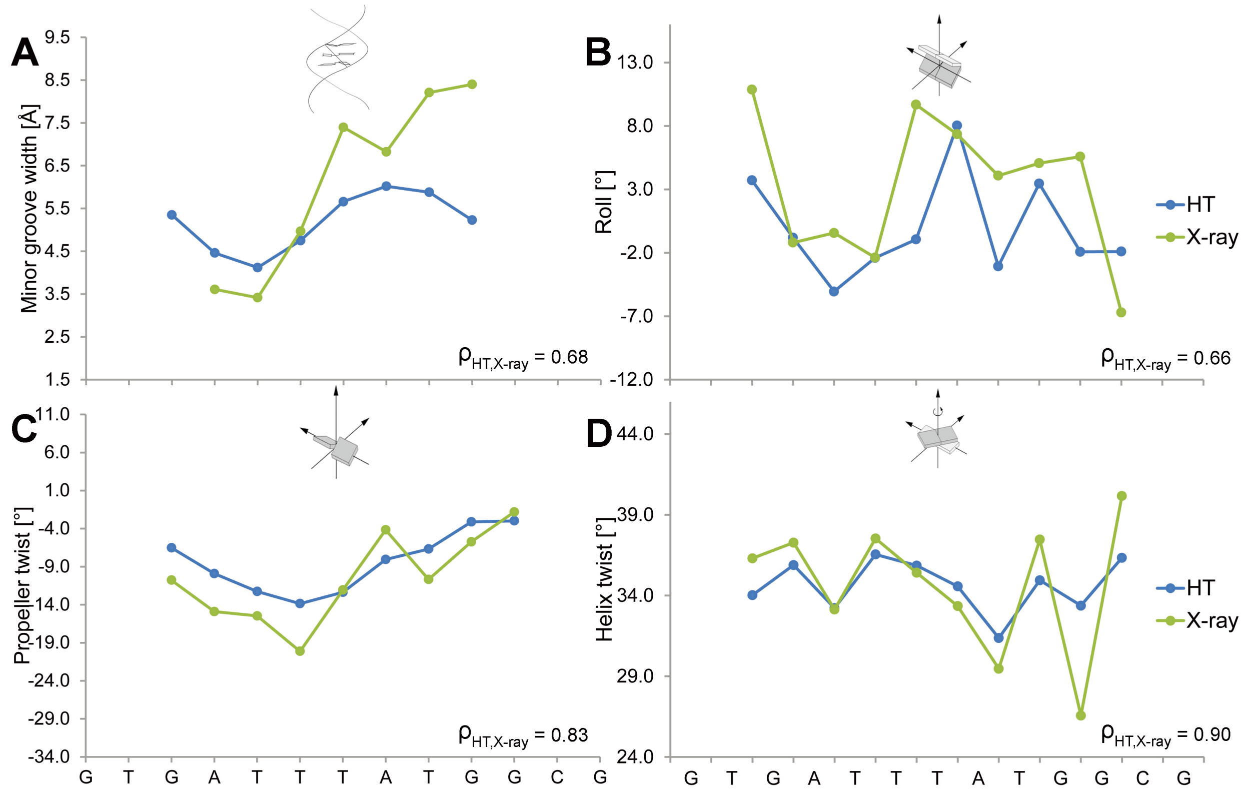
In addition, we previously showed the high correlation of the MC prediction of MGW with OH cleavage intensity measurements for the Dickerson dodecamer (19).

Supplementary Figure S3. High-throughput prediction of structural parameters for the DNA binding site of the phage Φ29 regulator p4.



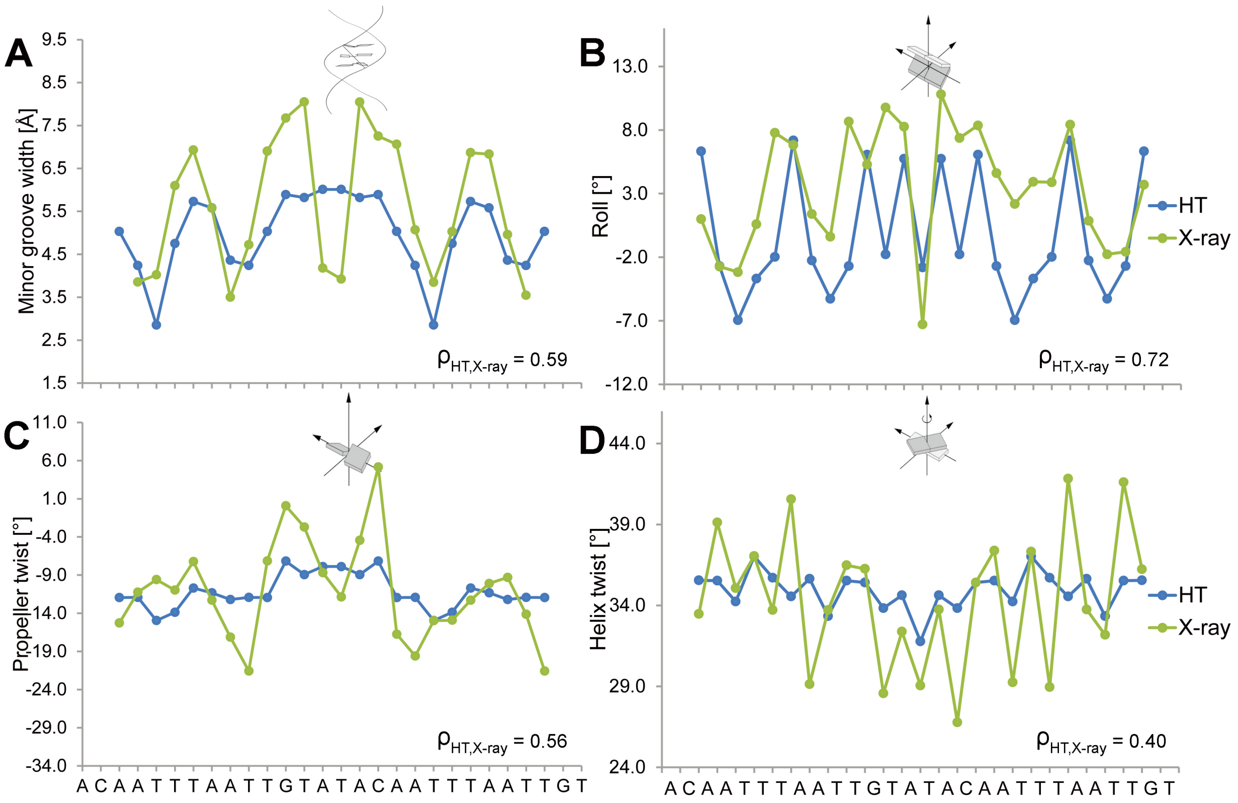
The structural features (A) MGW, (B) Roll, (C) ProT, and (D) HelT of the DNA binding site of the phage Φ29 regulator p4 are predicted using the HT approach (blue) and compared with the co-crystal structure (green) with the PDB ID 2fio. Spearman's rank correlation coefficients (ρ) demonstrate the statistical similarity between the predicted and experimental structural feature profiles. The more extreme HelT values observed in the X-ray structure are usually due to crystal packing, leading to a lower Spearman's rank correlation coefficient.

Supplementary Figure S4. High-throughput prediction of structural parameters for the DNA binding site of the Ubx-Exd heterodimer.



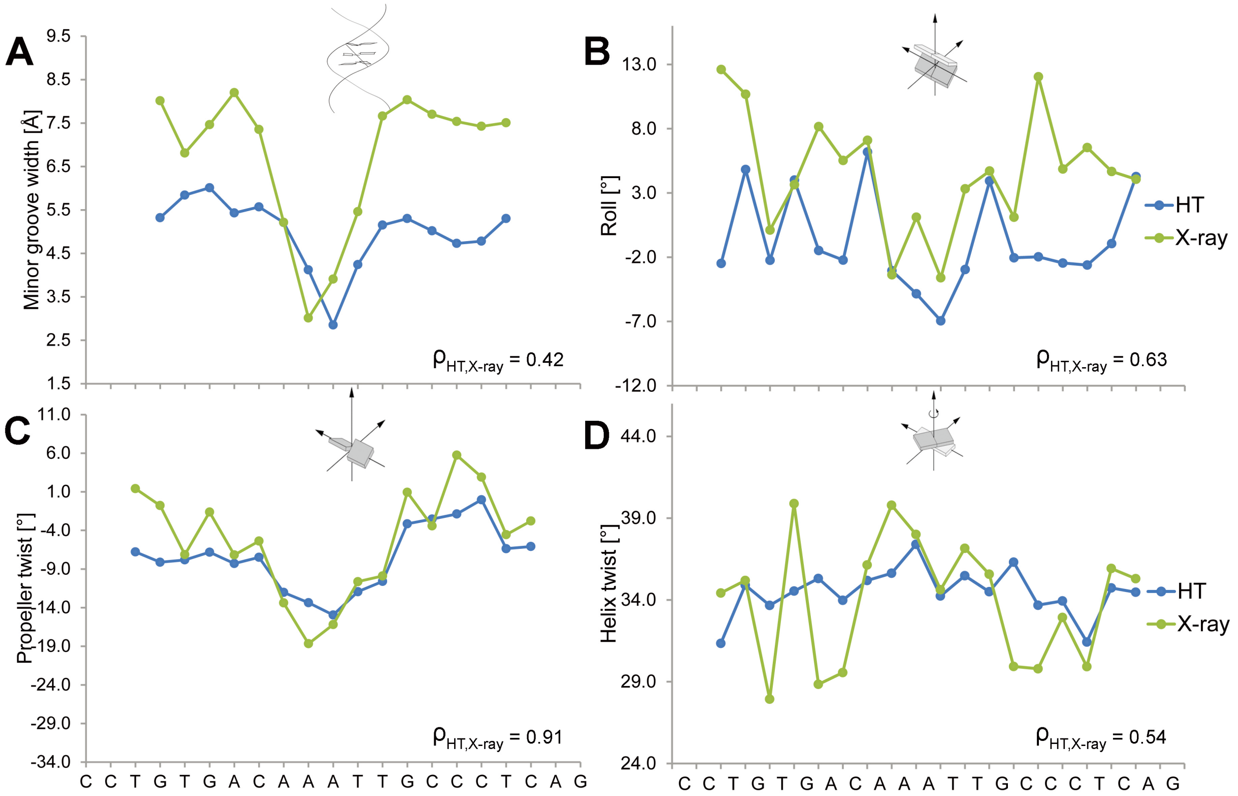
The structural features (A) MGW, (B) Roll, (C) ProT, and (D) HelT of the DNA binding site of the Ubx-Exd heterodimer are predicted using the HT approach (blue) and compared with the co-crystal structure (green) with the PDB ID 1b8i. Spearman's rank correlation coefficients (ρ) demonstrate the statistical similarity between the predicted and experimental structural feature profiles.

Supplementary Figure S5. High-throughput prediction of structural parameters for the DNA binding site of the OhrR regulator.



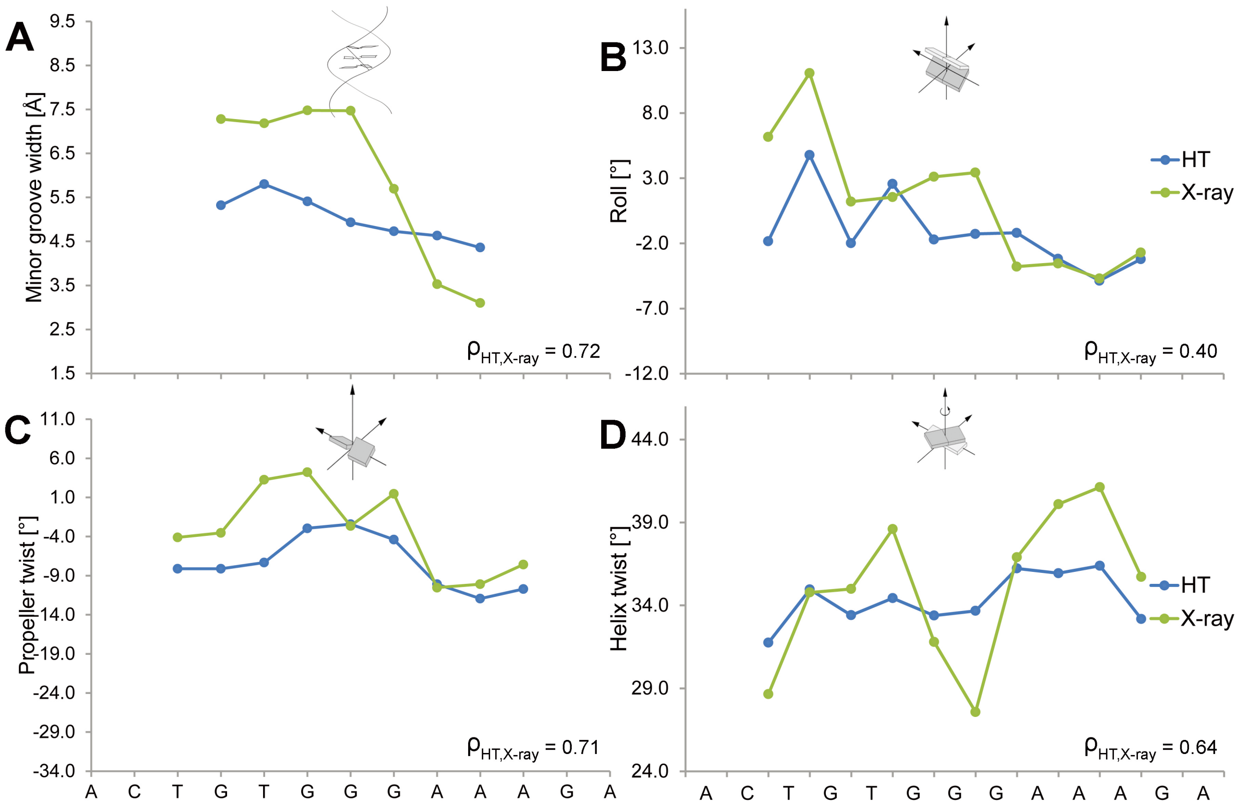
The structural features (A) MGW, (B) Roll, (C) ProT, and (D) HelT of the DNA binding site of the OhrR regulator are predicted using the HT approach (blue) and compared with the co-crystal structure (green) with the PDB ID 1z9c. Spearman's rank correlation coefficients (ρ) demonstrate the statistical similarity between the predicted and experimental structural feature profiles. The more extreme HelT values observed in the X-ray structure are usually due to crystal packing, leading to a lower Spearman's rank correlation coefficient.

Supplementary Figure S6. High-throughput prediction of structural parameters for the DNA binding site of the RepE initiator.



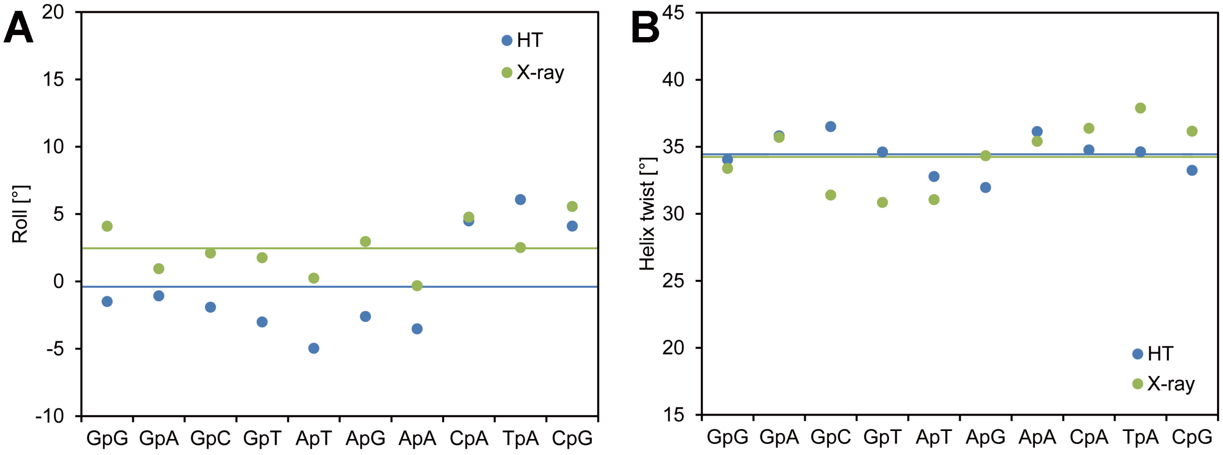
The structural features (A) MGW, (B) Roll, (C) ProT, and (D) HelT of the DNA binding site of the RepE initiator are predicted using the HT approach (blue) and compared with the co-crystal structure (green) with the PDB ID 1rep. Spearman's rank correlation coefficients (ρ) demonstrate the statistical similarity between the predicted and experimental structural feature profiles.

Supplementary Figure S7. High-throughput prediction of structural parameters for the DNA binding site of the CSL-Notch-Mastermind ternary complex.



The structural features (A) MGW, (B) Roll, (C) ProT, and (D) HelT of the DNA binding site of the CSL-Notch-Mastermind ternary complex are predicted using the HT approach (blue) and compared with the co-crystal structure (green) with the PDB ID 2fo1. Spearman's rank correlation coefficients (ρ) demonstrate the statistical similarity between the predicted and experimental structural feature profiles.

Supplementary Figure S8. Validation of HT prediction of Roll and HelT for the 10 unique dinulcleotides based on a dataset of co-crystal structures of protein-DNA complexes.



The agreement between HT predictions and X-ray data of a large dataset of 760 co-crystal structures is apparent. Each unique dinucleotide occurs on average 262 times in these structures in regions that are not drastically deformed (see Methodology). (A) Roll is slightly underestimated in HT predictions, while the sequence-dependent variations between dinucleotides are well captured. (B) HelT is estimated very well. Noteworthy is that the known systematic underestimation of HelT in MD simulations (13,26,27) has been overcome in HT predictions, including the HelT values for CpA, CpG, and TpA dinucleotides for which MD simulations even with refined force fields (28) report notoriously low HelT values (13).

SUPPLEMENTARY TABLES

Supplementary Table S1. 2,121 sequences in MC training dataset.

We predicted all-atom structures using Monte Carlo (MC) simulations for the 2,121 DNA sequences of length 12-27 bp listed in the attached spreadsheet (included as separate file). A subset of the MC data was previously published (18). The remaining sequences were designed to achieve a statistically significant coverage of all 512 unique pentanucleotides.

Supplementary Table S2. Seven sequences of Fis-DNA binding sites.

Binding site	Sequence (variable central 5 bp underlined)
F1	AAATTTGTTTGAATTTTGAGCAAATTT
F24	AAATTTGTTTGTTTTTTGAGCAAATTT
F25	AAATTTGTTTGTTAAATGAGCAAATTT
F26	AAATTTGTTTGAAAAATGAGCAAATTT
F27	AAATTTGTTTGAACTTTGAGCAAATTT
F28	AAATTTGTTTGAGCGTTGAGCAAATTT
F29	AAATTTGTTTGGGCGCTGAGCAAATTT

Binding affinities for these seven sequences were previously published (6).

Supplementary Table S3. PDB IDs for X-ray and NMR structures of DNA duplexes and protein-DNA complexes used for validation.

Dataset	**PDB IDs**
X-ray bound	1p3o, 1s97, 1c0w, 2h1o, 2er8, 2c6y, 1n48, 1r0a, 2nvx, 2ady, 1efa, 1akh, 2r5y, 2e2i, 1jj6, 1srs, 1rzr, 1gu5, 1qn4, 2is4, 1njx, 2j6u, 1ign, 2gli, 1owg, 1iaw, 2iie, 1oh6, 2ost, 1s32, 1zx4, 2pi5, 1nwq, 1pue, 1kb2, 1h88, 2au0, 1qln, 1k78, 1mm8, 1p3m, 1nvp, 1ozj, 1bp7, 1jkq, 1ddn, 1muh, 2agq, 2ppb, 1dsz, 2h27, 2ief, 1ic8, 1rh6, 2aoq, 1s10, 1lbg, 1f4k, 1oh7, 2euv, 1p34, 2d5v, 1xo0, 1r8d, 1lwt, 1t8i, 1cqt, 1d2i, 2o8e, 2dy4, 2d45, 1r4r, 1id3, 2e2j, 1ng9, 1tw8, 2i9t, 4ktq, 1le8, 1p3b, 1qnc, 2hmi, 1gji, 1zme, 2c2r, 2erg, 1cdw, 1skn, 2owo, 1hdd, 2c28, 1mus, 2o93, 1tl8, 1kbu, 2i13, 2bnz, 3pvi, 2jej, 1jkr, 1ma7, 1ig9, 2i9k, 1njy, 1k79, 1kx5, 1aoi, 1p3g, 1lmb, 1evw, 1iu3, 1f66, 2as5, 1n6j, 1z19, 1j1v, 2ewj, 2hot, 1mow, 1c9b, 2aor, 1z63, 1fjl, 1tkd, 2atl, 1pp8, 1hjc, 1s0n, 1oh8, 2p0j, 2a3v, 1ttu, 1hw2, 1au7, 2bq3, 1cz0, 1hcr, 1l3u, 1r0n, 1q9y, 1g2f, 1g9y, 2o5i, 2hvr, 2irf, 2euz, 2ktq, 1llm, 1qn8, 2cax, 1u8b, 1nkb, 2evg, 1skm, 2uvr, 2j6s, 1nzb, 1u3e, 1ej9, 1dux, 1n6q, 2aq4, 2hos, 1t8e, 2gm4, 1rpe, 1a74, 1ipp, 1sax, 2hof, 1p7d, 2i3q, 1dh3, 2uvv, 1apl, 1r0o, 1pvp, 1p3f, 1p3k, 1j59, 1w7a, 1eoo, 1mur, 1p8k, 1fjx, 1cez, 1bc7, 2og0, 2cgp, 1ytf, 1odh, 1a0a, 1p4e, 1wbd, 1ksy, 1ga5, 4crx, 1zre, 1ksx, 1sc7, 2jef, 1zlk, 2ajq, 1dfm, 1hlz, 1nk0, 2hdd, 2iif, 2rve, 1eqz, 1bl0, 1ecr, 2ezv, 1k7a, 1pvi, 1pyi, 1u0d, 1lpq, 1h0m, 1r7m, 2f8x, 2evh, 2ere, 1tf6, 1jj4, 1vrr, 1yo5, 1tk8, 1cf7, 2asj, 1m5r, 2asd, 2hhq, 1zg1, 1i3j, 1q0t, 1p7h, 1vkx, 1b72, 3crx, 1ixy, 1qne, 1hf0, 2oh2, 2br0, 1xbr, 1nkc, 2cv5, 2or1, 1sxq, 1f0o, 1qn9, 1p3l, 1kb6, 2p6r, 2ivh, 1a6y, 1flo, 9ant, 8mht, 1tqe, 2a66, 1zrf, 1w0u, 2bnw, 1r4o, 2f5p, 1lli, 1nk5, 2pi4, 1vol, 2gie, 1jey, 1hwt, 1le9, 2f8n, 1per, 2oaa, 2nvz, 1xns, 1tk0, 2ayb, 1bpz, 1ngm, 2evi, 1w0t, 1m1a, 1nne, 2fo1, 1cit, 2hvs, 2h8r, 2bsq, 1xhv, 2hoi, 1p47, 2i3p, 1ua1, 2nll, 1h9t, 3mht, 1h89, 2hhv, 1u8r, 2b9s, 1xhu, 2hr1, 1run, 1fok, 1qss, 1z1g, 1k6o, 1q3v, 1zs4, 1g2d, 1d66, 2h1k, 1io4, 1wbb, 2evf, 1l5u, 1n5y, 1rys, 1h8a, 1qn6, 1fos, 1zrc, 1zyq, 2etw, 1a36, 1ysa, 1rep, 1u35, 1hlv, 1f5t, 1t3n, 2gig, 1bdt, 2c7a, 1nkp, 2r5z, 1r9t, 1tgh, 2han, 1k61, 1zns, 1d5y, 1fw6, 1l3l, 1rz9, 1m5x, 1g3x, 2crx, 1o3t, 2ht0, 2o61, 2acj, 1glu, 1l3v, 1ijw, 1qp9, 1nlw, 1tx3, 2bqu, 1wb9, 1egw, 1m19, 1g9z, 1qna, 1zrd, 1awc, 1q3u, 1gtw, 1am9, 1jko, 2jeg, 1nh2, 2evj, 1mnn, 3cro, 1ouz, 1eyu, 2eux, 1pvr, 2is6, 1u78, 2ex5, 2bqr, 2hhs, 2it0, 1le5, 1a3q, 1qn3, 1wte, 1s9f, 1by4, 1m6x, 1puf, 1oh5, 1pvq, 2uvu, 1d3u, 1e3m, 1r49, 1j5o, 2dtu, 1lq1, 2f5n, 2a07, 1ihf, 2imw, 1trr, 2dpd, 1rm1, 2ntc, 1lat, 2euw, 4bdp, 1du0, 1r71, 2p5o, 2gih, 1owr, 1t05, 1zr4, 2q2t, 1bf5, 1f2i, 1gu4, 1qn7, 1lrr, 1cgp, 1crx, 2fio, 1q9x, 1yf3, 2nra, 1hlo, 1lb2, 1nk8, 1mjq, 1n3e, 1z1b, 2c9l, 2fld, 1gd2, 1m18, 3hdd, 1kc6, 1pp7, 2hzv, 2ivk, 1ig7, 1qnb, 1drg, 1rio, 1njw, 1jkp, 1mnm, 1cyq, 1gt0, 1jx4, 1a02, 2geq, 5crx, 1tsr, 2q2u, 2h7h, 1zbb, 1mj2, 2uvw, 1jj8, 1mey, 1mhd, 1jnm, 1kb4, 1ckt, 1u0c, 1n3f, 1b3t, 1ram, 1yfh, 2asl, 2c2e, 1n56, 1ewq, 1lws, 1zr2, 2c2d, 1ea4, 1oct, 2drp, 1o3r, 1bdv, 1imh, 1kx3, 1s0o, 1s0m, 1gxp, 1gdt, 2isz, 2ntz, 2j6t, 2o6g, 1p3p, 1p3a, 1o3q, 1ouq, 1tro, 1if1, 2bgw, 2odi, 2jei, 1par, 1rr8, 2hap, 1tup, 2is2, 2ago, 1b8i, 2c22, 2q10, 1ubd, 1yfi, 2gii, 1pt3, 2np2, 1rtd, 2e2h, 1qn5, 1dc1, 2ata, 1pzu, 1nk9, 1mjo, 1p3i, 1jfi, 1t2k, 1e3o, 1t2t, 1w36, 1r4i, 2f5o, 1je8, 1k4t, 1ytb, 2fj7, 2nzd, 1jt0, 1ryr, 1kx4, 2o5j, 1o3s, 1hbx, 1zla, 2gij, 1t9j, 1a73, 1s9k, 1ruo, 1f44, 1qtm, 1ynw, 2hht, 1mdm, 1hcq, 2ayg, 2ac0, 1tc3, 1owf, 1yrn, 1zg5, 1pdn, 1h6f, 2ahi, 1mdy, 1z9c, 1m0e, 1hjb, 2is1, 1k82, 1sa3, 6pax, 1yfj, 1qsy, 2agp, 1jwl, 1t9i, 1an4, 3ktq, 1rrj
X-ray unbound	1s2r, 1jgr, 2b1d, 424d, 9bna, 249d, 1fq2, 1ilc, 355d, 1dpn, 1lp7, 423d, 388d, 287d, 1d98, 428d, 461d, 1bna, 1hq7, 1dou, 1dn9, 389d, 425d, 455d, 1d29, 1d65, 194d, 1d28, 1bdn, 460d, 119d, 2bna, 1d89, 1sgs, 436d, 1dc0
NMR unbound	1tqr, 1naj, 1rvh, 1ss7, 1duf, 1fzx, 1x2o, 1x2u, 1rvi, 1x2s, 1g14, 1ssv
X-ray Dickerson	1bna, 2bna, 355d, 428d, 455d, 1dou, 1fq2, 1jgr
NMR Dickerson	1duf, 1naj

Supplementary Table S4. Spearman’s rank correlation coefficients for comparison of HT predictions with experimental data.

Spearman’s rank correlation coefficients for comparison of HT predictions with structural features derived from experimental structures solved by X-ray crystallography for DNAs bound by proteins (X-ray bound) and unbound DNA (X-ray unbound). Due to the small number of available structures in the X-ray unbound dataset (11), additional structures of unbound DNA solved by NMR spectroscopy (NMR unbound) were included in the validation study. MGW, Roll, ProT, and HelT were calculated with Curves (20). The datasets of experimental structures retrieved from the Protein Data Bank (PDB) are identical to the ones used in our recent study (19). HT predictions of the Dickerson dodecamer are compared individually because it is the best-studied DNA molecule (11). Crystal structures of the Dickerson dodecamer (X-ray Dickerson) have been symmetrized prior to the comparison in order to remove crystal-packing effects. NMR structures of the Dickerson dodecamer (NMR Dickerson) have only been included if residual dipolar coupling (RDC) data has been used in the refinement. The datasets are characterized by their PDB IDs listed in Supplementary Table S3. The numbers of PDB IDs and structures are not identical because several molecules can be contained in a single asymmetric unit (X-ray) or the refinement can result in multiple alternative configurations (NMR).

Dataset	**#PDB IDs**	**#Structures**	**MGW**	**Roll**	**ProT**	**HelT**
X-ray bound	591	760	0.67	0.63	0.55	0.54
X-ray unbound	36	46	0.55	0.31	0.49	0.29
NMR unbound	12	90	0.68	0.68	0.33	0.31
X-ray Dickerson	8	8	1.00	0.63	0.95	0.54
NMR Dickerson	2	10	1.00	0.95	0.95	0.95

AUTHOR CONTRIBUTIONS

T.Z. developed and validated the HT method, designed the web server, performed and analyzed MC simulations, and analyzed experimental structures and MD simulations. L.Y. analyzed nucleosome-binding sites, performed MD simulations, and assisted with the web server documentation. Y.L. modified MC analysis tools. I.D. contributed to the conceptual design of the HT method. A.C.D.M. performed and analyzed MC simulations. T.G. and R.D.F. performed and analyzed MD simulations. T.Z. and R.R. wrote the manuscript. R.R. conceived, designed, and supervised the project.