JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Evolutionary Algorithm in the Optimization of a Coarse-Grained Force Field

Filip Leonarski,*[,†,‡] Fabio Trovato,[§] Valentina Tozzini,[∥] Andrzej Leś,[‡] and Joanna Trylska*[,†]

[†]Centre of New Technologies, University of Warsaw, Żwirki i Wigury 93, Warsaw 02-089, Poland

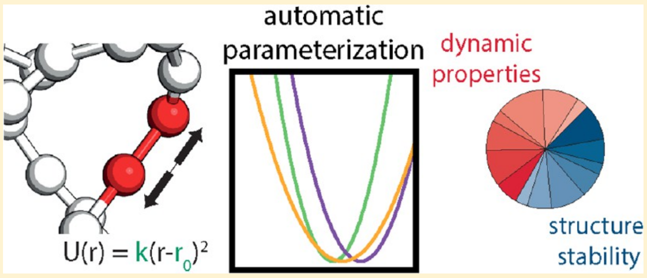
[‡]Faculty of Chemistry, University of Warsaw, Pasteura 1, Warsaw 02-093, Poland

[§]NEST, Istituto Nanoscienze - Cnr, Scuola Normale Superiore and Center of Nanotechnology and Innovation, IIT, Piazza San Silvestro 12, 56127 Pisa, Italy

[∥]NEST, Istituto Nanoscienze - Cnr and Scuola Normale Superiore, Piazza San Silvestro 12, 56127 Pisa, Italy

Supporting Information

ABSTRACT: Simulations using residue-scale coarse-grained models of biomolecules are less computationally demanding than simulations employing full-atomistic force fields. However, the coarse-grained models are often difficult and tedious to parametrize for certain applications. Therefore, a systematic and objective method to help develop or adapt the coarse-grained models is needed. We present an automatic method that implements an evolutionary algorithm to find a set of optimal force field parameters for a one-bead coarse-grained model. In addition to an optimized force field, parameter correlations and significance of the potential energy terms can be determined. The method is applied to two classes of problems: the dynamics of an RNA helix and the RNA structure prediction.



1. INTRODUCTION

In vivo dynamical processes involving nucleic acids last from nanoseconds to hours and involve even hundreds millions of atoms.[1] The common *in silico* method to investigate the dynamics of biomolecules is all-atom molecular dynamics (MD),[2] but with this method, either millions of atoms or microsecond time scales can be achieved, if a state-of-the-art supercomputer is used.[3,4] Not only computational resources but also the nature of the MD methods limit the simulation scales. All-atomistic force fields used in MD are not bare of artifacts that lead to erroneous behavior in long simulations, especially for RNA (e.g., see the formation of RNA ladder structures in ref 5).

In classical all-atom MD simulations of solvated systems, 90% of computational time is devoted to simulating the water molecules.[6] Thus, the methods that use implicit representation of interactions with solvent are often applied to decrease the simulation time. Implicit water descriptions are achieved by using continuum solvation. The generalized Born model or distance-dependent or sigmoidal dielectricts can be applied.[7] Solvation effects can be also included in the potential energy function through the solvent accessible surface area dependent term or by adding viscosity to the equations of motion as in Langevin dynamics.[2,7] To further speed up the computations, besides eliminating the solvent, the representation of a solute can be simplified, using the so-called coarse-grained (CG) models. In the CG models, sets of atoms are grouped into single interacting centers called pseudoatoms. This grouping reduces the number of degrees of freedom and consequently the simulation time by one to 2 orders of magnitude depending on the degree of simplification. Furthermore, the exploration of the conformational phase space is intrinsically enhanced as a consequence of smoothing of the free energy surface. Various CG models for polypeptides and nucleic acids have been designed for and successfully used in MD simulations (e.g., refs 8−16 or in review articles 17−20). Nevertheless, it is difficult to combine accuracy, transferability and predictivity of these models in such a small number of parameters. This task requires using rather complex potential energy functions, often including multiwell potential energy terms.[21−23]

Coarse-grained RNA models differ in the number and placement of the interacting centers and the functional forms associated with the force fields terms.[19,24] The FF is chosen based on a particular application and biological problem one wants to examine. Some FFs aim for predicting a 3D structure of RNA,[25−31] and others have been designed to determine the internal dynamics of an already folded RNA.[32−35] Universality of the models decreases with the reduction of degrees of freedom. Only the model with eight-beads-per-nucleotide is able to simulate at the same time and with the same set of FF parameters both the dynamics and folding of RNA,[28] which is impossible with a one-bead-per-nucleotide representation. Typically, these CG models do not have that wide range of

Received: June 13, 2013
Published: September 27, 2013

applications as full-atomistic ones such as Amber[36] or CHARMM.[37] The FF transferability depends on many factors such as the number of beads and their placement and explicitly represented interactions. The procedure to design a FF is often based on manual trial and error tests driven by the knowledge of physics and chemistry.[19] We propose to replace these time-consuming procedures by automatic design and parametrization.

To automatize the FF design procedure, two main components (besides the FF itself) have to be included. First, one needs an efficient way to assess the quality of the tested FF. A fitness (or objective) function has to be defined that provides a metric of the quality of a CG FF. Moreover, the terms considered in this metric function need to be balanced, for example, numerically to the same scale or effectively normalized so large terms such as energy do not swamp the others. The fitness function is crucial for the automatized FF optimization because it is a measure to compare CG FFs. In most cases, depending on the FF purpose, this function will include many terms, for example, the root-mean-square deviation (RMSD) to compare the deviation of the 3D structure from a known template, the CG FF forces[38] or energies[39] to compare with their respective values from all-atom MD. Distribution functions of the internal variables are additional important quality measures that can be compared with reference experimental or atomistic simulations. In the case of manual FF optimization, the shapes of the distributions are typically visually analyzed, but in an automated case, the comparison must be automatic and also reduced to a single number estimating the similarity.[40] For biomolecules, a single term taken to assess the quality of the FF is not sufficient and best results are achieved by combining different measures into a single final score.

The second component of the automatized procedure is the FF optimization algorithm, that is, the way how a FF parameter set (with the optimal score) is found by the program. Many methods are available, and some of them are available as software packages.[41−44] In general, there are three different approaches that might be applied: analytic solution, systematic local search, and metaheuristics.

The analytical approach is based on the assumption that one can define a direct mathematical relation between the scoring function and the FF parameters. A typical example of this approach is the Boltzmann inversion (BI). The BI method introduces a relation between the distance distribution function $d(r)$ and the potential energy $U(r)$,

$$U(r) = -k_B T \ln\left(\frac{d(r)}{d_0(r)}\right) \tag{1}$$

where k_B is the Boltzmann constant, T denotes temperature and $d_0(r)$ is a reference distribution. A few CG FFs for nucleic acids[23,32,45] have been parametrized by applying eq 1 to distributions taken from a set of X-ray resolved structures. However, this approach has two major limitations. First, parametrizing FFs in that way neglects the correlations between different potential energy terms. Second, it requires *a priori* knowledge of the reference distribution $d_0(r)$, which can be easily derived for simple polymers but not for biomolecules with complicated 3D architectures. These limitations of the BI method necessitate manual optimization followed by trial and error postprocessing. However, BI can be also used in an iterative manner[40,43] to at least partially overcome the mentioned problems. In the iterative BI method eq 1 is used to correct the potential energy function until the reference and calculated distance distributions agree reasonably well. The iterative BI method has been successfully applied to polymers.[40,43] However, besides problems with the reference distributions and correlations, this approach generates FFs consistent with the distribution functions while does not target other important quantities of biological systems. Another example of analytical approach is the force-matching method in which the parameters for a coarser model are found by comparing the forces with a reference finer MD simulation. Even though the method was successfully applied for the FF parametrization,[46] it shares the limitation of the iterative BI method by relying only on a single type of reference data.

In systematic local optimization methods, the best FF is also found by iteratively updating the scoring function until a set of parameters that cannot be further improved is found. However, the systematic local search, in contrast to the analytical method, uses general purpose mathematical algorithms and does not assume any relation of the FF parameters with the scoring function. These make the calculations more expensive. CG-OPT[41] is a CG automatic optimization workflow that uses a simplex algorithm and was shown to effectively find the CG FF for polyethylene. A more sophisticated local optimization is implemented in GROW[42] where the user can choose from multiple gradient-based algorithms. These algorithms were also proven effective for polymers but they would be hard to apply for large biomolecules. Directed search (specifically the gradient-based one) becomes computationally too expensive when the number of parameters increases. Also, it requires a good first guess of parameter values because only the closest local minimum will be found. One more important local optimization approach is the renormalization group method.[47] In this approach, a covariance matrix between the FF parameters is used to explicitly account for the correlations between different potential energy contributions. This method is resistant to MD simulation noise and has been successfully applied to develop a one-bead DNA model. The potential energy function is defined as a linear combination of terms and an optimal value of each term coefficient is found but the parameters inside a particular term are not optimized. This limitation may lead to overparameterization because to compensate for inaccuracies of nonlinear parameters more linear terms are added complicating the potential.

Metaheuristic methods, instead of a systematic and analytical way of finding an optimal solution, perform an educated guess. These methods do not guarantee finding an optimal solution but by smart design of the algorithm may find a good enough result in the cases where systematic methods fail. Also, metaheuristic methods are beneficial for the problems with a multiple minima potential energy landscape[48] and inaccurate first guess of the parameters. The examples of applications range from bioinformatics[49] and finances[50] to art[51] and gaming.[52] The most widely used are evolutionary algorithms (EA) based on Darwinian natural selection scheme.[48] A population is introduced, each consisting of a parameter set (genome) and a fitness function value (phenotype). Each iteration of the algorithm represents a generation — organisms with the best phenotype have a higher chance of mating since the mixture of good solutions has a higher chance to be the best one. The probabilistic selection of organisms for mating introduces additional variety to the population and prevents falling into a single local minimum. After mating, random

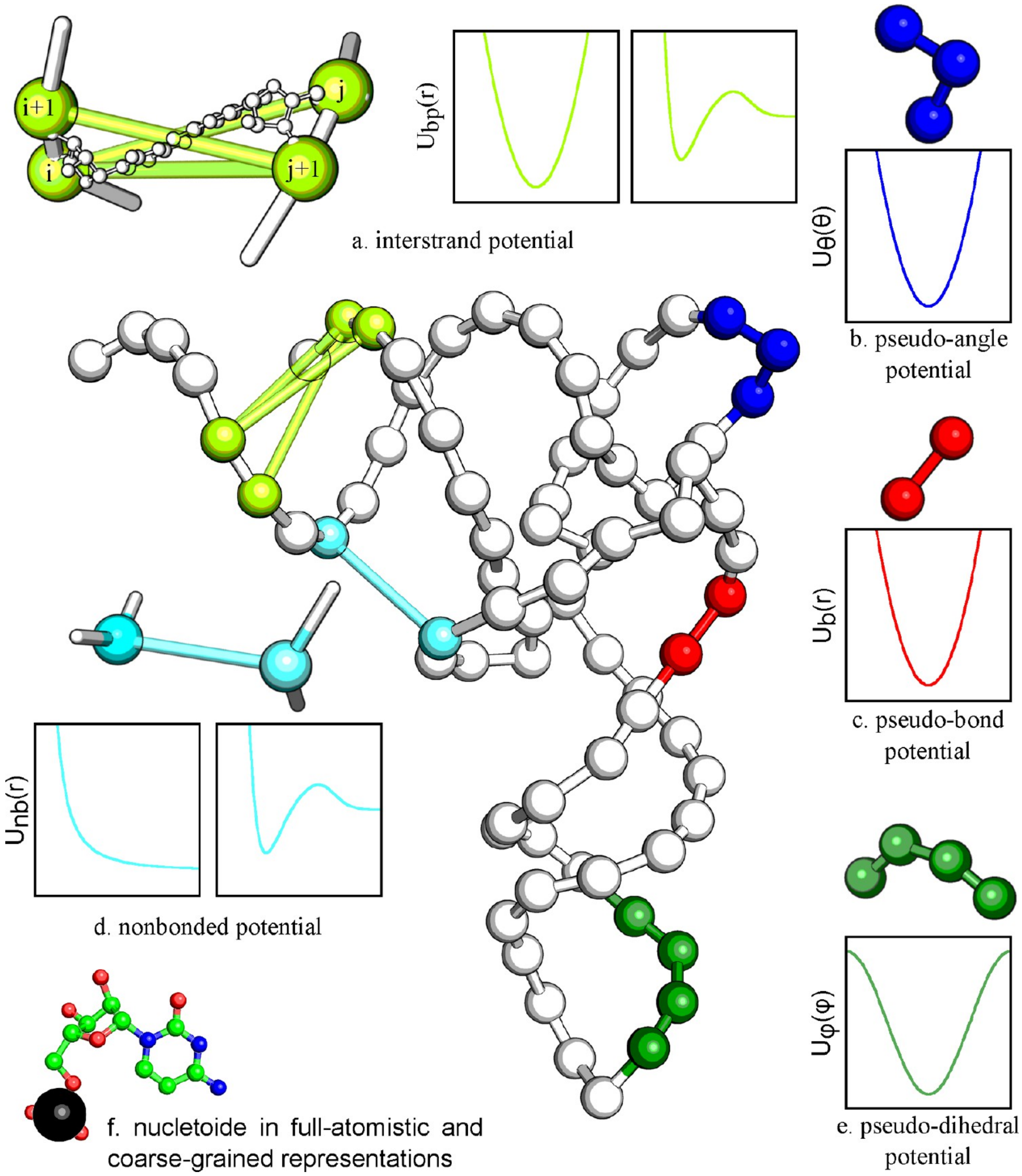


Figure 1. Coarse-grained model of the yeast phenylalanine tRNA (PDB ID: 6TNA). The insets sketch the functional forms of the interactions used in the CG simulations (*x*-axes are the respective variables and *y*-axes are potential energy functions, units are omitted and arbitrary). In the lower left corner, the heavy atom representation of a single base is shown with the black bead (positioned on the P atom) denoting the pseudoatom used in this CG model.[55,56]

mutations are applied to introduce new genes and extend the reach of the algorithm outside of the pool of values from the previous generation. A variation of this method, genetic programming (GP), has been already applied to a CG FF optimization problem.[53,54] GP finds the numerical parameters of a CG FF and automatically constructs and tests the mathematical expressions that describe the interactions. However, the computational cost involved with GP makes it unfeasible for biological molecules since GP is useful only to optimize the potential energy functions for theoretical two or three body model systems.

On the other hand, EAs might work well in optimizing the complex multiargument functions used to describe the potential energies of biomolecular conformations, and they are easy to parallelize for high performance computing. With EAs, one can generate data for statistical analysis, and EAs can be extended to account for multiobjective nature of the problem. In this work, we report an EA specifically designed to parametrize one-bead CG models for ribonucleic acids. We describe its implementation and applications to several different RNA structures showing its performance and potential. Although the present implementation focuses on RNA, the extension to DNA and proteins is straightforward.

2. METHODS

2.1. One-Bead Coarse-Grained RNA Potential Energy Function.

To test the applicability of the EAs to the parametrization of CG FFs, we selected a one-bead per nucleotide RNA model. This resolution allows for microsecond time scales, at the same time keeping the possibility to explicitly describe structural transitions. The previous one-bead DNA model[23] has proven successful in the studies of the supercoiling and denaturation of DNA, and for this reason, it was extended and adapted here to study RNA. The details of the CG FF are presented in Figure 1.

Pseudoatoms are localized in the positions of phosphorus atoms. The nucleotide type, associated with a particular pseudoatom, is stored in a CG representation file, but the FF does not differentiate between the base types. The FF is inspired by classical all-atom FFs with a CG model-specific term describing the hydrogen bond interactions found in the all-atom structure (see Figure 1 for the topology of the hydrogen bonds and functional forms of the FF):

$$U = U_b + U_\theta + U_\phi + U_{bp} + U_{nb} \tag{2}$$

$$U_{bp} = U_{i:j} + U_{i:j+1} + U_{i+1:j+1} \tag{3}$$

2.1.1. Intrastrand Potential. The intrastrand interactions among two, three, and four subsequent beads are effective interaction potentials that aim to capture the physics of covalent bonding, base stacking, and ultimately to account for the local flexibility of the single strands. Pseudobonds and pseudoangles have a harmonic potential form:

$$U_b(r) = \frac{1}{2}k_r(r - r_0)^2 \tag{4}$$

$$U_\theta(\theta) = \frac{1}{2}k_\theta(\theta - \theta_0)^2 \tag{5}$$

where k_r and k_θ denote force constants, r_0 and θ_0 are respectively the equilibrium distance and angle.

For the pseudodihedral term, we use a single-peaked cosine potential of the form:

$$U_\phi(\phi) = k_\phi(1 - \cos(\phi - \phi_0)) \tag{6}$$

where k_ϕ is half of the energy difference between the minimum and maximum and ϕ_0 denotes an equilibrium angle (the sign of this angle differentiates between the right- and left-handed RNA helix). As shown in Figure 1, this potential is similar to the harmonic one but better describes the periodicity of the pseudodihedral potential. This single-peaked formula is consistent with the data collected from the ensemble of PDB structures[55,57] containing RNA where the pseudodihedral distribution is composed of mainly one wide peak.

2.1.2. Base Pairing. The base pairing (interstrand) term describes the stabilizing forces in the helix (pairing of complementary bases and partially the stacking forces) and the electrostatic repulsion of the backbones. Since the model and interactions have no base specificity, the secondary structure has to be externally supplied. Each complementary base pair is modeled using three pseudobonds to correctly maintain the helical twist. These are: *i*:*j*, *i*:*j*+1, *i*+1:*j*+1 where *i* and *j* are the indices of the phosphorus atoms belonging to the bases *i* and *j* that define a complementary pair (Figure 1). We describe these interactions using two different potential energy functions, either a harmonic potential

$$U_{i:j}(r) = k^{i:j}\frac{1}{2}(r - r_0^{i:j})^2 \tag{7}$$

with symbols defined as in eq 4 or a Morse potential with a barrier and repulsive tail (further referred to as a well−barrier potential)

$$U_{i:j}(r) = (U_{\text{Morse}}^{i:j}(r) - U_0^{i:j}c^{i:j})sw^{i:j}(r) \tag{8}$$

$$U_{\text{Morse}}^{i:j}(r) = U_0^{i:j}[1 - \exp(-\alpha^{i:j}(r - r_0^{i:j}))]^2 \tag{9}$$

$$sw^{i:j}(r) = \frac{1}{2}[1 - \tanh(\lambda^{i:j}(r - r_1^{i:j})) \tag{10}$$

where $U_0^{i:j}$, $r_0^{i:j}$ and $\alpha^{i:j}$ control the shape of the original Morse potential, $U_0^{i:j}{\cdot}c^{i:j}$ is the energy difference between the bound and unbound state, $\lambda_{i:j}$ controls the slope of the switch function, $U_1^{i:j}$ controls the switch function energy difference, and $r_1^{i:j}$ the position of the switch. For a graphical description of $U_{\text{Morse}}^{i:j}(r)$ and $sw^{i:j}(r)$, see Supporting Information Figure S1a. Similar equations (either harmonic or well−barrier) are used for *i*:*j*+1 and *i*+1:*j*+1 pseudobonds. The harmonic form (eq 7) is faster to calculate but does not allow for interstrand bond breaking. Equation 8 requires more computer time but is more general because enables interstrand bond breaking leading to duplex denaturation. Another advantage of using this well−barrier function is that one can describe anharmonic fluctuations.

RNA, in contrast to DNA, forms many hydrogen bonds that do not follow the Watson−Crick base pairing scheme.[58] These noncanonical bonds are crucial for the stabilization of RNA tertiary motifs[59−61] and lead to an abundance of RNA structures. In their work on the CG RNA model, Jonikas et al.[27] suggest that the knowledge of at least few tertiary interactions is necessary to correctly predict the 3D structure of RNA. Therefore, in our model, we included the possibility of adding noncanonical bonds. These noncanonical bonds have to be inputted *a priori*, for example, based on experimental data[62−64] or crystal structure.[65] These bonds are weaker compared to canonical ones so we represent them as single *i*:*j* bonds. For each noncanonical bond its equilibrium distance is taken either from the X-ray data or atomistic simulation, and the energy parameters are optimized independently from the canonical bond parameters.

2.1.3. Nonbonded Potential. The nonbonded potential energy term is the most challenging one to parametrize because it has to implicitly account for the hydrophobic and steric effects as well as the electrostatic repulsion.[2] Similar as for base pairing we implemented the nonbonded potential using two different functional forms. The simpler form is the Coulomb repulsive function:

$$U_{nb}(r) = \frac{q_1q_2}{4\pi\varepsilon_0\varepsilon_r r} \tag{11}$$

where the only parameter that changes during the optimization phase is the relative permittivity (ε_r). The charges on the beads are set to $q_1 = q_2 = -1$ (in electronic charge units), and ε_0 is the vacuum permittivity. This potential accounts for the repulsive interactions of negatively charged phosphorus groups. Due to the presence of relative permittivity, it can also implicitly account for dielectric effects of the environment.

We have also implemented the more complex well−barrier potential. As previously shown for DNA,[23] this form allows fine-tuning of the hydrophobic and repulsive interactions, leading to correct minor and major groove sizes and global stability at different temperatures (ranging from 300 K to the melting point). We have used a potential similar in shape to the one defined in eqs 8 and 10, but to achieve faster calculations (as this is the most frequently calculated potential), we have applied a piecewise function with a well and barrier shape, composed of four parts:

$$U_{nb}(r) = \begin{cases} k(r - r_0)^2 + U_{nb}^0 & \text{if } r < r_0 \\ (U_{nb}^{\text{bar}} - U_{nb}^0)\exp(-\sigma_1(r - r_1)^2) + U_{nb}^0 & \text{if } r_0 < r < r_1 \\ U_{nb}^{\text{bar}} & \text{if } r_1 < r < r_2 \\ U_{nb}^{\text{bar}}\exp(-\sigma_2(r - r_2)^2) & \text{otherwise} \end{cases} \quad (12)$$

where U_{nb}^0 denotes the potential well depth (negative number), U_{nb}^{bar} is the energy barrier height, r_0 is the energy minimum position, r_1 is the barrier starting position, r_2 is the barrier end position, k controls the slope on the left side of the minimum, σ_1 and σ_2 control respectively the left and right slope of the barrier (see Supporting Information Figure S1b). To ensure the continuity, the $(r_0 - r_1) > 3/(2\sigma_1)^{1/2}$ condition has to be satisfied. This implementation is computationally faster and better for automatic optimization in contrast to the form used in the DNA model.[23] Furthermore, eq 12 better describes the broad, RNA-specific nonbonded BI distributions used as a first guess of the corresponding interaction potential.

2.2. Fitness Function.

Along with the potential energy formula we have introduced the fitness (or objective) function that describes how well the results of simulations using the CG FF correspond to experimental data or full-atomistic simulations. Various aspects of the potential energy terms are considered in the total score and in our implementation the fitness function is a weighted sum of multiple terms:

$$f_{\text{tot}} = \sum_{i=1}^{n} w_i f_i' \quad (13)$$

$$f_i' = \begin{cases} 0 & \text{if } f_i < l_i \\ \dfrac{u_i - f_i}{u_i - l_i} & \text{if } l_i < f_i < u_i \\ 1 & \text{if } f_i > u_i \end{cases} \quad (14)$$

where w_i is a weight satisfying $\sum_{i=1}^{n} w_i = 1.0$, f_i' is the normalized fitness function (phenotype) with l_i and u_i the fixed lower and upper bounds used to rescale f. The functions, the number of terms, as well as weights can be chosen by the user from a set of predefined objective functions, depending on a particular application of the CG FF. The bounds influence the value of the fitness function; extending the l_i and u_i bounds makes f_i' function less susceptible to the f_i function variations.

In the following, we present the functions that are the building blocks of the total fitness function. Each term can be used many times.

2.2.1. Distance Distribution Functions.

Distance distribution functions are often used as target observables to parametrize CG FFs.[23,27,40,43] The aim of the optimization is to find a CG FF that best resembles the reference distance distribution. The source of the reference distribution may be an ensemble of experimentally determined structures or generated by full-atomic simulation. Here, specific contributions to the total radial distribution function coming from single distances (e.g., minor groove distances) are used as targets of the EA.

For the automated approach, one needs a measure that describes the difference between two distributions. This numerical value has to capture multiple features of the distribution. Even if we compare distribution functions that have a simple bell-shape, similar to most distribution functions for RNA that correspond to a particular pairwise interaction, there are two independent similarity measures, the position of the maximum and the variance, but using both would increase the number of terms. We have therefore chosen the Hodgkin index (HI)[66] because it gave the best balance of sensitivity between the average and variance:

$$f_{\text{HI}} = \frac{2\sum_{i=1}^{N} d_1(i)d_2(i)}{\sum_{i=1}^{N} d_1^2(i) + \sum_{i=1}^{N} d_2^2(i)} \quad (15)$$

where d_1 and d_2 are the compared distributions.

2.2.2. Root Mean Square Deviation.

The root-mean-square deviation (RMSD) measures the structural difference between two conformations of the same set of atoms and is defined as

$$\text{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mathbf{r}(n, t) - \mathbf{r}_{\mathbf{ref}}(n))^2} \quad (16)$$

where $\mathbf{r}(n, t)$ is a position of n^{th} pseudoatom (out of N) after simulation time of t and $\mathbf{r}_{\mathbf{ref}}(n)$ is its reference value (in our case taken from the crystal or NMR structure). RMSD averaged over simulation time is taken for the scoring function:

$$f_{\text{RMSD}} = \frac{1}{T}\sum_{t=0}^{T}\text{RMSD}(t) \quad (17)$$

Rigid molecule translation and rotation are removed before calculating RMSD to account only for the contributions coming from internal dynamics. This is done analytically using the Kabsch algorithm.[67] The quality of the CG FF in the case of structure prediction is measured by comparing RMSD between the obtained prediction (starting from an unfolded structure) and the crystallographic reference. However, using a single reference structure and considering only RMSD in the fitness function bias the simulation toward small fluctuations. For this reason additional target quantities must be considered in the fitness function.

2.2.3. Root Mean Square Fluctuations.

Root mean square fluctuation (RMSF) measures the internal mobility of atoms:

$$\text{RMSF}(n) = \sqrt{\frac{1}{T}\sum_{t=0}^{T}(\mathbf{r}(n, t) - \mathbf{r}_{\mathbf{ref}}(n))^2} \quad (18)$$

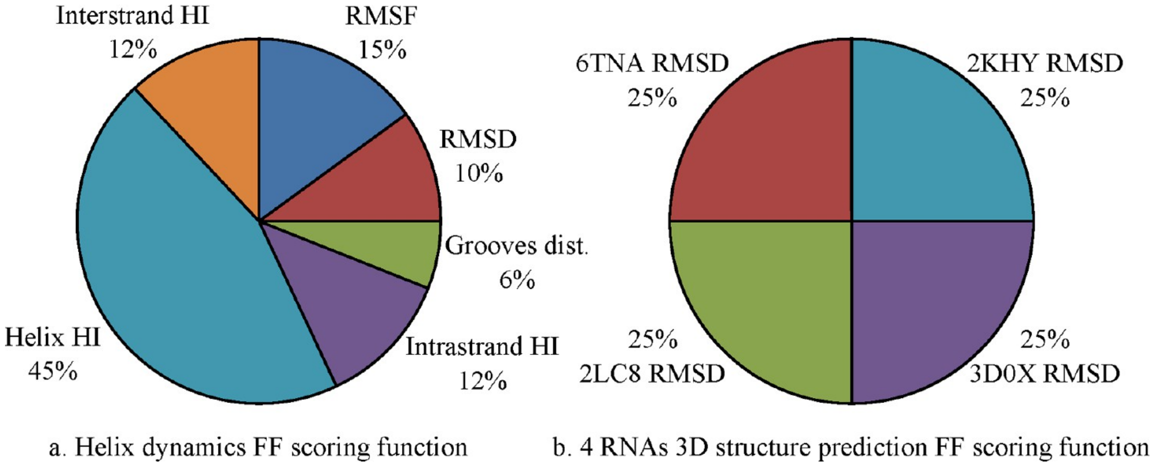
here $\mathbf{r}(n, t)$ is a position of n^{th} pseudoatom (out of N) after simulation time of t and $\mathbf{r}_{\mathbf{ref}}(n)$ is a reference value (in our case it is the average position of the pseudoatom during MD). Crystallographic β-factors of i^{th} atom are related to its RMSF according to the formula:

$$\beta(n) = \frac{8\pi^2}{3}\text{RMSF}(n)^2 \quad (19)$$

We define the fitness function as the difference between the simulation- and crystallography-derived RMSF, averaged over all atoms:

$$f_{\text{RMSF}} = \frac{1}{N}\sum_{n=1}^{N}\left|\text{RMSF}_{\text{simulation}}(n) - \sqrt{\frac{3}{8\pi^2}\beta(n)}\right| \quad (20)$$

Chart 1. Scoring Function Weights (in %) Used in the Two Categories of the EA Optimization Tests. (a) Fitness Function Applied in the Studies of Equilibrium Dynamics of the RNA Helix (for the definition of terms see Supporting Information Table S1). (b) Fitness Function Used for the Structure Prediction (see Figure 2 for 3D structures corresponding to PDB codes)



2.2.4. Distance Constraints. The previously defined functions concentrate on global, averaged properties of a molecule. In some cases, the ability of a CG FF to stabilize a particular bond or motif may be important. We have therefore added a distance constraining term that compares the average distance of two pseudoatoms $\bar{x}$ with a reference distance x_{ref}:

$$f_{\text{dist}} = |\bar{x} - x_{\text{ref}}| \quad (21)$$

2.3. Evolutionary Algorithm. EA is a metaheuristic optimization algorithm that was invented based on Darwinian evolution of species.[48,68] A general outline of the EA procedure, used in the presented work is the following:

1. A set of CG FFs with random parameters is created (first generation). The number of the FFs in the set is controlled by the population size, which is constant for all subsequent iterations (generations).
2. CG MD simulations of selected RNA molecules are performed for each of the FFs created in step 1.
3. The value of the fitness function (the phenotype) is calculated based on the outcome of the CG MD simulations.
4. Steps 2 and 3 can be repeated using the same FFs to average out random effects.
5. Based on the fitness function the FFs are ranked and a new generation of CG FFs is generated by propagating without any alterations the best FFs (elitism operation), supplemented by a new set of FFs based on the previous one (operations of mating and mutation).
6. The algorithm is iterated from step 2 for a number of generations defined at the beginning.

For more technical details, see Section 1 and 2 in the Supporting Information.

2.4. Parameterization Protocol. To find out if EA is suitable for automatic parametrization of CG FFs, we have applied it to determine a one-bead FF for two different simulation problems: an MD simulation of equilibrium (near-native state) dynamics and tertiary structure prediction. All CG simulation were performed using an in-house CG MD engine, RedMD[69] (http://bionano.cent.uw.edu.pl/Software/RedMD).

2.4.1. Equilibirum Dynamics Simulations of an A-RNA Helix. In this test, a one-bead model was parametrized to best mirror the dynamics of an RNA helix in comparison with an all-atom MD trajectory. A 35 base pair A-RNA helix was created using Nucleic Acid Builder in Ambertools.[70] A random sequence was used, with a comparable number of nucleotide types to avoid a bias for the A−T or C−G pairs since the CG model is not base specific. The helix was immersed in a truncated octahedral box of explicit TIP3P[71] water and neutralized by adding 70 Na^+ ions. Two full-atomistic MD simulations were performed, one with only neutralizing ions included and one with excess Na+ and Cl^- added to yield an approximately 0.1 M ionic strength.

Full-atomistic MD simulation was performed using Amber ff99 FF[36] with the bsc0 and χ-OL corrections[5,72] and NAMD 2.8.[73] The thermalization and equilibration protocols were based on our previous study.[74] The SHAKE[75] algorithm was used with a 2 fs time step. The NVT ensemble was used with Langevin thermostat and a damping constant of 1 ps^{-1} at 310 K. Periodic boundary conditions were applied with the Particle Mesh Ewald method[76] for long-range interactions with a grid spacing of about 1 Å and a 10 Å short-range cutoff for nonbonded interactions. The simulation consisted of a 100 ps thermalization and 900 ps equilibration phase, with a gradual decrease in restraints on the RNA heavy atoms and continued with 100 ns production runs.

RMSF, RMSD, and distance distribution data were extracted from the production phase trajectory to compose the scoring function (see Chart 1a and Supporting Information Table S1). The scoring function was constructed to emphasize the dynamic properties, that is, the RMSF and distance distributions of atoms interacting via a minor or major groove. Also, structural parameters were taken into account, such as RMSD and average groove distances, to make sure that the helix is stable in the CG MD simulation.

The CG MD simulation was carried out in the NVT ensemble with Langevin thermostat (damping constant 2 ps^{-1}) at 310 K with a 10 fs time step. Translations and rotations were removed every 200 MD steps. The CG MD simulation was preceded by minimization, but no thermalization was applied. The CG MD simulation time varied from 1 ns to 50 μs. The total score was averaged over 10 consecutive runs of the same system with the simulation differing only by a random number generator seed. This procedure was used to eliminate any

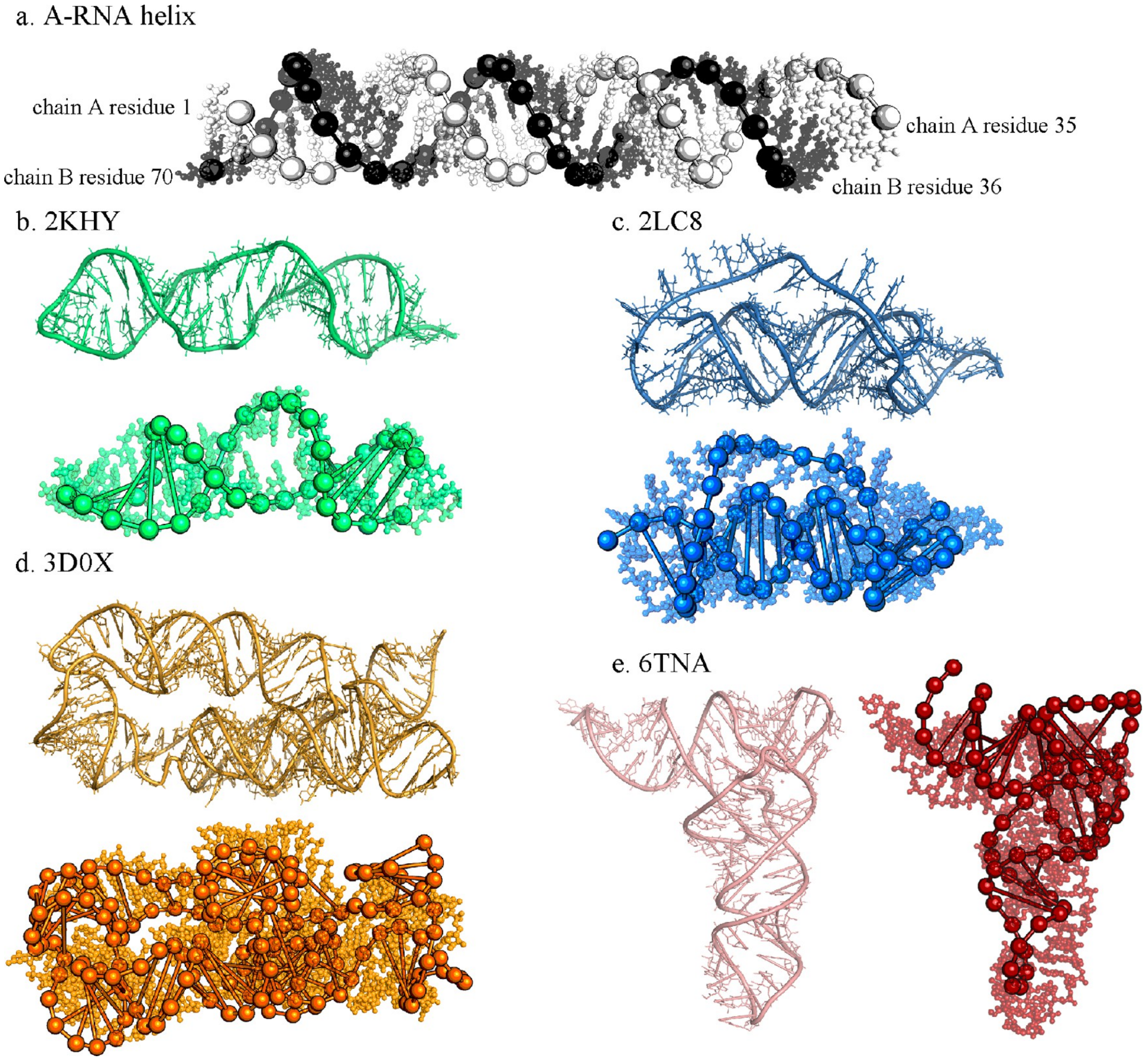


Figure 2. Following RNA structures were used as a training set of RNA CG FF: (a) 35 base pair A-RNA helix created using Ambertools[70] in original conformation, (b) specifier loop domain of tyrS T box leader RNA (PDB ID: 2KHY),[77] (c) murine leukemia virus readthrough pesudoknot (PDB ID: 2LC8),[78] (d) unbound lysine riboswitch (PDB ID: 3D0X),[79] and (e) yeast phenylalanine tRNA (PDB ID: 6TNA).[56] Structures b−e are shown in full-atomistic representation and in a CG representation that was found using the structure prediction CG FF developed in this work. Structures b−e are later denoted as "4 RNAs" learning/training set.

random effects coming from the initial velocity assignment and thermostat.

2.4.2. Tertiary Structure Prediction. In this test, the ability of the CG model to predict the 3D structure of selected RNA molecules was assessed. The task was to find the final structure without investigating the mechanism of RNA folding.

For the structure prediction test, a training set of 4 RNAs was selected from the Protein Data Bank:[55] a yeast phenylalanine tRNA (PDB ID: 6TNA),[56] a specifier loop domain of the tyrS leader RNA (PDB ID: 2KHY),[77] a murine leukemia virus readthrough pesudoknot (PDB ID: 2LC8)[78] and an unbound lysine riboswitch (PDB ID: 3D0X)[79]—see Figure 2. 6TNA and 3D0X were solved using X-ray diffraction methods and 2KHY and 2LC8 using solution NMR. These structures are of varying size (respectively 76, 38, 58, and 161 nucleotides) and complexity. 2KHY is the simplest with only a hairpin loop. 2LC8 includes also the interaction of two helices via a pseudoknot. 6TNA contains multiple motifs and a complex interplay of noncanonical bonds. 3D0X is the largest and contains helix−helix interaction via noncanonical bonds. The topologies of hydrogen bonds for these structures were obtained with RNAView.[65]

The CG FF MD simulation was carried out in the NVT ensemble with Berendsen thermostat ($\tau = 1ps$) and a 10 fs time step. Three simulation variants were tested: constant temperature of 310 or 100 K and a gradual decrease from 310 to 0 K. The last variant employs the idea of simulated annealing—high temperature in the beginning enabling conformational freedom and final low temperature for convergence to a local minimum. The simulation time varied from 10 ps to 1 ns. Translations and rotations were removed every 200 steps. The simulation was preceded by minimization but no thermalization was applied.

MD simulation started from an unfolded state. The starting structure either formed a circle (with the circumference equal to the number of nucleotides times 5 Å) or the CG beads were placed in a cubic grid, trying to position each bead close to its sequence and secondary structure neighbors. As a control, we have also started the tertiary structure prediction simulation from an already folded reference structure to find the limits of the performance of a CG model.

Since the fidelity of the predicted structure is the most important measure, the scoring was based only on RMSD. The RMSD scoring terms referred either to all four prediction targets with an equal weight (see Chart 1b) or only to 6TNA. This was done to find out if narrowing the FF task improves the results. RMSD was calculated for the last simulation frame (50% of the score) and 10 times through the whole production run (50% of the score). The score was averaged over 20 consecutive runs with the same CG MD simulation setup.

2.5. Parameter Ranges. To define physical ranges of parameters we have performed the BI procedure on 342 RNA structures from PDB[55] with resolution of at most 3.5 Å. The nonredundant RNA 3D structures were chosen according to the FR3D database.[57] For this set, we calculated the distance distributions (eq 1) and the first guess of the potential energy function parameters. Since these were used only as a rough estimate, the simplest reference distribution, $d_0(r) = 4\pi r^2$, was applied (although more appropriate forms exist for small and nonglobular systems[80,81]). The parameter ranges that could not be found by BI were taken from the CG DNA model;[23,82] these were some parameters of the well−barrier potential of eqs 8−10 and 12. The parameter ranges are shown in Supporting Information Tables S2−S5.

3. RESULTS AND DISCUSSION

We present the results of the EA optimization for the RNA FFs applied to study the RNA dynamics and folding. We describe the single best potential and then show how analyzing multiple FFs may contribute to understanding of the CG model. The designed parametrization procedure is completely automatic and the FFs do not require any manual postprocessing of parameters. The convergence of the EA algorithm was tested and fine-tuning of the applied EA parameters is described in Section 3 of Supporting Information and in Figure S2.

3.1. Equilibrium Dynamics of the RNA Helix. We optimized the CG FF for the RNA helix and addressed three specific questions: (a) Does the well−barrier potential for interstrand and nonbonded interactions influence the results of a CG RNA simulation compared to simpler functional forms? (b) What are the one-bead CG FF parameters for a physically reliable MD simulation of the RNA helix? (c) Are the potential energy terms correlated with each other and do we capture these correlations in the CG model?

3.1.1. Analysis of the Intrastrand and Nonbonded Potentials. To address question a, we performed six concurrent FF optimization runs, three with the harmonic potential (eq 7) and three with the well−barrier potential (eq 8) for the interstrand base pairing. For each group a different nonbonded potential was used: an electrostatic repulsive potential (eq 11) or a well−barrier (eq 12). The results of this comparison are presented in Table 1 and the optimized interstrand and nonbonded potentials in Figure 3.

Table 1. Score for the Best CG FF Parameterized Using the EA Procedure after 64 Generations with 1024 Population Members[a]

interstrand potential	nonbonded potential	best score
harmonic (eq 7)	well−barrier (eq 12)	0.167 ± 0.023
harmonic (eq 7)	repulsive (eq 11)	0.065 ± 0.009
well−barrier (eq 8)	well−barrier (eq 12)	0.129 ± 0.010
well−barrier (eq 8)	repulsive (eq 11)	0.061 ± 0.007

[a] Optimization was performed for the RNA helix presented in Chart 1a. The scoring function was calculated based on 1 ns MD simulation of the helix. Different functions were used for interstrand and nonbonded potential, as well as different range of parameters. Respective equations are in parentheses.

Table 1 shows the lowest score of 0.061 ± 0.009 for the well−barrier interstrand potential and repulsive nonbonded. However, using a harmonic interstrand potential leads to a similar score of 0.065 ± 0.007. No difference between the two might be attributed to simulation conditions, especially moderate temperatures, in which the A-RNA helix forms a stable duplex. Therefore, the breakable well−barrier potential is not crucial because denaturation is not expected. The other difference between the potentials, anharmonicity of the well−barrier potential, in contrast to symmetric harmonic potential, does not play a major role in the final score. The two interstrand potentials are visually compared in Figure 3a for the *i*:*j* interaction. All potentials, except the harmonic one parametrized together with the well−barrier nonbonded, show the energy minimum position at approximately 19 Å and similar force constants.

For the nonbonded potential, as shown by the best score in Table 1, the repulsive potential outperforms the more complex well−barrier potential by a factor of 1.3. A possible explanation is shown in Figure 3b. For the unbound beads the optimized repulsive potential provides a short-range repulsion, but further than 10 Å, the interactions vanish. For the well−barrier potential, an additional stabilization is provided for the beads that interact through the minor groove,[23] that is, in the range 9−13 Å (depending on the structure source, with lower values for X-ray structures and higher for NMR resolved ones). The worse score obtained for the nonbonded well−barrier potential suggests that additional groove stabilization is not necessary. However, one cannot rule out that, since this potential depends on 8 parameters and correlations between the nonbonded and interstrand terms do exist, EA was not able to find the parameter combination leading to a lower score.

If the repulsive nonbonded potential is used, the best score is not influenced by the type of the interstrand term (Table 1). On the contrary, if the well−barrier is used for the nonbonded term the best score depends on the interstrand term (either harmonic or well−barrier). The use of the well−barrier interstrand term improves the score by 22% in comparison to the harmonic interstrand term.

Overall, the analysis of the nonbonded potential functions optimized for the RNA near-native state dynamics shows that the electrostatic repulsive potential outperforms a more tailored well−barrier nonbonded interaction in contrast with the previous work on the DNA model.[23,82] This discrepancy can be due to (i) different physical quantities used to optimize the

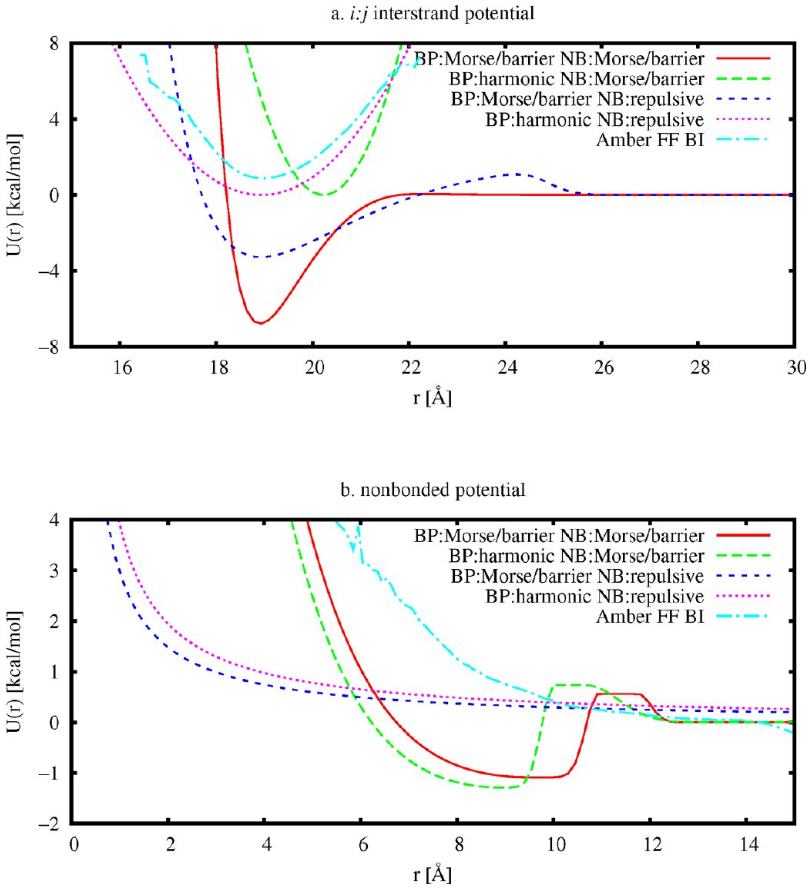


Figure 3. Optimal potential energy functions for the RNA helix to simulate its equilibrium dynamics. The fitness function is shown in Chart 1a. The following potentials are shown: (a) the *i*:*j* interstrand potential and (b) the nonbonded potential. BP stands for the interstrand base pairing potential and NB for the nonbonded one. Amber all-atom FF BI represents the potential of mean force derived using BI method from all-atom MD simulation of this helix.

two FF parameters (the DNA model included also thermodynamic information) and (ii) the incomplete parameter space exploration in the case of the well−barrier nonbonded potential. Note, that the presented scores depend on the scoring function, simulation conditions, and RNA structure. Here, the interstrand bonds were not stretched far from equilibrium. If higher temperatures were applied, then the difference between the well−barrier and harmonic potential would be larger because the well−barrier term could break. In the presented MD case, there is no significant improvement upon introducing the well−barrier potential for the interstrand or nonbonded terms; however, such potential can be better for other tasks than the repulsive or harmonic functions. Correlations among the nonbonded and base-pair potential parameters, described by 26 parameters in total, can be an additional source of incomplete sampling.

3.1.2. One-Bead CG FF Parameters for the RNA Helix. To address question b, we provide the optimal set of CG FF parameters optimized by EA for the harmonic interstrand potential (eq 7) and repulsive nonbonded potential (eq 11). The parameters are shown in Supporting Information Table S6 (first column). To verify the quality of the CG FF, we performed 100 ns, 1 μs, and 50 μs CG MD simulations (respectively 100, 1000, and 50 000 times longer than the MD simulation used in the optimization procedure to calculate the FF score). The RNA helix was stable during the simulation with RMSD for the phosphate beads of 5.0 ± 0.25 Å for the 100 ns case, 5.2 ± 1.3 Å for the 1 μs (the reference to calculate RMSD was the all-atomic starting model, see Figure 4), and 5.4 ± 1.5 Å for the 50 μs case. The RMSD for the phosphorus atoms in the all-atom MD simulation with Amber FF was equal to 4.3 ± 1.0 Å. Figure 4 shows periodic fluctuations of RMSD, which can be attributed to bending of the whole helix. The movie of 1 μs CG MD simulation is presented as Supporting Information Movie S1. The shape of the RMSF (Figure 4 bottom) is similar for all-atom and CG MD simulations even though the fluctuations are lower in the all-atom simulation.

In Figure 5, we present the comparison of the distance distribution functions between the full-atomistic and CG MD simulations. Overall, the positions of the minima are similar, but the widths of the curves (corresponding to force constants) are different for the major groove region (see Figure 5d). The nonbonded distance distribution function (see Figure 5e) and the total radial distribution function (see Figure 5f) have the same overall shape in both cases but more energy minima are present in the distributions obtained from the full-atomistic simulations. This is expected because the reduction of the degrees of freedom must lead to less complex distance distribution functions.

3.1.3. Parameter Correlations. The last question (c) pertains to the correlations between parameters, which is important because the BI method fails to predict them. Finding the correlations can help in selecting the parameters that can be easily calculated from fitting to structural data and those that

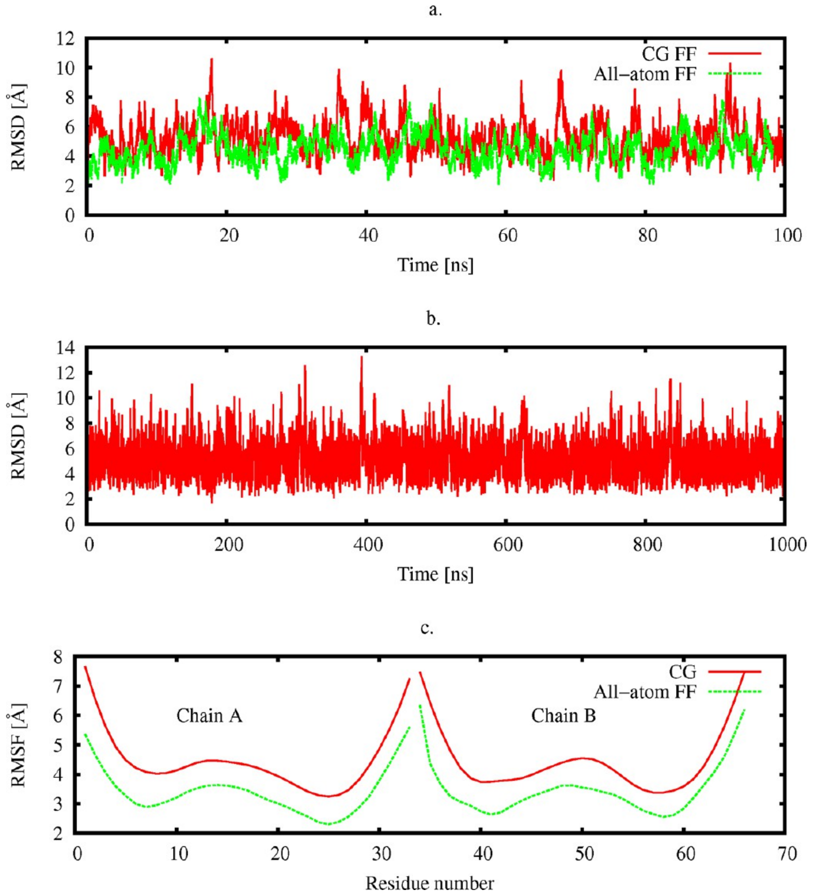


Figure 4. RMSD and RMSF of the P atoms for the 35 base pair A-RNA helix (see Figure 2a) calculated from MD simulations using the Amber all-atom FF and CG FF.

are hard to determine and require additional automatic or manual parametrization. The correlation analysis is an additional benefit of using the EA method population approach. In simulated annealing or local search methods a single solution is improved in a sequence of iterations leading to an optimal solution. However, in EA in each iteration a set of solutions is created based on a previous one so multiple pathways can be explored during the optimization. Therefore, a single generation of EA is a good input for the correlation analysis, especially for a large population size. Table 2 presents the Pearson correlation coefficients between the CG FF parameters taken from the best 20% parameter solutions, which provides the first information about possible linear correlations. In this set, there were no strong correlations (higher than 0.9 or lower than −0.9). There were, however, 42 correlation coefficients higher than 0.1 or lower than −0.1. The highest correlations, of 0.64, are between the energy minima of the interstrand potential. The first one, between interstrand minima, can be attributed to the fact, that shortening or lengthening *i*:*j* pseudobond equilibrium distance, without analogous change to *i*:*j*+1 distance, leads to a tension in the helical conformation: these interactions, together with the *j*:*j*+1 one (pseudobond on the complementary strand) form a triangle of pseudobonds connected by strong harmonic interactions and changing the distance of one has to be accommodated by extending the second one. This is consistent with the positive sign of the correlation. On the other hand dihedral angle force constant and the interstrand potential force constants are anticorrelated with the coefficient in the range from −0.31 to −0.17. Both interactions are crucial for the stabilization of the helical conformation so if the dihedral is weakened then the interstrand bonds have to be strengthened and vice versa. In principle, there might be nonlinear correlations not captured by the Pearson coefficients. However, a visual inspection of the scatter diagrams of pairs of variables does not indicate that nonlinear correlations or associations are present. A more detailed study on a possible nonlinear correlations should be performed in the future to verify whether some important relations were overlooked.

3.2. RNA 3D Structure Prediction.

We have also parametrized a CG FF suitable for the RNA tertiary structure prediction using a CG MD simulation. We have applied a scoring function described in Chart 1b and performed the following steps: (a) setting up the EA method, (b) verifying the importance of tertiary contacts in the RNA structure prediction process, and (c) performing correlation analysis.

EA Setup. The EA parameters, such as population size, mutation rate, and elitism percentage that have been already tuned to study the internal near-native state dynamics of RNA were not changed because they do not affect the final outcome. However, since the folding of RNA is a different problem compared to its internal dynamics close to equilibrium, we have verified the MD simulation protocol on the RNA structures presented in Figure 2. Three factors in the MD protocol were tested: temperature, starting structure, and simulation length.

The tests of the simulation temperature used for folding (i.e., constant 310 or 100 K or gradual cooling from 310 to 0 K) showed that the difference in the final EA score was smaller

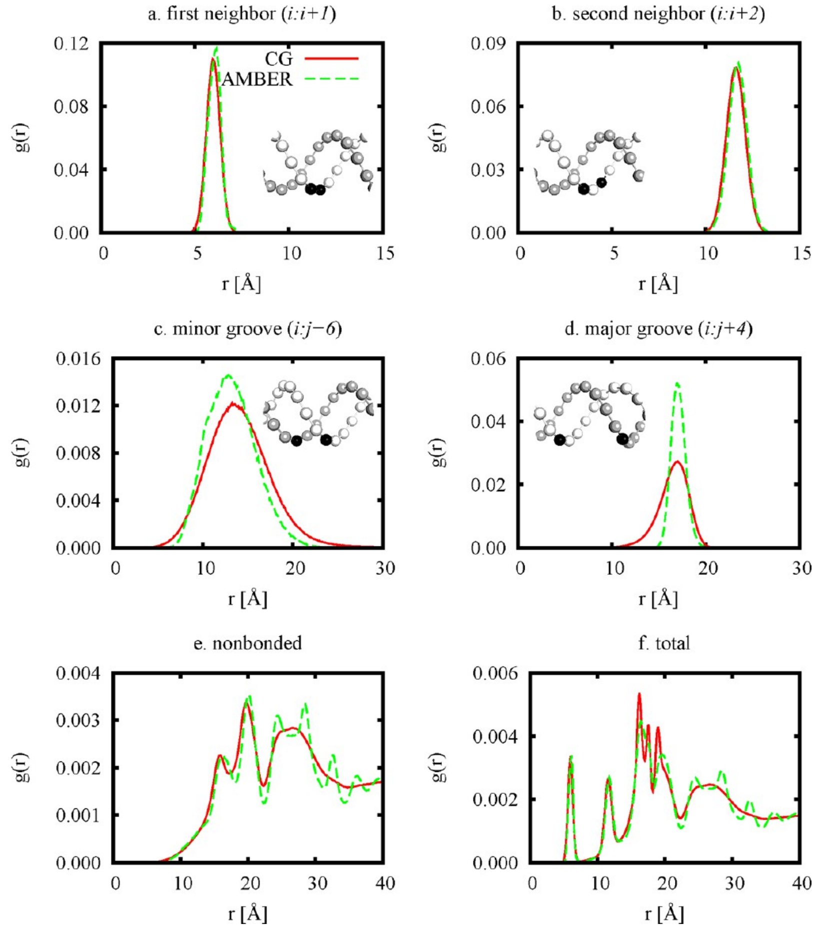


Figure 5. Comparison of distance distribution functions obtained from 100 ns MD simulation of a 35 base pair A-RNA helix (see Figure 2a) with the Amber all-atom FF (dashed line) and CG FF (solid line). Black beads represent a test pair of pseudoatoms used to calculate the corresponding distance distribution function.

Table 2. Pearson Correlation Coefficients between the CG FF Parameters in the Best 20% Models Tuned for the Scoring Function Applied in the Equilibrium Dynamics of the RNA Helix[a]

	r_0	k_r	θ_0	k_θ	α_0	k_α	$k^{i:j}$	$k^{i:j+1}$	$k^{i:j+2}$	$r_0^{i:j}$	$r_0^{i:j+1}$	$r_0^{i:j+2}$	ε_r
r_0		0.23						0.10					
k_r	0.23			−0.15					−0.10				
θ_0				−0.14						−0.19			0.19
k_θ		−0.15	−0.14				−0.10	−0.14			−0.10		
α_0												−0.13	
k_α			0.16				−0.31	−0.29	−0.17				
$k^{i:j}$				−0.10		−0.31		0.16					
$k^{i:j+1}$	0.10			−0.14		−0.29	0.16			0.13			
$k^{i:j+2}$		−0.10				−0.17				0.13	0.12		
$r_0^{i:j}$			−0.19					0.13	0.13		0.45		
$r_0^{i:j+1}$				−0.10					0.12	0.45		0.64	
$r_0^{i:j+2}$					−0.13						0.64		
ε_r			0.19										

[a] See Chart 1a and Figure 2. Parameters are those of eq 4−7 and eq 11. The values lower in magnitude than 0.1 are not shown for clarity. The coefficients are for a CG FF with the following potential energy terms: interstrand interactions were described with a harmonic potential for the canonical Watson−Crick pairs and the repulsive nonbonded interactions with a Coulomb function.

than the calculated error. Therefore, the subsequent MD simulations were performed at a constant temperature of 310 K. Next, the effect of the RNA starting structure was tested. Ten picasecond MD folding simulations that started from circular positioning of CG beads scored better than the grid placing method probably because the latter method had trouble with pseudoknots. However, for a 1 ns MD folding simulation the difference resulting from the starting structure became insignificant. The starting structure affects only the first frames of the simulation and this effect is reduced in the subsequent MD steps. So, in the consecutive MD runs, we have used the circular starting topology because it is easier to apply. Finally, we have found that the 3D structure prediction outcome depends on the length of the MD simulation. As shown in Table 3, the lowest RMSDs in 10 ps simulations are above 17 Å, but extending the folding simulations 100-times drops the RMSDs from the reference crystal structure to about 10 Å. Achieving RMSD of 10 Å is sufficient for such a simple one-bead RNA model[83] and within the limits of the CG model. The CG predictions can be further improved by remapping the CG structure to the full-atomistic one and running an all-atom MD simulation.[45] Figure 6 and Supporting Information Figure S3 show the RMSD changes during MD simulations of various RNAs shown in Figure 2. The corresponding trajectories are shown in Supporting Information Movies S2−S5. Typically, in the beginning of the simulation when a fold close to the reference one is found, a major drop in RMSD is observed followed by the fluctuations around the optimal position. Due to these oscillations the RMSD averaged over 10 frames is smaller than the RMSD of the final frame so more than one structure from the CG trajectory should be taken for further refinement.

3.2.2. Tertiary Contacts. Further, we have verified if including the tertiary contacts affects the prediction quality. Jonikas et al.,[83] who described an RNA one-bead CG model, stated that tertiary contacts are essential for proper structure prediction and took such contacts from RNA chemical probing experiments. To test their hypothesis in our case, we have used all noncanonical hydrogen bonds, found in the reference X-ray or NMR structure, as tertiary contacts. These noncanonical hydrogen bonds were identified by RNAView[65] and are reported in Supporting Information Table S7. Due to diverse nature of these contacts, their equilibrium distance r_0 was independently set as in the reference structure. The results are presented in Tables 3 and 4 and suggest that the effect of

Table 3. RMSD of the CG MD Folding Simulation of tRNA Showing the Dependence on the Learning/Training Set Used to Tune the FF, MD simulation time, and the Presence of Non-canonical Hydrogen Bonds in the Potential Energy Terms

		RMSD [Å] for tRNA[a]	
		tertiary contacts[b]	tertiary contacts
learning set	folding time [ps]	included	not included
tRNA	10	17.11 ± 0.01	18.80 ± 0.03
	1000	9.2 ± 00.5	10.8 ± 0.2
4 RNAs[c]	10	17.43 ± 0.03	20.5 ± 0.1
	1000	9.1 ± 2.6	12.0 ± 1.3

[a]The lowest RMSD from the reference tRNA structure among the 10 best-optimized models. [b]The noncanonical tertiary contacts that were included are listed in Supporting Information Table S7. [c]The four RNAs learning set is shown in Figure 2.

Table 4. RMSD for the Structure Prediction Tests in 1 ns MD Simulations with Non-canonical Hydrogen Bonds Turned on and off

	RMSD[b] [Å]			
tertiary contacts	tRNA (6TNA)[c]	hairpin loop (2KHY)	pseudoknot (2LC8)	ribozyme (3D0X)
included[a]	11.6 ± 1.0	8.2 ± 0.7	10.2 ± 0.7	15.8 ± 0.9
not included	13.6 ± 1.1	7.4 ± 0.8	10.6 ± 0.8	16.0 ± 1.4

[a]The noncanonical tertiary contacts that were included are listed in Supporting Information Table S7. [b]RMSDs are an average over 10 best-optimized models. [c]PDB codes are in parentheses, and the structures are shown in Figure 2.

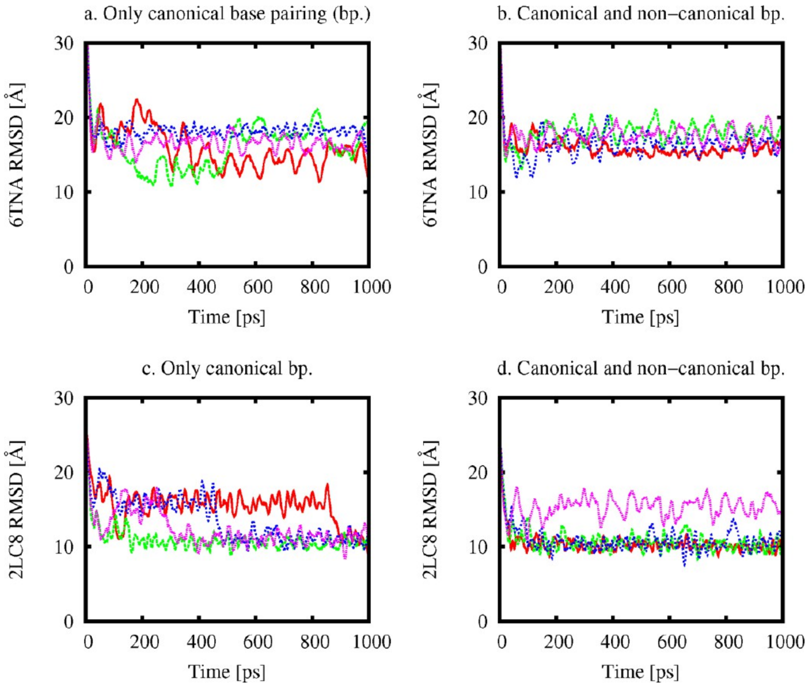


Figure 6. RMSD as a function of the CG MD simulation time for the structure prediction of tRNA (Figure 2b) and a pseudoknot (Figure 2d). The plots on the left present RMSD for a CG FF optimized without including the interactions for noncanonical hydrogen bonds (different line styles/colors represent different trajectories of the same CG FF) and plots on the right present RMSD for a CG FF with noncanonical hydrogen bonds included.

Table 5. Pearson Correlation Coefficients between CG FF Parameters in the Best 10% Models Tuned Using the Structure Prediction Scoring Function Defined for Four RNAs Training Set with Canonical Hydrogen Bonds Included[a]

	r_0	k_r	θ_0	k_θ	α_0	k_α	$k^{i:\,j}$	$k^{i:\,j+1}$	$k^{i:\,j+2}$	$r_0^{i:j}$	$r_0^{i:j\,+\,1}$	$r_0^{i:j\,+\,2}$	ε_r
r_0													0.13
k_r				0.18		−0.12	0.24					0.16	
θ_0											0.10		0.10
k_θ		0.18				0.14	0.11	0.13					
α_0													−0.10
k_α		−0.12		0.14				0.18					
$k^{i:\,j}$		0.24		0.11							−0.15	0.17	−0.11
$k^{i:\,j+1}$				0.13		0.18							
$k^{i:\,j+2}$													
$r_0^{i:\,j}$											0.10	−0.13	
$r_0^{i:\,j\,+\,1}$			0.10				−0.15			0.10			
$r_0^{i:\,j\,+\,2}$		0.16					0.17			−0.13			
ε_r	0.13		0.10		−0.10		−0.11						

[a] See Chart 1b and Figure 2. Parameters are those of eq 4−7 and eq 11. For clarity the values lower in magnitude than 0.1 are not shown. The following potential energy terms were used: interstrand interactions were described with a harmonic function for canonical Watson−Crick pairs and the nonbonded interactions were described with a repulsive shielded Coulomb potential.

including the tertiary contacts varies among molecule classes. For tRNA, which has a complicated tertiary fold, RMSD improves on average by 2 Å if the tertiary constraints are turned on. For a pseudoknot and ribozyme the noncanonical bonds do not affect the folding outcome and for a hairpin loop including these tertiary bonds results in a 0.8 Å rise in RMSD. Since these interactions have the largest impact on the structure prediction for tRNA, we have checked further the effect of limiting the training set to only tRNA. As shown in Table 3 for the "tRNA specialized" CG potential there is no difference in the prediction quality for both FF variants. For tRNA we have also tested the same constraints as were used by Jonikas et al.[83] (contacts involving nucleotides: 8 and 14, 15 and 48, 54 and 58, 10 and 45, according to numbering in the 6TNA.pdb structure) and in this case a tRNA-only potential resulted in a 8.1 Å best RMSD (the average over 10 best FFs was 8.6 ± 0.2).

3.2.3. Parameter Correlations. The last question (c) refers to parameter correlations. The Pearson correlation coefficients for the structure prediction cases are presented in Table 5. The highest absolute correlation coefficient was 0.24 between the pseudobond and *i:j* interstrand force constants. For structure prediction without noncanonical bonds, 27 pairs had a correlation coefficient higher in magnitude than 0.1 (see Table 5). The strongest correlation with the Pearson coefficient of 0.24 was observed between the $k_{i:j}$ base pairing force constant and the k_r pseudobond force constant. Second in magnitude, with a coefficient of 0.18, was the correlation between the k_θ pseudoangle force constant and the k_r pseudobond force constant. This suggests that a stronger pseudobond potential hinders adopting a proper tertiary structure and has to be balanced by increasing the force constants for the angle and base pairing terms. However, due to overall relatively low Pearson coefficients, such considerations have to be considered with caution.

3.3. Comparison of FF Parameters for Different Tasks.

Our results allow a comparison of the optimal FF parameters for different tasks, such as RNA structure prediction and equilibrium internal dynamics. We compare the FFs that differ only by numerical parameter values but have the same mathematical form (harmonic interstrand and repulsive nonbonded), include the same types of interactions (no tertiary or noncanonical base pairings) and have the same reasonable range of parameters. The optimal numerical values of the FF parameters are presented in Supporting Information Table S6 and the parameter value−score relationship is presented in Figure 7. The FF parameters show that the equilibrium positions for bonded and interstrand interactions are universal. However, the corresponding force constants differ significantly; the structure prediction favors higher force constants and stronger electrostatic repulsion. Such tendency can be also seen in Figure 7. Equilibrium dynamics is scored by, among others, the distance distributions and therefore too strong and too weak potentials are equally unfavorable. Structure prediction, on the other hand, requires a more restrictive potential to prevent misfolding in the beginning of the simulation. If the structure prediction simulation is started from a reference crystal structure, the pseudoangle force constant drops 7-fold and the repulsive interaction becomes twice as weaker (see Supporting Information Table S6)—reverting to parameters close to those for the equilibrium dynamics.

Linear correlation coefficients analysis shown in Tables 2 and 5 also suggests that parameters optimized for the equilibrium dynamics are more correlated than the ones for tertiary structure prediction. Therefore, the correlations between the parameters are more of an issue of proper dynamical behavior than structure stabilization.

The relationship between the performance of the CG model in an MD simulation and the FF parameters (shown in Figure 7) provides a systematically obtained proof that certain parameter values better suit a particular task (described by a particular fitness function) and helps to predict if a single model can be applied for multiple tasks. Even though this analysis can be treated only qualitatively, it gives an estimate of the relationship between FF parameters and FF performance.

4. CONCLUSIONS AND OUTLOOK

We presented an EA based method for automatic optimization of one-bead CG models for ribonucleic acids and parametrized the CF FFs for equilibrium dynamics and tertiary structure prediction of RNA. The automatic procedure is less time-consuming and troublesome and can be applied to test different bead mappings and FF parameters. The parameter correlation analysis and the relationship between the FF parameters and the performance of the CG MD simulation can be also derived.

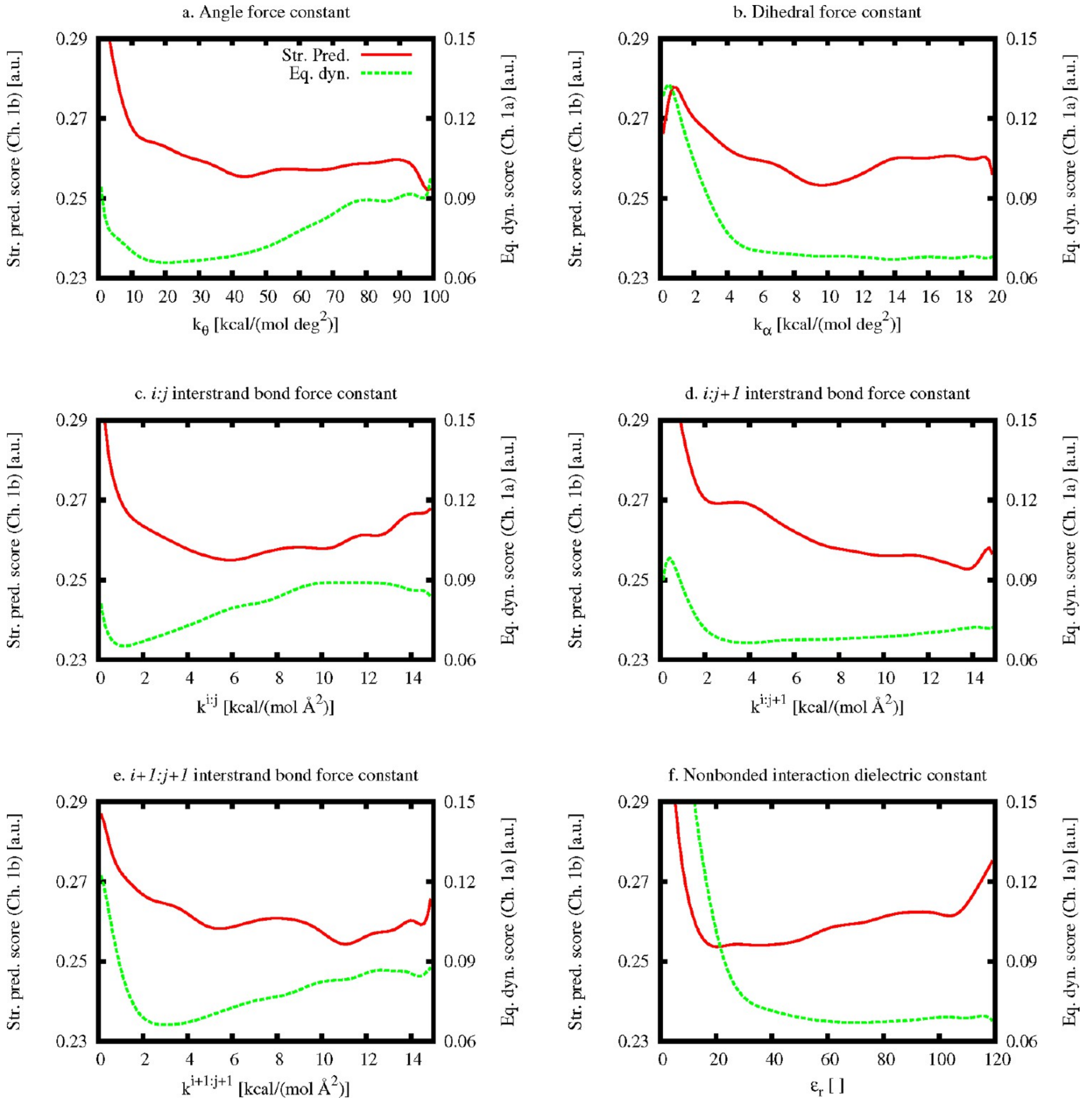


Figure 7. Best score (in arbitrary units) as a function of the parameter for equilibirum dynamics (Chart 1a) and structure prediction (Chart 1b) scoring functions. Graphs are presented for potentials with repulsive nonbonded, harmonic interstrand and without any additional non-Watson–Crick or tertiary interactions. The parameter range (x-axes) was divided into 64 equally distributed bins. The score is the best one achieved for a particular parameter (from the values in this bin range) and any value for all the other parameters. The values were calculated over 8192 FFs forming a single population. Graphs were smoothened by extrapolating with Bezier curves.

The quality of the obtained CG FFs is limited by numerous factors. The first is the fitness function. The EA algorithm finds only the FF that is optimal according to given criteria. If our assessment of the criteria is faulty then the algorithm might consider an unphysical solution as an optimal one, for example, leading to a potential energy function that is noncontinuous. This problem can be avoided by visual inspection of results, that is, trajectories and potential energy function plots. The fitness function can affect the results also in a more subtle way—by the choice of the weights. The fitness functions used were chosen in an arbitrary way to prevent the domination of a single scoring term in the optimization procedure. However, there is no strict mathematical or physical justification of this choice. This problem could be overcome at least to some extent by the Pareto method[48] in which different fitness function terms (corresponding to measures such as RMSD, RMSF, distance distributions) are not mixed into a total fitness function. Rather than finding a single optimal solution for the FF optimization process, a set of Pareto-optimal models is found. Further, a researcher (or a different computational method) have to choose an optimal FF from the Pareto-optimal set. This approach is successful for few objectives but if the size of the set is large, such as 25 distinct fitness function terms in the present MD case, the resulting Pareto-optimal set would be also large. The Pareto method has also another obstacle. Because of random noise affecting the fitness function there is no single optimal result for a single set of weights but rather a set of statistically indistinguishable solutions, increasing the population size even more. In the future we plan to include a hybrid approach—the user will be able to choose sets of fitness function terms that can be mixed using weights (e.g., multiple distribution scores or two different RMSDs). A comparison between these sets would be carried out using the Pareto front method.[48]

The presented EA-based methodology provides a systematic and automatic way to find optimal CG FF parameters. In the future not only the parameters of the potential energy functions but also the forms of the functions could be optimized in a single procedure. This approach, a genetic programming method[48] was already applied to find an optimal function that solves the Schroedinger equation[53] and to find interatomic potential energy function for a simple molecular system.[54] However, current computational resources are not sufficient to apply this procedure to a CG MD simulation because it is much more expensive than the parameter optimization. Further, we have used the EA framework with the MD method, but it can be easily transformed to Monte Carlo simulations, which are also used in the structure prediction tools. Also, this approach can be extended to more complex CG models, including more beads, sequence dependence, and interactions with other molecules such as proteins. With the increasing complexity, it might be hard to optimize all the parameters at the same time. However, for the optimization less important parameters can be fixed (e.g., to values taken from the BI method) or different subsets of parameters can be optimized during one run. With respect to the scoring function further studies are needed to include and assess the importance of the denaturation temperature on the FF parameters and to find an efficient way to assign the weights to the scoring terms.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

EA implementation details, FF parameters and ranges, MD FF optimization scoring function weights, and movies of the optimized MD FF (1 μs) and folding of four RNA molecules (100 ns). This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: f.leonarski@cent.uw.edu.pl.
*E-mail: joanna@cent.uw.edu.pl.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Piotr Krupiński for performing all-atomistic MD simulations of the RNA helix. F.L. and J.T. acknowledge support from ICM University of Warsaw (G31-4), National Science Center (DEC-2011/03/N/NZ2/02482 and DEC-2012/05/B/NZ1/00035), Foundation for Polish Science TEAM/2009-3/8 project, and HPC Infrastructure for Grand Challenges of Science and Engineering POWIEW project (the last two are cofinanced by European Regional Development Fund operated within Innovative Economy Operational Programme). F.T. and V.T. thank Paolo Mereghetti for useful discussions. The authors acknowledge use of PyMOL Molecular Graphics System, version 1.5.1 by Schrödinger, LLC, to create illustrations for this paper.

■ REFERENCES

(1) Al-Hashimi, H. M.; Walter, N. G. *Curr. Opin. Struct. Biol.* **2008**, *18*, 321−329.

(2) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide (Interdisciplinary Applied Mathematics)*, 2nd ed.; Springer: New York, 2010.

(3) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.

(4) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75−L77.

(5) Banáš, P.; Hollas, D.; Zgarbova, M.; Jurec, P.; Chetham, I.; Thomas, E. *J. Chem. Theory Comput.* **2010**, *6*, 3836−3849.

(6) Guvench, O.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1005−1014.

(7) Feig, M.; Brooks, I.; Charles, L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.

(8) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Phys.: Condens. Matter* **2004**, *16*, R481−R512.

(9) Koliński, A.; Skolnick, J. *Polymer* **2004**, *45*, 511−524.

(10) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.

(11) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.

(12) Seo, M.; Rauscher, S.; Pomès, R.; Tieleman, D. P. *J. Chem. Theory Comput.* **2012**, *8*, 1774−1785.

(13) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.

(14) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073−5083.

(15) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(16) Paliy, M.; Melnik, R.; Shapiro, B. *Phys. Biol.* **2010**, *7*, 036001.

(17) Müller-Plathe, F. *ChemPhysChem* **2002**, *3*, 754−769.

(18) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.

(19) Leonarski, F.; Trylska, J. In *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes. From Bioinformatics to Molecular Quantum Mechanics*; Liwo, A., Ed.; Springer Verlag: Berlin/Heidelberg, 2013; pp 109−149.

(20) Trylska, J. *J. Phys.: Condens. Matter* **2010**, *22*, 453101.

(21) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123−128.

(22) Tozzini, V.; Trylska, J.; Chang, C. E.; McCammon, J. A. *J. Struct. Biol.* **2007**, *157*, 606−615.

(23) Trovato, F.; Tozzini, V. *J. Phys. Chem. B* **2008**, *112*, 13197−13200.

(24) Laing, C.; Schlick, T. *J. Phys.: Condens. Matter* **2010**, *22*, 283101.

(25) Das, R.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14664−14669.

(26) Sharma, S.; Ding, F.; Dokholyan, N. V. *Bioinformatics* **2008**, *24*, 1951−1952.

(27) Jonikas, M. A.; Radmer, R. J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R. B. *RNA* **2009**, *15*, 189−199.

(28) Pasquali, S.; Derreumaux, P. *J. Phys. Chem. B* **2010**, *114*, 11957−11966.

(29) Denesyuk, N. A.; Thirumalai, D. *J. Phys. Chem. B* **2013**, *117*, 4901−4911.

(30) He, Z.; Chen, S.-J. *J. Phys. Chem. B* **2013**, *117*, 7221−7227.

(31) Xia, Z.; Bell, D. R.; Shi, Y.; Ren, P. *J. Phys. Chem. B* **2013**, *117*, 3135−3144.

(32) Trylska, J.; Tozzini, V.; McCammon, J. A. *Biophys. J.* **2005**, *89*, 1455−1463.

(33) Hyeon, C.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6789−6794.

(34) Tan, R. K. Z.; Petrov, A. S.; Harvey, S. C. *J. Chem. Theory Comput.* **2006**, *2*, 529−540.

(35) Cragnolini, T.; Derreumaux, P.; Pasquali, S. *J. Phys. Chem. B* **2013**, *117*, 8047−8060.

(36) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2005**, *25*, 1157−1174.

(37) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(38) Ercolessi, F.; Adams, J. B. *Europhys. Lett.* **1994**, *26*, 583−588.

(39) Bernauer, J.; Huang, X.; Sim, A. Y. L.; Levitt, M. *RNA* **2011**, *17*, 1066−1075.
(40) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624−1636.
(41) Reith, D. *Comput. Phys. Commun.* **2002**, *148*, 299−313.
(42) Hülsmann, M.; Köddermann, T.; Vrabec, J.; Reith, D. *Comput. Phys. Commun.* **2010**, *181*, 499−513.
(43) Sun, Q.; Faller, R. *Comput. Chem. Eng.* **2005**, *29*, 2380−2385.
(44) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. *J. Chem. Theory Comput.* **2009**, *5*, 3211−3223.
(45) Jonikas, M. A.; Radmer, R. J.; Altman, R. B. *Bioinformatics* **2009**, *25*, 3259−3266.
(46) Izvekov, S.; Violi, A.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 17019−17024.
(47) Savelyev, A.; Papoian, G. A. *Biophys. J.* **2009**, *96*, 4044−4052.
(48) Haupt, R. L.; Haupt, S. E. *Practical Genetic Algorithms*, 2nd ed.; Wiley-Interscience, 2004.
(49) González-Álvarez, D. L.; Vega-Rodríguez, M. A.; Gómez-Pulido, J. A.; Sánchez-Pérez, J. M. In *Evolutionary Computation, Machine Learning, and Data Mining in Bioinformatics*; Pizzuti, C., Ritchie, M. D., Giacobini, M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, 2011; Vol. *6623*, pp 89−100.
(50) Vassiliadis, V.; Thomaidis, N.; Dounias, G. In *Applications of Evolutionary Computation*; Chio, C., Brabazon, A., Caro, G., Drechsler, R., Farooq, M., Grahl, J., Greenfield, G., Prins, C., Romero, J., Squillero, G., Tarantino, E., Tettamanzi, A., Urquhart, N., Uyar, A., Eds.; Lect. Notes Comput. Sci.; Springer: Berlin/Heidelberg, 2011; Vol. *6625*, pp 131−140.
(51) Byrne, J.; Fenton, M.; Hemberg, E.; McDermott, J.; O'Neill, M.; Shotton, E.; Nally, C. In *Applications of Evolutionary Computation*; Chio, C., Brabazon, A., Caro, G., Drechsler, R., Farooq, M., Grahl, J., Greenfield, G., Prins, C., Romero, J., Squillero, G., Tarantino, E., Tettamanzi, A., Urquhart, N., Uyar, A., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, 2011; Vol. *6625*, pp 204−213.
(52) Quadflieg, J.; Preuss, M.; Rudolph, G. In *Driving Faster Than a Human Player Applications of Evolutionary Computation*; Di Chio, C., Cagnoni, S., Cotta, C., Ebner, M., Ekárt, A., Esparcia-Alcázar, A., Merelo, J., Neri, F., Preuss, M., Richter, H., Togelius, J., Yannakakis, G., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, 2011; Vol. *6624*; Chapter 15, pp 143−152.
(53) Makarov, D. E.; Metiu, H. *J. Med. Phys.* **1998**, *108*, 590−598.
(54) Brown, W. M.; Thompson, A. P.; Schultz, P. A. *J. Chem. Phys.* **2010**, *132*, 024108.
(55) Berman, H. M.; et al. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899−907.
(56) Sussman, J. L.; Holbrook, S. R.; Warrant, R. W.; Church, G. M.; Kim, S. H. *J. Mol. Biol.* **1978**, *123*, 607−630.
(57) Sarver, M.; Zirbel, C.; Stombaugh, J.; Mokdad, A.; Leontis, N. *J. Math. Biol.* **2008**, *56*, 215−252.
(58) Leontis, N. B.; Westhof, E. *RNA* **2001**, 499−512.
(59) Brion, P.; Westhof, E. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 113−137.
(60) Leontis, N. B.; Westhof, E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 300−308.
(61) Reiter, N. J.; Chan, C. W.; Mondragón, A. *Curr. Opin. Struct. Biol.* **2011**, *21*, 319−326.
(62) Galas, D. J.; Schmitz, A. *Nucleic Acids Res.* **1978**, *5*, 3157−3170.
(63) Tullius, T. D. *Nature* **1988**, *332*, 663−664.
(64) Merino, E. J.; Wilkinson, K. A.; Coughlan, J. L.; Weeks, K. M. *J. Am. Chem. Soc.* **2005**, *127*, 4223−4231.
(65) Yang, H.; Jossinet, F.; Leontis, N.; Chen, L.; Westbrook, J.; Berman, H.; Westhof, E. *Nucleic Acids Res.* **2003**, *31*, 3450−3460.
(66) Hodgkin, E. E.; Richards, W. G. *Quantum Biol. Symp.* **1987**, *14*, 105−110.
(67) Kabsch, W. *Acta Crystallogr., Sect. A* **1976**, *32*, 922−923.
(68) Yang, X.-S. *Int. J. Bio-Inspired Computation* **2011**, *3*, 77−84.
(69) Górecki, A.; Szypowski, M.; Długosz, M.; Trylska, J. *J. Comput. Chem.* **2009**, *30*, 2364−2373.
(70) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(71) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335−340.
(72) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817−3829.
(73) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(74) Panecka, J.; Mura, C.; Trylska, J. *J. Phys. Chem. B* **2011**, *115*, 532−546.
(75) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comput. Phys.* **1977**, *23*, 327−341.
(76) Darden, T.; Perera, L.; Li, L.; Pedersen, L. *Structure (Oxford, U.K.)* **1999**, *7*, R55−60.
(77) Wang, J.; Henkin, T. M.; Nikonowicz, E. P. *Nucleic Acids Res.* **2010**, *38*, 3388−3398.
(78) Houck-Loomis, B.; Durney, M. A.; Salguero, C.; Shankar, N.; Nagle, J. M.; Goff, S. P.; D'Souza, V. M. *Nature* **2011**, *480*, 561−564.
(79) Garst, A. D.; Héroux, A.; Rambo, R. P.; Batey, R. T. *J. Biol. Chem.* **2008**, *283*, 22347−22351.
(80) Tozzini, V. *Q. Rev. Biophys.* **2010**, *43*, 333−371.
(81) Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714−2726.
(82) Trovato, F. Modello minimalista per Acidi Nucleici: studio delle proprietà e delle transizioni strutturali del DNA tramite dinamica molecolare. M.Sc. thesis, Università di Pisa, Italy, 2007.
(83) Jonikas, M. A.; Radmer, R. J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R. B. *RNA* **2009**, *15*, 189−199.