Fitting corrections to an RNA force field using experimental data

Andrea Cesari,[†] Sandro Bottaro,[‡] Kresten Lindorff-Larsen,[‡] Pavel Banáš,[¶] Jiří Šponer,[§,¶] and Giovanni Bussi[*,†]

† *Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, 34136 Trieste, Italy*

‡ *Structural Biology and NMR Laboratory and Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark*

¶ *Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacký University, tř. 17 listopadu 12, 771 46, Olomouc, Czech Republic*

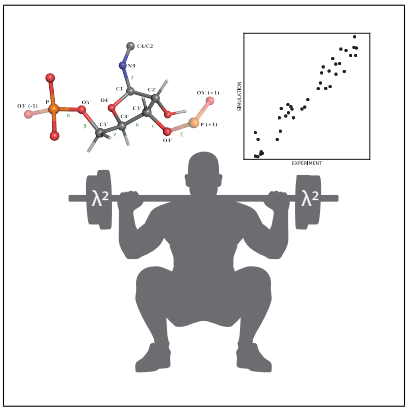
§ *Institute of Biophysics of the Czech Academy of Sciences, Kralovopolska 135, Brno 612 65, Czech Republic*

E-mail: bussi@sissa.it

Abstract

Empirical force fields for biomolecular systems are usually derived from quantum chemistry calculations and validated against experimental data. We here show how it is possible to refine the full dihedral-angle potential of the Amber RNA force field by using solution NMR data as well as stability of known structural motifs. The procedure can be used to mix multiple systems and heterogenous experimental information, and crucially depends on a regularization term chosen with a cross-validation procedure. By fitting corrections to the dihedral angles on the order of less than 1kJ/mol per angle, it is possible to increase the stability of difficult-to-fold RNA tetraloops by more than one order of magnitude.

Graphical TOC Entry



Molecular simulations using empirical force fields allow the characterization of the structural dynamics of RNA systems at atomistic detail, thus complementing and aiding the interpretation of experimental findings.[1,2] In combination with enhanced sampling methods,[3] they enable the study of conformational transitions ranging from the formation of base pairing and stacking to binding of ions and other ligands. However, their predictive capability is limited by the accuracy of the employed force fields. There is growing evidence that recent RNA force fields are not capable to model conformational dynamics in agreement with solution experiments on flexible oligonucleotides[4,5] or to predict the native structure of short hairpins.[6,7] It is thus becoming common practice to take advantage of experimental data in order to enforce agreement with experiment.[8–12] However, the quality of the underlying force field has still a significant impact on the reliability of the results. Over the last ten years, a number of attempts have been made to improve the accuracy of the Amber RNA force field.[13–23] Most of these studies used quantum chemistry-based calculations in order to parametrize dihedral terms.[13–17,21,22] A recent large re-parametrization of multiple charges, Lennard-Jones, and dihedral terms was shown to be capable to correctly fold some RNA hairpins[22] and made the proposed force field compatible with a 4-points water model.[24] The main difficulty of simultaneously changing a large number of parameters, however, is that the potential side effects are difficult to predict, and a validation on a large dataset might be required. Some of these side effects were already identified in Ref.,[23] where a more conservative approach was taken correcting specific hydrogen bonds.

Dihedral angle parameters in empirical potentials are fitted against quantum-mechanical (QM) calculations performed on very small model systems, up to dozens of atoms. Notably, dihedral reparametrizations were usually motivated by inaccuracies observed when performing simulations on larger systems and comparing them against experimental data. For example, the QM-based corrections bsc0[13] and χ_{OL3}[15] were proposed to avoid spurious transitions in the DNA backbone substates and in the RNA helix geometry, respectively. It is thus intriguing to evaluate the possibility to directly use failures in reproducing experimental

data in order to derive the force field terms, rather than choosing the dihedrals to be refined based on empirical observations and relying on QM calculations for their parametrization. Procedures have been introduced in the past in order to iteratively refine a force field so as to match some experimental observation.[25–29] A particularly critical issue in the usage of experimental data to derive force field parameters is the necessity to take into account the error in the experimental data, the error in the *forward models* used to back-calculate experiments from simulations, and, more generally, the need to avoid overfitting on specific datasets.

We here propose a procedure to derive force-field corrections by incorporating an arbitrary number of heterogenous experimental data measured for an arbitrary number of systems. The procedure is based on a likelihood maximization scheme where individual experiments can be assigned arbitrary weights. A regularization term is introduced to avoid overfitting. Particular care is dedicated to the choice of this term using a cross-validation procedure where datasets are iteratively excluded from the fitting procedure. We provide a practical example by refining all the torsions in the Amber RNA force field so as to improve the agreement with NMR data and with the observed stability of tetraloops. Multiple systems and different experimental types are included in order to improve the transferability of the corrections. Our results show that small corrections to dihedral angle parameters on the order of less than 1 kJ/mol per torsion can increase the stability of the native structure of a hairpin loop by orders of magnitude without significant side effects on the tested systems.

The systems considered in our refinement procedure were four RNA tetranucleotides (AAAA, CCCC, UUUU and GACC) and two RNA tetraloops (ccGAGAgg and ccUUCGgg). The tetranucleotide simulation data were obtained using parallel tempering[30] simulations and were taken from Ref.[11] Simulations of tetraloops were performed using the same protocol as in Ref.,[7] involving a combination of metadynamics[31–33] applied on the eRMSD from native structure[34] and parallel tempering.[30,35] Following Ref.,[7] eRMSD was computed using a cutoff larger than the standard value defined in Ref.[34] in order to increase its capability to accelerate

folding events. The final bias was used in order to compute weights for metadynamics simulations.[36] All systems were simulated with GROMACS,[37] using the ff99bsc0 + χ_{OL3} Amber force field[13,15,38] with corrections to van der Waals oxygen radii[39] and using the OPC water model.[40] The ff99bsc0 + χ_{OL3} force field was chosen as a starting point since, despite many further attempts, it still remains overall the most reliable force-field version for general simulations of diverse RNAs.[1,23] In combination with the OPC water model and modified van der Waals oxygen radii, it was shown to moderate the population of intercalated structures in RNA tetranucleotides.[11,18] We will simply refer to this force field as Amber. The parameters are available in GROMACS format at `https://github.com/srnas/ff`. Data for tetranucleotides involve both NOE and scalar couplings NMR measurements.[4,11,41] For tetraloops we require the native state to be the most populated one. Native conformations were arbitrarily chosen as those with `eRMSD`<0.8 from the X-ray reference structure.[34,42,43] Whereas we are not aware of direct measurements of the thermodynamic stability of the two investigated hairpin loops, experimental data for other hairpins of similar length suggest that the folded structure should have a non-negligible population (see e.g.[44]).

We consider a system described by a potential energy function V_0, to which corresponds a Boltzmann probability $P_0(\boldsymbol{x}) \propto \exp\left(-\beta V_0(\boldsymbol{x})\right)$, where $\beta = \frac{1}{k_B T}$, T is the temperature and k_B is the Boltzmann constant. Our aim is to construct a refined probability distribution of the form $P(\boldsymbol{x}, \{\boldsymbol{\lambda}\}) \propto P_0(\boldsymbol{x}) \exp\left(-\beta \sum_i^N f_i(\boldsymbol{x})\lambda_i\right)$. The correcting potential is thus expanded on a set of N basis functions, that might be for instance dihedral angles or non-bonded interaction terms. A factor λ_i is associated to each of the N basis functions. These factors must be found in order to simultaneously reproduce M experimental observables. The correcting potential was thus chosen with the following form:

$$V_{corr} = \sum_{t \in \{torsions\}} \sum_{i=1}^{N_t} \sum_{n=1}^{3} \left(\lambda_{1tn} \cos\left(n\phi_{ti}\right) + \lambda_{2tn} \sin\left(n\phi_{ti}\right)\right), \tag{1}$$

where $torsions = \{\alpha,\ \beta,\ \gamma,\ \delta,\ \epsilon,\ \zeta,\ \chi_{Pur},\ \chi_{Pyr}\}$ is the set of torsion types subject to the

correcting potential, N_t is the number of nucleotides involved in the refinement, $\lambda_{1tn}(\lambda_{2tn})$ is the weight associated to the cosine (sine) with multiplicity n relative to the torsion type t and ϕ_{ti} is the torsion of type t in the nucleotide i.

It is important to notice the difference with respect to maximum-entropy-based methods,[45,46] in which the basis functions are by construction identical to the forward-models used to backcalculate the experimental observables, and thus the number N of parameters is equal to the number of enforced experiments. In our approach the number of experimental observables M is in general different from N, usually being much larger ($M \gg N$). In order to find the optimal weights λ_i, we define an error function E encoding the overall discrepancy between observable averages in the refined ensemble and the related experimental values. The error function is built such that $E = 0$ if all observables are exactly reproduced. Given a set of M experimental observables denoted by $O_j (j = 1, \ldots, M)$, it is possible to enforce both equalities (i.e. $\langle O_j \rangle = O_j^{exp}$) or inequalities (i.e. $\left(\langle O_j \rangle < O_j^{exp}\right)$ and/or $\left(\langle O_j \rangle > O_j^{exp}\right)$). The averages are meant to be taken in the refined probability distribution $P(\boldsymbol{x}, \{\boldsymbol{\lambda}\})$. In this work we will compute such averages by reweighting the unrefined ensemble. The accuracy of the procedure will then depend on how close the refined ensemble is to the unrefined one. The error function E, which depends on the observables averages, will indirectly depend on $\boldsymbol{\lambda}$. Similarly to Ref.,[28] we introduce a *regularized* error function $\tilde{E}$ defined as:

$$\tilde{E}\left(\langle O_1 \rangle (\boldsymbol{\lambda}), \ldots, \langle O_M \rangle (\boldsymbol{\lambda})\right) + \alpha |\boldsymbol{\lambda}|^2 \tag{2}$$

which must be minimized in order to enforce the M ensemble averages. We will denote with $\boldsymbol{\lambda}^\star$ the set of parameters which minimize Eq. 2. Forward models and the applied restraints for all systems are summarized in Tab. S1. The second term in Eq. 2 is a l^2 regularization term needed to avoid over-fitting. The strength of the regularization can be tuned with the parameter α, choosing a value in the range $0 \leq \alpha < \infty$. Setting $\alpha = 0$ will maximally fit the data at the cost of a very large correcting potential. This will lead to a new ensemble

which will be potentially very different from the unrefined one, generating poor reweighting performance, and will have a large chance to be overfit on the utilized datapoints. In the opposite case ($\alpha \to \infty$) the data will not be fitted. In order to minimize $\tilde{E}(\boldsymbol{\lambda})$ we compute its gradients

$$\begin{aligned} \frac{\partial \tilde{E}}{\partial \lambda_j} &= \sum_{i=1}^{M} \frac{\partial E}{\partial \langle O_i \rangle} \frac{\partial \langle O_i \rangle}{\partial \lambda_j} + 2\alpha\lambda_j = \\ &= \sum_{i=1}^{M} \frac{\partial E}{\partial \langle O_i \rangle} \left(\langle f_j \rangle \langle O_i \rangle - \langle f_j O_i \rangle \right) + 2\alpha\lambda_j \end{aligned} \tag{3}$$

where $j = 1, \ldots, N$. The term $\frac{\partial E}{\partial \langle O_i \rangle}$ depends on the specific functional form used to combine the errors. We then minimize $\tilde{E}$ using the limited memory version of the Broyden-–Fletcher—Goldfarb—Shanno algorithm (L-BFGS). The employed code is available at `https://github.com/bussilab/ff-fitting-tools`.

Once the optimal $\boldsymbol{\lambda}^\star$ are found, the final estimation of the observable averages can be found by reweigthing the unrefined ensemble:

$$\langle O_i \rangle = \frac{\sum_{t=1}^{N_{frames}} O_i(t)\, e^{\sum_{d=1}^{N} f_d(t)\lambda_d}}{\sum_{t=1}^{N_{frames}} e^{\sum_{d=1}^{N} f_d(t)\lambda_d}} \tag{4}$$

where t denotes the t^{th} frame of the unrefined ensemble. As we anticipated the efficiency of the reweighting procedure is inversely related to the distance between the refined and unrefined ensembles. Such distance can be kept relatively small by tuning the regularization parameter α in Eq. 2. The optimal value of α, to which we will refer as $\alpha^\star$, can be found via cross validation strategies. In this work we use the k-fold cross validation method, where the data set is split in k blocks. For each trial value of α, k minimizations are performed. During the i^{th} minimization, with $i = 1, \ldots, k$, the i^{th} block is left out as validation set while the remaining $k-1$ blocks are used as training set. After the optimal $\lambda^\star$ are found, the un-regularized error function E is evaluated on the validation set. At the end of the k^{th} minimization, a final cross validation error E_{cv}, for the given α, is computed as the average

of the validation errors on each of the k blocks. The optimal value of α will then be the one minimizing the cross validation error E_{cv}. The rationale behind this procedure is that in this way we will choose the more conservative value of α which will generalize better than other values of α to data that were not seen in the training set. In the present context, the optimal α is thus expected to result in force field corrections that will be better transferable to systems and experiments not considered in the fitting procedure. The regularization parameter was estimated using a k-fold cross validation procedure with $k = 3$. More precisely, we divided the training set in 3 blocks, each of which containing data from a different type of experiment. The first block contained all the ^3J scalar couplings, the second block contained all the NOE data, and the third block contained the stability of the tetraloops (see Table S1). The cross-validation error function evaluated for different values of α shows a minimum at $\alpha^\star \approx 1500$ (Fig. 1). This means that, for the given dataset, 1500 represent the optimal value of α giving the best balance between overfitting and predictiveness on different data and, presumably, systems not seen in the training set.

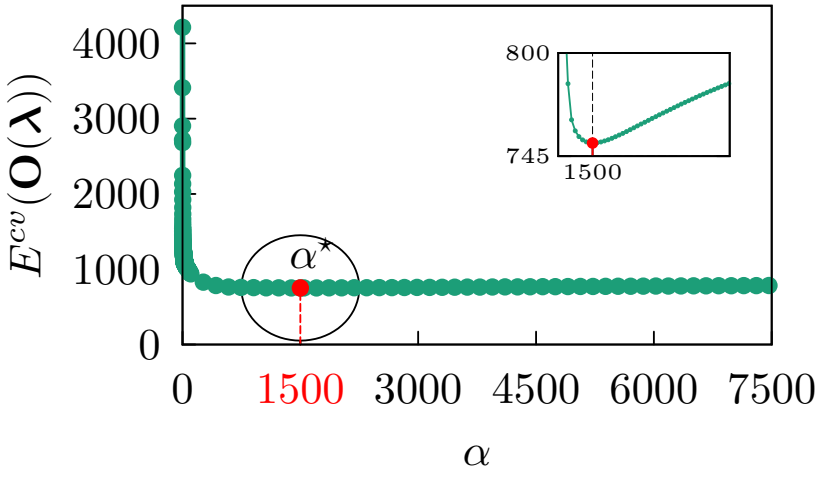


Figure 1: Cross-validation results. The unregularized error functions is evaluated on the validation set for parameters obtained using different trial values of α.

Notice that since we are enforcing a large number of data on multiple systems and using a limited number of parameters mutual compatibility is not *a priori* guaranteed. We thus checked the restraints enforced during the training phase. We will call $\texttt{Amber}_{\texttt{RW}}$ the ensemble

obtained by reweighting the reference `Amber` simulation using the optimal weights obtained by re-fitting all the training set (without leaving out any data) using the optimal value of $\alpha^\star = 1500$. In the case of tetranucleotides we computed the value of the enforced ^3J scalar couplings and `NOE` distances. We then validated the estimated corrections by computing the RMSE and the percentage of violations for ^3J couplings and `NOE` respectively (Fig. 2). ^3J couplings computed in the reweighted ensemble better reproduce experimental data for all the considered tetranucleotides, although in some cases the improvement is limited. All the resulting RMSEs are compatible with the expected error for the forward model used to compute scalar couplings. The percentage of violated `NOE` is decreased by the corrections for all systems except `GACC` tetranucleotide. The improvement is particularly visible for `CCCC`. The decrease in violated NOEs is paralleled by an increase in the population of A-form structures (Fig. S1). We notice that a similar increase has been observed in Ref.[11] as well. However, in this case the population shift is a consequence of small corrections on all dihedral angles, whereas in Ref.[11] the structures were explicitly weighted based on their agreement with NMR data.

In order to assess the importance of introducing a regularization term, it is instructive to consider the results of a reweighting performed setting $\alpha = 0$ (i.e., without any regularization term). Results obtained fitting all the datapoints are reported in Fig. S2 (to be compared with Fig. 2). Here it can be appreciated that the agreement with experiments is improved for all the ^3J scalar couplings. However, this is obtained at the price of overfitting the data. Indeed, the fraction of NOE violations instead is even higher than that obtained with $\alpha = 1500$. Overfitting can also be systematically assessed by considering the cross-validation datasets (Fig. S3). On the contrary, when using $\alpha = 1500$, agreement with experiment for individual datapoints is largely independent of which datapoints are discarded in the training phase (Fig. S4).

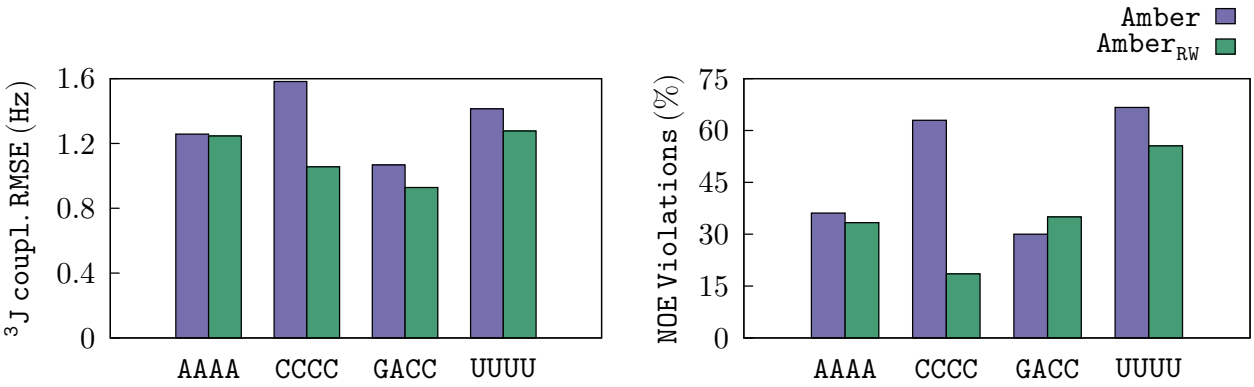


Figure 2: ^3J scalar coupling RMSE and NOE violations for RNA tetranucleotides. For both RMSE and percentage of violations, the lower the better.

As regards the two tetraloops, we computed the folded fraction and the free energy as function of the `eRMSD` from native structure, both with the unrefined `Amber` force field and with the reweighted ensemble $\texttt{Amber}_{\texttt{RW}}$. Results are reported in Fig. 3. The fraction of folded structures is significantly increased for both systems, indicating that the introduced correction reduces the relative weight of some of the unfolded structures observed in the `Amber` ensemble. The effect of the regularization term can be appreciated in Fig. S5, where populations obtained without regularization are reported. Strikingly, when the tetraloop stabilities are not included in the training set, their value becomes much lower than that obtained with the intial `Amber` force field. On the other hand when choosing $\alpha = 1500$ the stabilities of the tetraloops are moderately increased with respect to the intial `Amber` force field even when the tetraloop stabilities are not included in the training set.

All the results shown so far were obtained performing a reweighting of a given simulated ensemble. The statistical accuracy of the reweighting procedure depends however on the distance between the unrefined ensemble and the reweighted one. In case the two ensembles are too different, reweighting might be inefficient since there may be very few frames in the original ensemble with a significant weight. A rough estimate of the reweighting accuracy is given by the Kish's effective sample size $n_{eff} = \frac{\left(\sum_{i=1}^{N_{frames}} w_i\right)^2}{\sum_{i=1}^{N_{frames}} w_i^2}$, that satisfies $1 \leq n_{eff} \leq N_{frames}$ and indicates how many frames, among all the available ones, are effectively used in the reweighting procedure. The value of the Kish's effective sample size can be controlled by

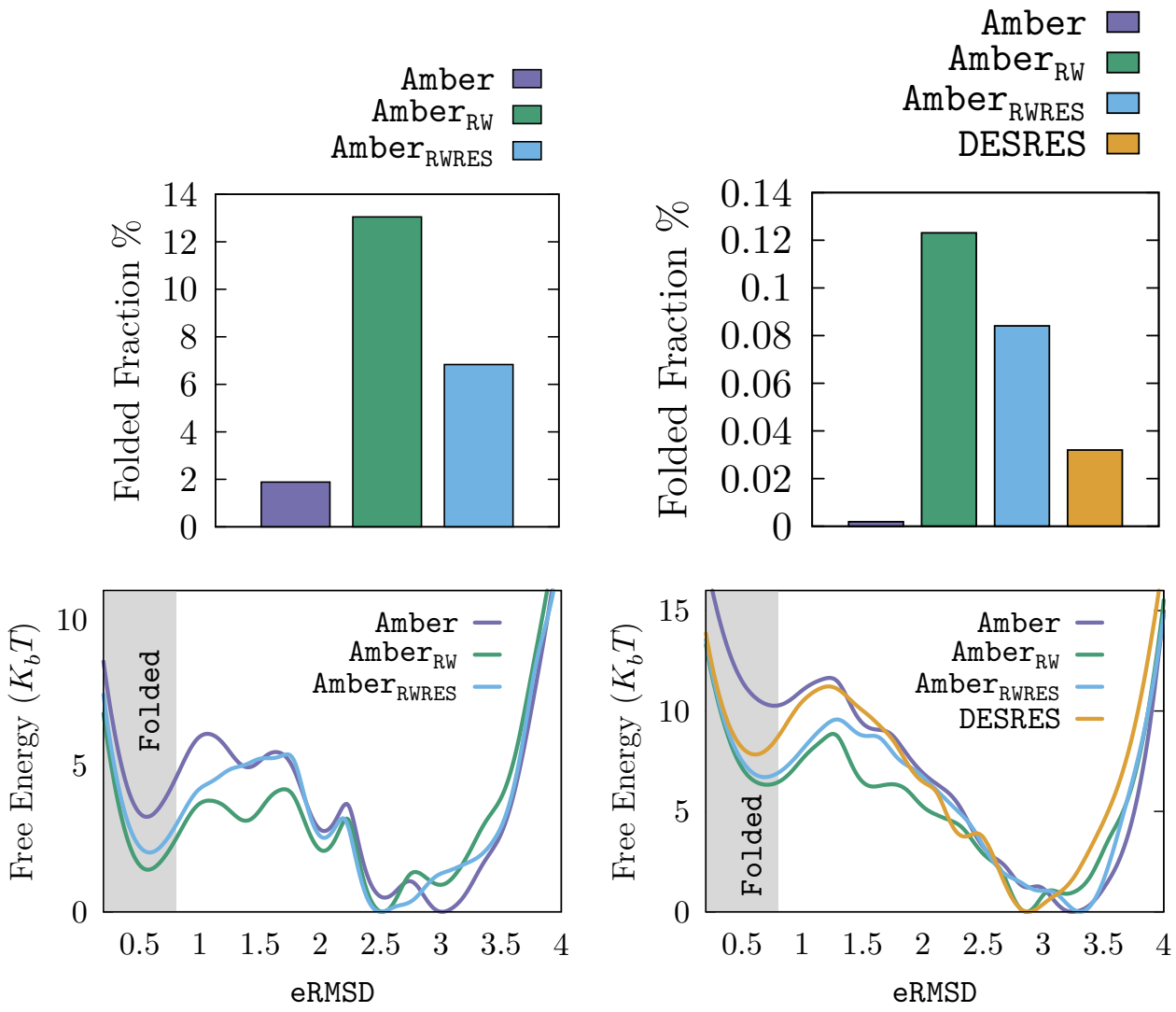


Figure 3: RNA `ccGAGAgg` (left) and `ccUUCGgg`(right) tetraloops. Fraction of folded structures (top) and free energy surface (bottom) with different force-field parametrizations. Unrefined Amber force-field (`Amber`) in purple, reweighted Amber force-field (`Amber`$_{\texttt{RW}}$) before resampling in green, refined Amber force-field (`Amber`$_{\texttt{RWRES}}$) after resampling in light blue, D. E. Shaw[22] (`DESRES`) force field in orange. `eRMSD` was computed here using the same distance cutoff used for enhanced sampling simulations. The threshold used to define the folded state is shaded. Whereas for GAGA it correctly identifies the native structure, for UUCG it is too strict and discards a fraction of the folded states. Numerical values for the populations, as well as alternative results obtained analyzing the simulations with `eRMSD` using the standard distance cutoff and a different threshold, are reported in Supporting Information (Table S2).

adjusting the regularization parameter α. Although it is difficult to set a general criterion for the acceptable values of n_{eff}, the $\alpha^{\star}$ chosen here guarantees at least 30% of effective samples, similar to previous iterative approaches.[25,26] A detailed analysis of reweighting performance, together with a critical comparison between reweighting and restraining methods, can be found in Ref.[47]

Squared fluctuations of the correcting potential scale proportionally to the number of nucleotides, and tetranucleotides are small enough for reweighting to be efficient. Indeed, in a recent work it was possible to directly reproduce experimental data by using a maximum-entropy-based reweighting[11,45,46] based on the same simulations. We therefore performed a resampling for the tetraloops only, that represent the most challenging case, by repeating simulations starting from the same initial conditions, with identical enhanced sampling schemes, but including the corrections corresponding to the optimized parameters (Table S3) directly in the potential energy function used to generate the trajectories. Results are reported in Fig. 3 and Fig. 4. In the top panels in Fig. 3 results obtained by reweighting ($\texttt{Amber}_{\texttt{RW}}$) and after resampling ($\texttt{Amber}_{\texttt{RWRES}}$) are compared. We notice that, although $\texttt{Amber}_{\texttt{RWRES}}$ increases the stability of both tetraloops, the improvement is not as large as the one obtained by reweighting only. This is a consequence of both the inaccuracy of the reweighting due to poor sampling of the reference ensemble and a small unavoidable overfitting on the specific conformations present in the original `Amber` ensembles. In any case, thanks to the regularization term, the correction is limited, the Kish's effective sample size is relatively large, and the trends observed in the reweighting calculations are the same of those observed in the resampling calculation. At an early stage we tried the same procedure without including any regularization term and, whereas the reweighting procedure was reporting a high stability for the tetraloops, the stabilities obtained at the resampling stage were significantly lower than those obtained during reweighting, and also lower than those obtained with the optimal regularization parameter, indicating that the derived parameters were highly overfitted on the conformations sampled in the specific run. The obtained weights were also validated

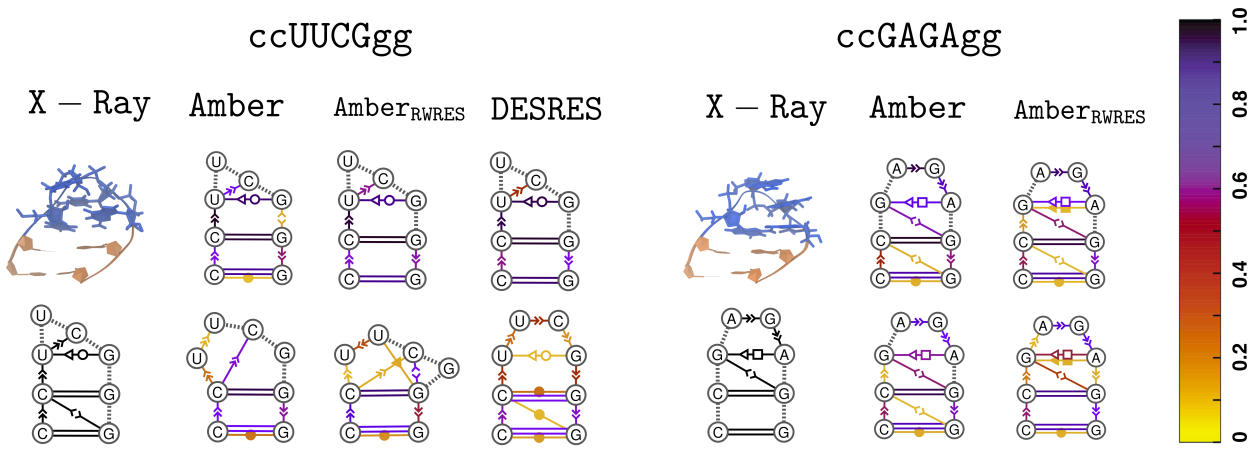


Figure 4: First and fifth column show, respectively, ccUUCGgg and ccGAGAgg tetraloops X-ray structures. Both three-dimensional and secondary structures are shown. Secondary structure were obtained with the BARNABA software.[48] For all the other columns, we show dynamic secondary structure representations obtained with different force fields as indicated in the column name. The color scheme shows the fraction of frames for which the interaction is formed. Structures were randomly sampled from the simulations with probabilities chosen in order to remove the effect of the metadynamics bias. In the first row, we report ensembles where the eRMSD of the whole system is within 0.8 from native. This ensemble was selected with the same criterion used to identify the native conformations in the force-field fitting procedure. In the second row instead we report ensembles where only the stem has eRMSD from native less than 0.8. In this second case the loop portion is then free to assume any conformation allowed by the employed force field.

using a WHAM procedure (see SI).

Figure 4 reports the dynamic secondary structure[48] for selected ensembles for both tetraloops as obtained using all the employed force fields, together with their native structures. When the ensemble is selected to only contain conformations where both the stem and the loop are formed, the dynamic secondary structure is highly homogeneous and, by construction, consistent with native. Conversely, the ensembles where only the stem is selected to be formed report on the capability of the employed force fields to reproduce the native loop structure assuming the stem to be formed. For the GAGA tetraloop, the secondary structure is consistent with the native structure both using the Amber and the Amber$_{RESRW}$ force

fields. This indicates that already in the original force field the loop would have the correct structure if the stem is folded. Our correction does not perturb significantly the result. For the UUCG tetraloop, both the original Amber and the $\text{Amber}_{\text{RESRW}}$ force fields are not capable to reproduce the contacts present in the crystal structure of the loop, indicating that further corrections would be required to this aim. It is also possible that the reported conformations might be present in the experimental ensemble although with a low population.[49] For this system we also report results obtained using the DESRES force field,[22,24] simulated using an identical protocol. We notice that the native structure of this tetraloop is more stable using $\text{Amber}_{\text{RESRW}}$ compared to the DESRES force-field (Fig. 3). When conformations where the stem is formed are selected, the loop displays a partly correct native structure where the trans-sugar/Watson-Crick pair (U3-G6) is detected, although with a low population. On the other hand the parallel stacking U3-C5, that is reported both in crystal[42] and solution[50] structures, is not observed. A comparison between the results of the DESRES force field with the force field corrections derived here is however difficult since the DESRES force field modified most of the nonbonded interactions, whereas our correction only impact the torsional angles.

In conclusion, we introduce a method based on an existing procedure[25] to develop force-field corrections using experimental data. An important extension presented here is the introduction of a regularization procedure based on cross-validation that controls overfitting. The method is used to combine data of different types on multiple systems, which is crucial in order to achieve transferable parameters. Since the optimization of the parameters is done with a reweighting procedure, its application in a single iteration as shown here is limited to small corrections, such as dihedral terms or other solute-solute non-bonded terms. Fitting parameters that lead to larger changes in the ensemble might require the procedure to be applied in an iterative manner[25] by resampling new conformations at every change of the correction parameters. The method is very flexible in that arbitrary error functions can be optimized. In this specific case, we used NMR data and assumed populations of native struc-

tures. Other possible choices for nucleic acid systems could be helical parameters or other structural quantities for which ranges of acceptable values can be identified *a priori*. For the investigated systems, we have shown that very small corrections to dihedral angles can affect significantly the population of the native structure in RNA tetraloops. By only correcting dihedrals we were not able to obtain a force field capable to fold the investigated loops to the native structure with a significant population. However, the resulting populations were improved with respect to the original ones and, for the UUCG tetraloop, higher than those obtained with a recently proposed reparametrization.[22] As a word caution, before suggesting the derived corrections to be used on new systems they should be validated on a larger set of RNA motifs including more non-canonical interactions. The developed parameters are available for testing (see Table S3 and `https://github.com/bussilab/ff-fitting-tools`). Better results might be obtained if starting from a more accurate force field. Since torsional potentials are usually fitted as the last step in force field derivation, we suggest that a final refinement could be performed with the procedure introduced here on top of any a priori available parametrization, including in the minimized error function all the desired structural features.

Supporting Information Available

Detailed description of the experimental data used to define the error function (Table S1). Supplementary results for tetraloops showing the population of native structures using various metrics (Table S2). Population of A-form structures in tetranucleotide simulations (Fig. S1). Effect of regularization term on tetranucleotide (Fig. S2, S3, and S4) and tetraloop (Fig. S5) simulations. Coefficients of the corrections (Table S3). Validation of corrections using WHAM (Fig. S6). This information is available free of charge via the Internet at http://pubs.acs.org

References

(1) Sponer, J.; Bussi, G.; Krepl, M.; Banáš, P.; Bottaro, S.; Cunha, R. A.; Gil-Ley, A.; Pinamonti, G.; Poblete, S.; Jurečka, P. et al. RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem. Rev.* **2018**, *118*, 4177–4338.

(2) Dans, P. D.; Gallego, D.; Balaceanu, A.; Darré, L.; Gómez, H.; Orozco, M. Modeling, Simulations, and Bioinformatics at the Service of RNA Structure. *Chem* **2019**, *5*, 51–73.

(3) Mlỳnskỳ, V.; Bussi, G. Exploring RNA structure and dynamics through enhanced sampling simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 63–71.

(4) Condon, D. E.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Yildirim, I.; Turner, D. H. Stacking in RNA: NMR of Four Tetramers Benchmark Molecular Dynamics. *J. Chem. Theory Comput.* **2015**, *11*, 2729–2742.

(5) Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Cheatham, T. E. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **2015**, *21*, 1578–1590.

(6) Kuhrová, P.; Best, R. B.; Bottaro, S.; Bussi, G.; Šponer, J.; Otyepka, M.; Banáš, P. Computer Folding of RNA Tetraloops: Identification of Key Force Field Deficiencies. *J. Chem. Theory Comput.* **2016**, *12*, 4534–4548.

(7) Bottaro, S.; Banáš, P.; Šponer, J.; Bussi, G. Free Energy Landscape of GAGA and UUCG RNA Tetraloops. *J. Phys. Chem. Lett.* **2016**, *7*, 4032–4038.

(8) Borkar, A. N.; Bardaro, M. F.; Camilloni, C.; Aprile, F. A.; Varani, G.; Vendruscolo, M. Structure of a low-population binding intermediate in protein-RNA recognition. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7171–7176.

(9) Krepl, M.; Blatter, M.; Cléry, A.; Damberger, F. F.; Allain, F. H.; Sponer, J. Structural

study of the Fox-1 RRM protein hydration reveals a role for key water molecules in RRM-RNA recognition. *Nucleic Acids Res.* **2017**, *45*, 8046–8063.

(10) Podbevšek, P.; Fasolo, F.; Bon, C.; Cimatti, L.; Reißer, S.; Carninci, P.; Bussi, G.; Zucchelli, S.; Plavec, J.; Gustincich, S. Structural determinants of the SINE B2 element embedded in the long non-coding RNA activator of translation AS Uchl1. *Sci. Rep.* **2018**, *8*, 3189.

(11) Bottaro, S.; Bussi, G.; Kennedy, S. D.; Turner, D. H.; Lindorff-Larsen, K. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Science Advances* **2018**, *4*, eaar8521.

(12) Kooshapur, H.; Choudhury, N. R.; Simon, B.; Mühlbauer, M.; Jussupow, A.; Fernandez, N.; Jones, A. N.; Dallmann, A.; Gabel, F.; Camilloni, C. et al. Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nature Comm.* **2018**, *9*, 2479.

(13) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **2007**, *92*, 3817–3829.

(14) Yildirim, I.; Stern, H. A.; Kennedy, S. D.; Tubbs, J. D.; Turner, D. H. Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine. *J. Chem. Theory Comput.* **2010**, *6*, 1520–1531.

(15) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

(16) Yildirim, I.; Kennedy, S. D.; Stern, H. A.; Hart, J. M.; Kierzek, R.; Turner, D. H. Revision of AMBER torsional parameters for RNA improves free energy predictions

for tetramer duplexes with GC and iGiC base pairs. *J. Chem. Theory Comput.* **2011**, *8*, 172–181.

(17) Chen, A. A.; Garcia, A. E. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 16820–16825.

(18) Bergonzo, C.; Cheatham III, T. E. Improved force field parameters lead to a better description of RNA structure. *J. Chem. Theory Comput.* **2015**, *11*, 3969–3972.

(19) Gil-Ley, A.; Bottaro, S.; Bussi, G. Empirical Corrections to the Amber RNA Force Field with Target Metadynamics. *J. Chem. Theory Comput.* **2016**, *12*, 2790–2798.

(20) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J. Chem. Theory Comput.* **2016**, *12*, 6192–6200.

(21) Aytenfisu, A. H.; Spasic, A.; Grossfield, A.; Stern, H. A.; Mathews, D. H. Revised RNA dihedral parameters for the Amber force field improve RNA molecular dynamics. *J. Chem. Theory Comput.* **2017**, *13*, 900–915.

(22) Tan, D.; Piana, S.; Dirks, R. M.; Shaw, D. E. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E1346–1355.

(23) Kuhrova, P.; Mlynsky, V.; Zgarbova, M.; Krepl, M.; Bussi, G.; Best, R. B.; Otyepka, M.; Sponer, J.; Banas, P. Improving the performance of the Amber RNA force field by tuning the hydrogen-bonding interactions. *J. Chem. Theory Comput.* **2019**, 10.1021/acs.jctc.8b00955.

(24) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water dispersion interactions

strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

(25) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **2008**, *94*, 182–192.

(26) Li, D.-W.; Brüschweiler, R. Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins. *J. Chem. Theory Comput.* **2011**, *7*, 1773–1782.

(27) Wang, L.-P.; Chen, J.; Van Voorhis, T. Systematic parametrization of polarizable force fields from quantum chemistry data. *J. Chem. Theory Comput.* **2012**, *9*, 452–460.

(28) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building force fields: an automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

(29) Chen, J.; Chen, J.; Pinamonti, G.; Clementi, C. Learning effective molecular models from experimental observables. *J. Chem. Theory Comput.* **2018**, *14*, 3849–3858.

(30) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(31) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(32) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(33) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(34) Bottaro, S.; Di Palma, F.; Bussi, G. The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.* **2014**, *42*, 13306–13314.

(35) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

(36) Branduardi, D.; Bussi, G.; Parrinello, M. Metadynamics with adaptive Gaussians. *J. Chem. Theory Comput.* **2012**, *8*, 2247–2254.

(37) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(38) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(39) Steinbrecher, T.; Latzer, J.; Case, D. A. Revised AMBER Parameters for Bioorganic Phosphates. *J. Chem. Theory Comput.* **2012**, *8*, 4405–4412.

(40) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **2014**, *5*, 3863–3871.

(41) Tubbs, J. D.; Condon, D. E.; Kennedy, S. D.; Hauser, M.; Bevilacqua, P. C.; Turner, D. H. The Nuclear Magnetic Resonance of CCCC RNA Reveals a Right-Handed Helix, and Revised Parameters for AMBER Force Field Torsions Improve Structural Predictions from Molecular Dynamics. *Biochemistry* **2013**, *52*, 996–1010.

(42) Ennifar, E.; Nikulin, A.; Tishchenko, S.; Serganov, A.; Nevskaya, N.; Garber, M.; Ehresmann, B.; Ehresmann, C.; Nikonov, S.; Dumas, P. The crystal structure of UUCG tetraloop1. *J. Mol. Biol.* **2000**, *304*, 35–42.

(43) Trausch, J. J.; Xu, Z.; Edwards, A. L.; Reyes, F. E.; Ross, P. E.; Knight, R.; Batey, R. T. Structural basis for diversity in the SAM clan of riboswitches. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6624–6629.

(44) Proctor, D. J.; Ma, H.; Kierzek, E.; Kierzek, R.; Gruebele, M.; Bevilacqua, P. C. Folding thermodynamics and kinetics of YNMG RNA hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. *Biochemistry* **2004**, *43*, 14004–14014.

(45) Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451.

(46) Cesari, A.; Reißer, S.; Bussi, G. Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation* **2018**, *6*, 15.

(47) Rangan, R.; Bonomi, M.; Heller, G. T.; Cesari, A.; Bussi, G.; Vendruscolo, M. Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.* **2018**, *14*, 6632–6641.

(48) Bottaro, S.; Bussi, G.; Pinamonti, G.; Reisser, S.; Boomsma, W.; Lindorff-Larsen, K. Barnaba: Software for Analysis of Nucleic Acids Structures and Trajectories. *RNA* **2019**, *25*, 219–231.

(49) Nichols, P. J.; Henen, M. A.; Born, A.; Strotz, D.; Güntert, P.; Vögeli, B. High-resolution small RNA structures from exact nuclear Overhauser enhancement measurements without additional restraints. *Commun. Biol.* **2018**, *1*, 61.

(50) Nozinovic, S.; Fürtig, B.; Jonker, H. R.; Richter, C.; Schwalbe, H. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **2009**, *38*, 683–694.