Boosting Ensemble Refinement with Transferable Force-Field Corrections: Synergistic Optimization for Molecular Simulations

Boosting Ensemble Refinement with Transferable Force Field Corrections: Synergistic Optimization for Molecular Simulations

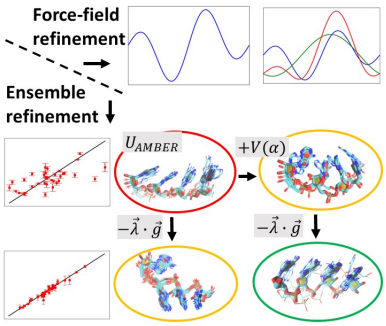
Ivan Gilardoni,[†] Thorben Fröhlking,[†,‡] and Giovanni Bussi[*,†]

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

‡ *Current address: Université de Genève, Switzerland*

E-mail: bussi@sissa.it

Abstract



A novel method combining the ensemble refinement by maximum entropy principle and the force field fitting approach is presented. Its formulation allows to continuously interpolate in between these two methods, which can thus be interpreted as two limiting cases. A cross-validation procedure enables to correctly assess the relative weight of both of them, distinguishing scenarios where the combined approach is meaningful from those in which either ensemble refinement or force field fitting separately prevails. The efficacy of their combination

is examined for a realistic case study of RNA oligomers. Within the new scheme, molecular dynamics simulations are integrated with experimental data provided by nuclear-magnetic-resonance measures. We show that force field corrections are in general superior when applied to the appropriate force field terms, but are automatically discarded by the method when applied to inappropriate force field terms.

Molecular dynamics (MD) simulations play a crucial role in resolving the underlying conformational dynamics of molecular systems.[1] However, their capability to reproduce and predict dynamics in agreement with experiments is limited by the statistical significance of the sampled trajectory and the accuracy of the force field model. While the first issue can be addressed by using enhanced sampling techniques,[2] the second one can be faced by suitable integration of MD simulations and experimental data.[3] To this aim, two main philosophies for experiment-based refinement were proposed in the literature.[4] The first one is the so-called ensemble refinement (ER) approach.[5–13] Developed from the maximum entropy principle, this technique selects the ensemble which best describes the experimental measures and is, at the same time, as close as possible to the initially hypothesized one. In doing so, ER is agnostic with respect to the knowledge of the force field parametrization: the functional form of the corrections carried to the initial ensemble only depends on the selected observables[5,14] and thus the corrections are not transferable to different systems. The second philosophy is force field refinement (FFR).[15–24] Based on a reasonable guess of the force-field correction terms, their optimal coefficients are determined by minimizing a loss function that includes the discrepancy from experimental data and, in modern implementations, a regularization term that penalizes moving away from the initial force field, in a Bayesian line of thought. This approach enables one to encode prior information about the reliability of a given force field term, by choosing which specific term should be refined, and makes the resulting corrections transferable to other systems.[25] However, adding the same force-field correction terms to all the copies of a given residue could be over-limiting and not able to capture further relevant differences among them. Indeed, the functional form of the force-field is limited and might be intrinsically unable to reproduce experimental data. These two categories of methods have been

traditionally derived in a different manner. Only recently, a formulation of the FFR approach that formally relates to the maximum entropy principle has been proposed.[21] Importantly, methods of the two classes have been so far used in a disjoint fashion. The user is thus expected to decide based on experience if transferable or non-transferable corrections are performing best for a given system.

In this Letter, we introduce a procedure to seamlessly combine the ER method with FFR. This allows to preserve the flexibility of ER while at the same time ensuring the transferability of the resulting force-field corrections to different molecules as in FFR. The procedure is here applied to the refinement of conformational ensembles of RNA oligomers, for which nuclear-magnetic-resonance (NMR) experimental data are available, but can be applied to reweight conformational ensembles of arbitrary systems for which solution data are available. In a nutshell, the method works as follows. In traditional FFR approaches, the original ensemble P_0 is reweighted to include force-field corrections resulting in a new ensemble P_ϕ, which is then compared with experiment (for example through the χ^2). Corrections are chosen so as to maximise the agreement. Here, before comparing with experiment, we perform an additional ER step, which fine tunes the resulting weights in a new ensemble P. The former step is expected to take into account any transferable contribution, and to leverage on the knowledge of which force-field terms might benefit a refinement. The latter step makes sure the final ensemble averages agrees with experiment. The combination of the ER and FFR approaches is controlled by two hyperparameters (α and β), as prescribed by the loss function that we adopt here:

$$\mathscr{L}[P,P_\phi] = \frac{1}{2}\chi^2[P] + \alpha D_{KL}[P|P_\phi] + \beta D_{KL}[P_\phi|P_0], \tag{1}$$

with D_{KL} the Kullback-Leibler divergence. While the first term quantifies the agreement of P with experiments, the following two regulate the closeness of P from P_ϕ and of P_ϕ from the original ensemble P_0, respectively (see Supporting Information for more details). We first show the behavior of this approach on a toy model. Then, we use the method to derive ensembles and force

field corrections for RNA oligomers. For the latter case, we show how a carefully performed cross-validation procedure is necessary to tune the hyperparameters. In our tests, we intentionally investigate the case where inappropriate force-field corrections are attempted, showing that our cross-validation procedure can detect this issue and automatically switch off the FFR step.

Proof of concept. To check the validity of the combined refinement, we set up a simple toy model which consists of a two-dimensional probability distribution with four peaks (see Fig. 1). The initial hypothesis sets most probability in the top-left peak ($x < 0$ and $y > 0$), while the ground truth probability is distributed also on the top-right peak ($x > 0$ and $y > 0$). So, the average value of the x observable is underestimated by the initial hypothesis, whereas the average value of the y observable is approximately correct. We then correct the initial hypothesis purely based on the value of the observed averages of x and y. Ensemble refinement shifts the probability from the two peaks at $x < 0$ to those at $x > 0$, perfectly matching the observed averages. According to the maximum entropy principle, this is the minimal correction to the prior ensemble that allows matching the observed averages. The resulting ensemble is guaranteed to be closer to the ground truth one,[26] but still not necessarily identical. We then assume that a physical knowledge of the system suggests the top-right and bottom-left peaks to be coupled, leading to a specific functional form for the force-field correction. By performing a force-field refinement with this additional information, the observed averages are not exactly matched, but the obtained ensemble is also getting closer to the ground truth. Including further flexibility through the combined approach introduced in this work allows to optimally combine the information used in the force field refinement approach with the maximum entropy principle, resulting in better agreement with the ground truth ensemble than the one obtained applying any of the two methods separately. Fig. 1b reports the distance from ground truth for the ensembles obtained using different possible values of α and β. The method interpolates between no correction to ensemble refinement, force-field refinement, and any combination of the two, as indicated in the figure. A similar figure reporting the discrepancy between the predicted and experimental observables (χ^2) is reported in Fig. S1. A suitable choice of the two hyperparameters α, β is required to avoid overfitting. This is particularly relevant considering that

experimental data are only known with a given uncertainty. As this effect is not present in this toy model, hyperparameters tuning can be better illustrated using a more realistic example.

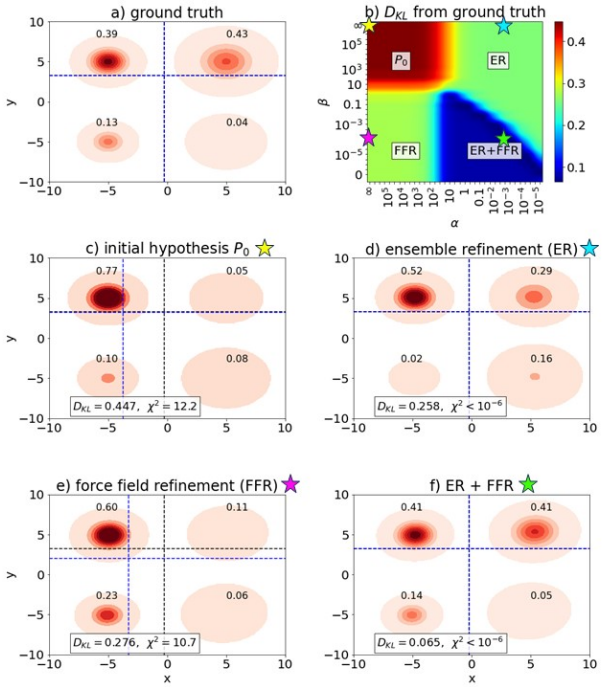


Figure 1: Results on a toy model. (a) Ground truth distribution. Populations are also reported for each peak. (c) Initially assumed distribution. (d) Ensemble refinement leads to an ensemble closer, but still not identical, to the ground truth. (e) An attentive choice of the force field correction term could result in a different refinement (FFR), as good as the previous one (compare the Kullback-Leibler divergences from the ground truth, indicated as D_{KL}), with the benefit to be transferable. If this correction is not sufficient, (f) further flexibility can be included through the combined approach. In panels a,c,d,e,f, black dashed lines report the averages computed using the ground truth, whereas blue dashed lines report the averages computed using the refined ensemble. (b) Hyperparameters scan. The hyperparameters values prove to be crucial for a proper balancing of the ER and FFR contributions. In panel (b), the values of the hyperparameters used to generate the ensembles in panels (c,d,e,f) are indicated with a star of a matching color.

Application to real systems, including cross validation. We then test the method on a set of RNA oligomers for which simulations were previously reported,[27] using the same experimental data set that was used in Ref. [27] Experimental data, corresponding to torsion angles $\beta, \varepsilon, \gamma$, are taken from Refs.[28–32] We perform the minimization of the loss function obtained combining all the oligomers in the training set (AAAA, CCCC, GACC, UUUU, UCAAUC – Fig. 2a) with a scan in the space of the hyperparameters α and β. The limiting cases of ensemble refinement (scan on the hyperparameter α at $\beta = \infty$) and force field fitting (scan on the hyperparameter β at $\alpha = \infty$) are

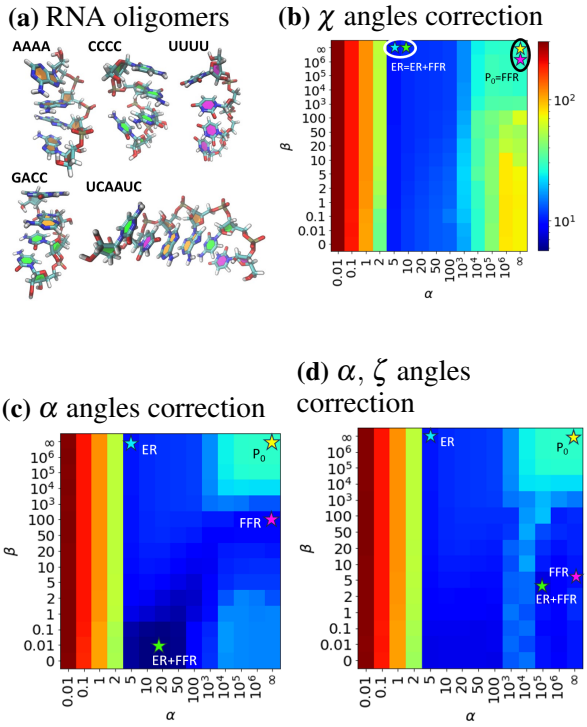


Figure 2: (a) Case study: RNA oligomers (4 tetramers and 1 hexamer). (b,c,d) Reduced χ^2 on validating observables using three different functional forms for the force field refinement step (cross validation averages). In particular: (b) the correction on χ angles is fruitless and the contribution of (non-transferable) ensemble refinement is essential; (c) the correction on α angles is profitable, however adding more flexibility results in a better agreement with experimental values; (d) the correction on α, ζ angles alone is enough and the inclusion of further flexibility is not necessary.

included, as well as the initially assumed ensemble, corresponding to $\alpha = \beta = \infty$ (i.e., no refinement). For each considered value of the hyperparameters α, β, we take 20 random choices (seeds) of 70% frames (of "demuxed" – continuous – trajectories, from replica exchange simulations in temperature range $275 - 400$ K, with the exception of UCAAUC for which a corrupted trajectory made it impossible to generate the continuous trajectories) and 70% observables to be used as a training set (same choices for all the sampled hyperparameters), implementing a bootstrap strategy.[33] The remaining observables are used to evaluate the reduced χ^2 on the full trajectory, i.e., including also frames which were employed in training (validating step). The achievement of the minimizations is reflected by an increase in the minimum value of the loss function at given seeds when α or β are increased. Given the way trajectories are bootstrapped, the reduced χ^2 on training observables and the one on validating observables coincide within their statistical error at $\alpha = \beta = \infty$, since it corresponds to the original ensemble.

The two hyperparameters (α and β) control the flexibility of the fitting. By decreasing one or both, the reliability given to the original assumptions on the force field is reduced in favour of the confidence on experimental measures. In other words, low values of the hyperparameters correspond to high flexibility in the correction to the ensemble, which means strong ability to fit the (training) experimental values. This is particularly true for the hyperparameter α, corresponding to the ensemble refinement direction. For β (force field refinement direction), the flexibility is instead intrinsically limited by the constrained functional form of the force field correction. Whereas a limited flexibility can provide a (physically meaningful) improvement in the description of the molecules, an uncontrolled flexibility may lead to overfit the data, disregarding their intrinsic experimental error. Cross validation is all about assessing the appropriate importance of this flexibility, which in the method here proposed plays on two different directions: the non-transferable ensemble refinement and the force field fitting ones. This task is performed by evaluating the error (the reduced χ^2) on left-out observables, namely those which are not used in this training step to determine the optimal ensemble.

Whereas the χ^2_{red} computed on training data is decreasing when decreasing the values of the

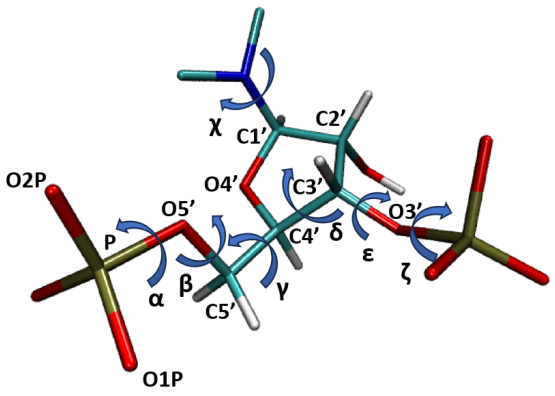


Figure 3: Sample RNA backbone structure, with standard atom names and dihedral angles indicated. The nucleobase is truncated so that only two carbon atoms from the cycle are shown.

two hyperparameters, the χ^2_{red} computed on validating observables (Fig. 2 b,c,d) does so only over a certain range of the hyperparameters, signaling when under/over-fitting occurs. Firstly, we consider the case of traditional ensemble refinement, which in the combined method corresponds to $\beta = \infty$ (first row in the plots, independent on the selected force field correction; Fig. S2a). One can notice how, starting from the prior ensemble P_0 at $\alpha = \infty$, the error χ^2_{red} decreases with α, up to a certain point, where it starts to increase. Such point of minimum (approximately at $\alpha \simeq 5$, with $\chi^2_{red} \simeq 10$) marks the transition from the under-fitting to over-fitting scenarios, respectively. We then move forward to transferable corrections. It is instructive to consider three different functional forms for the force field refinement approach. All of them are given by linear combinations of sine and cosine of selected dihedral angle, respectively: χ, O4′–C1′–N1′–C2′ for pyrimidines and O4′–C1′–N9′–C4′ for purines; α, $O3'_{i-1}$–P–O5′–C5′; and combined α, ζ, C3′–O3′–P_{i+1}–$O5'_{i+1}$ (see Fig. 3 for a representative representation of RNA backbone dihedral angles); in this last case we restricted to equal coefficients for sine terms and equal ones for cosine terms. Such force field corrections exhibit the three different behaviours which are expected when applying the combined ER+FFR method. The first attempted correction (χ angles, Fig. 2b and Fig. S2b) applied alone is fruitless, since the χ^2_{red} at $\alpha = \infty$, i.e. in the FFR regime, is larger than in the original ensemble P_0 for any choice of $\beta < \infty$. The contribution of (non-transferable) ensemble refinement is thus essential. Including both contributions, the minimum χ^2_{red} results at $\alpha = 5, \beta = \infty$. Hence, this case corresponds to the extreme in which adding the contribution of the force field refinement does

not improve the description with respect to ensemble refinement alone. Even in this particularly difficult case, the ensemble refinement step is able to report reasonable cross-validation observables (see Table 1). The second attempted correction (on α angles) is profitable, as shown at $\alpha = \infty$ (FFR only, Fig. 2c; see also Fig. S2c). However adding more flexibility with the combination of ER and FFR proves to be profitable with respect to the separate application of either the two methods since it results in a lower cross-validation error. Finally, the correction on α, ζ angles alone is sufficient and the inclusion of further flexibility is leading to a minor improvement in the agreement with experiment (Fig. 2d and Fig. S2d). This is close to the other extreme, in which FFR alone shows to be optimal.

Table 1: Results of cross validation on training molecules (see also Fig. 2). We compare ER, FFR and their combination ER+FFR, with three different force field corrections. For each case, we report the optimal α, β hyperparameters and the minimum value of χ^2_{red} on validating observables (cross validation averages). In the first line we report the χ^2_{red} before any corrections on the ensembles.

force field correction	method	optimal α, β	χ^2_{red}
-	no reweighting	$\alpha = \infty, \beta = \infty$	28.32
-	ER ($\beta = \infty$)	$\alpha = 5$	10.10
χ angles	FFR ($\alpha = \infty$)	$\beta = \infty$	28.32
χ angles	ER+FFR	$\alpha = 5, \beta = \infty$	10.10
α angles	FFR ($\alpha = \infty$)	$\beta = 100$	9.33
α angles	ER+FFR	$\alpha = 20, \beta = 0.01$	5.70
α, ζ angles	FFR ($\alpha = \infty$)	$\beta = 5$	9.81
α, ζ angles	ER+FFR	$\alpha = 10^5, \beta = 2$	8.02

Optimal force field corrections. Once the optimal values of the α, β hyperparameters have been determined through cross validation, we estimate the coefficients of the force field corrections by minimizing the loss function with such hyperparameters on the whole data set (without distinguishing between training and test set). The associated uncertainty is quantified by the standard deviation of the coefficients resulting from cross validation, as it is conventionally done in bootstrap analysis.[33] The average values on the bootstrap samples are compatible with the values from whole minimization within the obtained uncertainties. The results are reported in Table 2. The correction on the χ dihedral angles is null, since the optimal hyperparameters correspond to ensemble refinement only. This implies that relevant force field corrections are not in the χ bond,

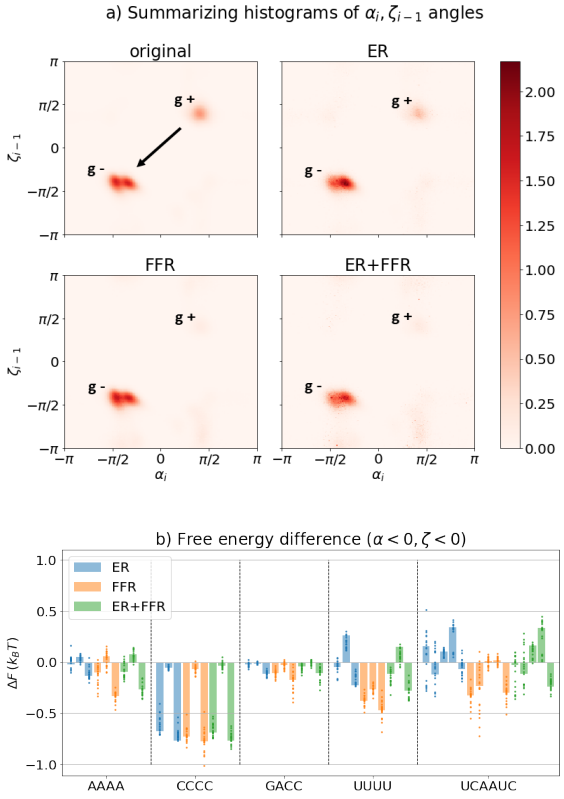


Figure 4: (a) Summarizing histograms of α, ζ angles before/after corrections on α dihedral angles, averaged over all the oligomers; (b) free energy difference for the $(\alpha, \zeta) \in (-\pi, 0) \times (-\pi, 0)$ region, for each molecule and phosphate (bars corresponding to minimization on the whole set of observables, dots corresponding to minimizations in cross validation).

rather in the backbone structure. Indeed, for the other two cases (correction on α or α, ζ) the force field correction coefficients are significantly different from zero, with the general effect of disfavouring gauche+ conformations for both angles.

We then study how the force-field corrections on the α angles modify the distribution of α, ζ dihedral angles, comparing the three different refinement methods described above. In these constructs, the phosphate group of the first nucleotide is absent (in agreement with experiments) so the $\alpha_i\, \zeta_{i-1}$ dihedral angles encompass the i-th phosphate group, with $i = 2, \ldots, N$. In Fig. 4a) we show the overall histogram of the α_i, ζ_{i-1} distributions, averaged over molecules and positions, which highlights two dominant peaks in the original ensembles, corresponding to gauche+ (g+) and gauche− (g−). The population in these two peaks is modified when introducing ensemble and/or force field corrections, with a general increase in the $\alpha_i(g-), \zeta_{i-1}(g-)$ region, to the disadvantage of the $\alpha_i(g+), \zeta_{i-1}(g+)$ area. This is in agreement with previous studies.[34–36] To better visualize these variations, Fig. 4b reports the free energy differences ΔF associated to the $\alpha_i(g-), \zeta_{i-1}(g-)$ region, separately for each oligomer and phosphate group. We notice how, for AAAA, CCCC, and GACC tetramers, the refinements have a significant impact only on the α_i, ζ_{i-1} angles corresponding to the first and last phosphate, while the population for the middle phosphate is almost unchanged. This can be explained on the basis that the employed force-field is well-suited for long RNA molecules, for which the first and last phosphate constitute a small fraction of the whole molecule, so that *ad hoc* correction at the termini might be convenient.[37] In particular, both the FFR and ER+FFR corrections go in the same direction as ER, favouring $\alpha(g-), \zeta(g-)$ angles as expected above. Results for the UUUU tetramer instead show significant free energy differences also for the intermediate phosphate. Here, ER and FFR suggest opposite corrections, with the former disfavouring $\alpha_3(g-)$, $\zeta_2(g-)$. Also for the UCAAUC hexamer, the α_i, ζ_{i-1} dihedral angles corresponding to $i = 4, 5$ tend to be modified by both the ER and ER+FFR methods, and left unchanged by FFR correction. Given the better performance of the ER+FFR approach in cross validation tests (see χ^2_{red} in Table 1), we argue that the ensembles obtained with the ER+FFR approach are more reliable than those obtained using the ER or the FFR approach

alone.

It is also instructive to monitor the effect of the corrections to the least visited region in the α,ζ domain (see Fig. S4). This test highlights the difficulty of performing cross-validation tests for poorly populated region, for which modified weights might have a minor impact in both the discrepancy with respect to experiment (χ_{red}) and the distance from the prior distribution (D_{KL}).

Table 2: Force-field correction coefficients (measure unit: $k_BT = 2.49 kJ/\text{mol}$). We report the coefficients resulting from the minimization on the whole data set together with their uncertainty. For the correction on χ angles, as shown in Fig. 2, FFR provides no correction and optimal ER+FFR corresponds to ER only, hence no transferable corrections. The functional form for the force field corrections is $V(\chi) = \phi_1 \sin\chi + \phi_2 \cos\chi$, $V(\alpha) = \phi_1 \sin\alpha + \phi_2 \cos\alpha$ and $V(\alpha,\zeta) = \phi_1(\sin\alpha + \sin\zeta) + \phi_2(\cos\alpha + \cos\zeta)$ respectively; sum over all the specified dihedral angles present in the molecule is implicit.

ff correction / method	ϕ_1	ϕ_2
χ angles	$\phi_1(\sin\chi)$	$\phi_2(\cos\chi)$
FFR ($\alpha=\infty,\beta=\infty$)	0	0
ER+FFR ($\alpha=5,\beta=\infty$)	0	0
α angles	$\phi_1(\sin\alpha)$	$\phi_2(\cos\alpha)$
FFR ($\alpha=\infty,\beta=100$)	0.91 ± 0.16	1.67 ± 0.85
ER+FFR ($\alpha=20,\beta=0.01$)	0.51 ± 0.08	1.65 ± 0.36
α,ζ angles	$\phi_1(\sin)$	$\phi_2(\cos)$
FFR ($\alpha=\infty,\beta=5$)	3.0 ± 1.5	4.0 ± 2.0
ER+FFR ($\alpha=10^5,\beta=2$)	3.0 ± 1.6	4.0 ± 1.8

Testing the force field corrections on left-out molecules. Finally, we compare the performance of the two force fields that we obtained with the FFR and ER+FFR procedure when transferred to new molecules, not considered in training, namely the CAAU and UCUCGU oligomers. To this aim, we employ the optimal coefficients of the force field corrections that were reported in Table 2 to reweight the CAAU and UCUCGU ensembles. We do so without performing a new ensemble refinement procedure. The corresponding χ^2_{red} is then evaluated on the whole set of observables (see Table 3). These two molecules are quite different and respond differently to the correction fitted on the training set. Specifically, the original ensemble of CAAU has a very large χ^2 which is dramatically decreased by the force field corrections. The larger the penalty on the gauche+ (g+) rotamers, the better the agreement with experiment. As a consequence, the correction applied on the α,ζ angles results in the best agreement with experiment. This result is partly unexpected,

because the FFR correction on α, ζ angles was performing worse than the FFR+ER approach in our cross-validation tests. We speculate that the inclusion in the original training set of the UCAAUC oligomer, which contains the CAAU sequence, makes the inferred FFR well suitable for CAAU, even though this system is not present in the training set. Conversely, the UCUCGU hexamer has a moderate χ^2. Interestingly, the force field corrections obtained on the training set are not capable on improving the agreement of the corresponding ensemble with experiment. In this case, all corrections lead to some degree of overfitting. The mixed FFR+ER approach on the α angle, which leads to more conservative force field corrections, results in smaller overfitting and in a χ^2 comparable to the one obtained with the original force field.

Table 3: Reduced χ^2 on validating molecules, based on the force-field corrections introduced above (coefficients resulting from minimization of the training molecules on the whole data set).

	CAAU	UCUCGU	both
no reweighting	243.8	14.1	211.3
α angles correction			
FFR	24.8	16.4	23.6
ER+FFR	57.5	14.3	51.4
α, ζ angles correction			
FFR	11.7	23.8	13.4
ER+FFR	11.6	23.7	13.3

In summary, this study reports a strategy to boost the efficiency of ensemble refinement methods by preceding them with a knowledge-based force field refinement step. This combined method allows also to obtain force-field corrections which can then be transferred to different systems, and can outperform normal ensemble refinement in cross validation tests. Differently from force field refinement by itself, these corrections are derived taking explicitly into account the fact that transferable corrections might not be able to match experiments simultaneously in multiple systems. This specificity on the system is guaranteed by the ensemble refinement step. Whereas the force field refinement and ensemble refinement methods have been separately applied in several works, we are not aware of any attempt made to combine their strengths in a single approach. We apply the proposed method to a realistic case study of RNA oligomers. Despite the appar-

ent simplicity of these small RNA molecules, current force-fields are still limited in correctly generating structural ensembles, which therefore can be used to improve molecular potentials. We use a robust cross validation protocol to select the suitable values of the two hyperparameters α, β and then analyze the predictions about both training and validation oligomers. The scripts used to perform the refinements discussed in this work can be found at `https://github.com/bussilab/force-field-ensemble-refinement`. The analyzed time series can be found at `https://doi.org/10.5281/zenodo.10185005`.

In this work we designed force field corrections to be partly independent from the measured observables, by correcting dihedral angles that were not directly measured. However, correlations indirectly appear through other interactions. Correlated observables used in ensemble refinement act cooperatively and are known not to be a problem, since corrections on the different observables lead to equivalent ensembles.[19] However, correlated corrections in force-field refinement or in the combined method introduced here might lead to multiple inequivalent solutions. In the extreme case of identical force-field corrections and observables, we can expect that the cross-validation procedure used here will make sure that a fraction of the correction that can be safely transferred to other copies of the same residue is included in the force-field correction, whereas the remaining part is included in the ensemble refinement part.

The introduced method interpolates between ER and FFR, reducing to either of the two whenever the other approach does not provide significant improvement of the χ^2 on computed on a validating set of observables. This happens with the force-field correction on glycosidic bond angles, which does not exhibit any improvement in cross validation, therefore the combined approach selects ER alone, resulting in better agreement with experiments than FFR alone. On the opposite, the correction on α, ζ dihedral angles, with identical coefficients, has enough flexibility so that the inclusion of system-specific refinement turns out to be irrelevant, and FFR alone is the optimal solution. However, as shown by comparison of Figs. 2 b, c, and d, the best agreement with experimental data is obtained through the flexible correction on α dihedral angles only using the combined method.

In order to minimize the loss function we employ a reweighting of the ensembles approach. This might be a weakness of the proposed approach, common to all reweighting methods, due to its potentially low statistical efficiency.[38,39] This concept can be quantified as the effective number of frames, computed through the Kish sample size or analogously the relative entropy. While for the examined oligomers the effective number of frames still remains a significant fraction of the whole amount (see Fig. S3), for more complex systems, it might not be the case. Performing new MD simulations during the minimization, for instance whenever the resulting ensembles move too far from the initial ones, would lead to greater statistical robustness. Also, on-the-fly restraining could be performed.[7,19,40–42] A second, related issue arises from regions of the conformational space with limited or no sampling. Due to the way our cross validation procedure is performed, left-out portions of the initial trajectory are used to test for overfitting. However, if samples from a region are never observed in the initial trajectory, it is impossible to use reweighting to predict which will be the effect of the correction on samples from that region. If the region has a very large energy, e.g. because it is sterically forbidden, any change to the potential energy function is irrelevant. But if the region can be sampled in a new simulation for the same or for a different system, overfitting issues will arise.[25] This is a well-known issue in force-field fitting strategies, where it is common to perform new simulations using the refined force-field parameters to test for these artifacts. It might also be an issue in ensemble refinement maximum entropy strategies, if consecutive simulations are performed including linear corrections to the energy function. This issue is instead not expected to be visible if no new simulations are performed and the resulting ensemble is reported *as is*.

Finally, the discrepancy may be due not only to incorrect structural ensembles but also to inaccurate forward models used to compute experimental observables from MD simulations. In the most extreme cases, the ensembles might be in perfect agreement with the ground truth, still having high χ^2_{red} due to wrong forward models (like, for example, the empirical coefficients of Karplus equations). In a recent work, we have shown how to simultaneously optimize ensembles and forward models.[27] This idea could be pushed further and, in combination with the ideas presented

here, lead to a simultaneous optimization of force fields, ensembles, and forward models.

Supporting Information

Supplementary Methods: Combining ensemble and force field refinements; Interpretation of the hyperparameters; Calculations using reweighting; Generalization to multiple systems; Minimization strategy; Cross validation; Simulation details; Experimental data. Supplementary Results: Toy model; RNA oligomers.

Data Availability

The scripts used to perform the refinements discussed in this work can be found at `https://github.com/bussilab/force-field-ensemble-refinement`. The analyzed time series can be found at `https://doi.org/10.5281/zenodo.10185005`.

References

(1) Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99*, 1129–1143.

(2) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced sampling methods for molecular dynamics simulations. *Living J. Comput. Mol. Sci* **2022**, *4*, 1583.

(3) Bottaro, S.; Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **2018**, *361*, 355–360.

(4) Orioli, S.; Larsen, A. H.; Bottaro, S.; Lindorff-Larsen, K. How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* **2020**, *170*, 123–176.

(5) Pitera, J. W.; Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comp.* **2012**, *8*, 3445–3451.

(6) Beauchamp, K. A.; Pande, V. S.; Das, R. Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys. J.* **2014**, *106*, 1381–1390.

(7) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, 243150.

(8) Brookes, D. H.; Head-Gordon, T. Experimental inferential structure determination of ensembles for intrinsically disordered proteins. *J. Am. Chem. Soc.* **2016**, *138*, 4530–4538.

(9) Capelli, R.; Tiana, G.; Camilloni, C. An implementation of the maximum-caliber principle by replica-averaged time-resolved restrained simulations. *J. Chem. Phys.* **2018**, *148*, 184114.

(10) Köfinger, J.; Stelzl, L. S.; Reuter, K.; Allande, C.; Reichel, K.; Hummer, G. Efficient ensemble refinement by reweighting. *J. Chem. Theory Comput.* **2019**, *15*, 3390–3401.

(11) Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. *Structural bioinformatics: methods and protocols* **2020**, 219–240.

(12) Medeiros Selegato, D.; Bracco, C.; Giannelli, C.; Parigi, G.; Luchinat, C.; Sgheri, L.; Ravera, E. Comparison of different reweighting approaches for the calculation of conformational variability of macromolecules from molecular simulations. *ChemPhysChem* **2021**, *22*, 127–138.

(13) Brotzakis, Z. F.; Vendruscolo, M.; Bolhuis, P. G. A method of incorporating rate constants as kinetic constraints in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2012423118.

(14) Cesari, A.; Reißer, S.; Bussi, G. Using the maximum entropy principle to combine simulations and solution experiments. *Computation* **2018**, *6*, 15.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

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

(20) Cesari, A.; Bottaro, S.; Lindorff-Larsen, K.; Banáš, P.; Šponer, J.; Bussi, G. Fitting corrections to an RNA force field using experimental data. *J. Chem. Theory Comp.* **2019**, *15*, 3425–3431.

(21) Köfinger, J.; Hummer, G. Empirical optimization of molecular simulation force fields by Bayesian inference. *Eur. Phys. J. B* **2021**, *94*, 245.

(22) Fröhlking, T.; Mlỳnskỳ, V.; Janeček, M.; Kührová, P.; Krepl, M.; Banáš, P.; Šponer, J.; Bussi, G. Automatic learning of hydrogen-bond fixes in an AMBER RNA force field. *J. Chem. Theory Comput.* **2022**, *18*, 4490–4502.

(23) Bolhuis, P. G.; Brotzakis, Z. F.; Keller, B. G. Optimizing molecular potential models by imposing kinetic constraints with path reweighting. *J. Chem. Phys.* **2023**, *159*, 074102.

(24) Kümmerer, F.; Orioli, S.; Lindorff-Larsen, K. Fitting Force Field parameters to NMR Relaxation Data. *J. Chem. Theory Comput.* **2023**, *19*, 3741–3751.

(25) Fröhlking, T.; Bernetti, M.; Calonaci, N.; Bussi, G. Toward empirical force fields that match experimental observables. *J. Chem. Phys.* **2020**, *152*, 230902.

(26) Dannenhoffer-Lafage, T.; White, A. D.; Voth, G. A. A direct method for incorporating experimental data into multiscale coarse-grained models. *J. Chem. Theory Comput.* **2016**, *12*, 2144–2153.

(27) Fröhlking, T.; Bernetti, M.; Bussi, G. Simultaneous refinement of molecular dynamics ensembles and forward models using experimental data. *J. Chem. Phys.* **2023**, *158*, 214120.

(28) 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 Comp.* **2015**, *11*, 2729–2742.

(29) 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.

(30) Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H. Benchmarking AMBER force fields for RNA: Comparisons to NMR spectra for single-stranded r (GACC) are improved by revised χ torsions. *J. Phys. Chem. B* **2011**, *115*, 9261–9270.

(31) Zhao, J.; Kennedy, S. D.; Berger, K. D.; Turner, D. H. Nuclear magnetic resonance of single-stranded RNAs and DNAs of CAAU and UCAAUC as benchmarks for molecular dynamics simulations. *J. Chem. Theory Comp.* **2020**, *16*, 1968–1984.

(32) Zhao, J.; Kennedy, S. D.; Turner, D. H. Nuclear magnetic resonance spectra and AMBER OL3 and ROC-RNA simulations of UCUCGU reveal force field strengths and weaknesses for single-stranded RNA. *J. Chem. Theory Comp.* **2022**, *18*, 1241–1254.

(33) Efron, B.; Tibshirani, R. An introduction to the bootstrap; Chapman & Hall, New York, 1993.

(34) Gil-Ley, A.; Bottaro, S.; Bussi, G. Empirical corrections to the AMBER RNA force field with target metadynamics. *J. Chem. Theory Comp.* **2016**, *12*, 2790–2798.

(35) Bottaro, S.; Banáš, P.; Špöner, J.; Bussi, G. Free energy landscape of GAGA and UUCG RNA tetraloops. *J. Phys. Chem. Lett.* **2016**, *7*, 4032–4038.

(36) Chen, J.; Liu, H.; Cui, X.; Li, Z.; Chen, H.-F. RNA-specific force field optimization with CMAP and reweighting. *J. Chem. Inf. Model.* **2022**, *62*, 372–385.

(37) Mlýnský, V.; Kührová, P.; Kühr, T.; Otyepka, M.; Bussi, G.; Banáš, P.; Špöner, J. Fine-tuning of the AMBER RNA force field with a new term adjusting interactions of terminal nucleotides. *J. Chem. Theory Comp.* **2020**, *16*, 3936–3946.

(38) 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 Comp.* **2018**, *14*, 6632–6641.

(39) Shen, T.; Hamelberg, D. A statistical analysis of the precision of reweighting-based simulations. *J. Chem. Phys.* **2008**, *129*, 034103.

(40) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138*, 094112.

(41) White, A. D.; Voth, G. A. Efficient and minimal method to bias molecular simulations with experimental data. *J. Chem. Theory Comp.* **2014**, *10*, 3023–3030.

(42) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2016**, *2*, e1501177.