DFFR: A NEW METHOD FOR HIGH-THROUGHPUT RECALIBRATION OF AUTOMATIC FORCE-FIELDS FOR DRUGS

David Moreno[1], Sanja Zivanovic[1], Francesco Colizzi[1], Adam Hospital[1], Juan Aranda[1],
Robert Soliva[2] and Modesto Orozco[1,3]*

We present DFFR (drug force-field recalibration) a new method for the refining automatic force-fields used to represent small drugs in docking and molecular dynamics simulations. The method is based on a fine tuning of torsional terms to obtain ensembles that reproduce observables derived from reference data. DFFR is fast, flexible and can be easily automatized for a high-throughput regime, making it useful in drug design projects. We tested the performance of the method in a few model systems and also in a variety of drug-like molecules using reference data derived from: i) DFT/SCRF (density functional theory coupled to a self-consistent reaction field representation of solvent) calculations on highly populated conformers and ii) enhanced sampling QM/MM where the drug is reproduced at the QM level while solvent is represented by classical force-fields. Extension of the method to include other source of reference data is discussed

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology (BIST). Barcelona

[2] Nostrum Biodiscovery, Nexus II Building. Barcelona

[3] Departament de Bioquímica i Biomedicina. Facultat de Biologia. Universitat de Barcelona.

* Correspondence to Prof. M.Orozco: modesto.orozco@irbbarcelona.org

INTRODUCTION

Force-fields (FFs) are a set of simple mathematical expressions connecting the coordinates of a system with its potential energy [1]–[3]. They are the central element in any classical molecular simulation and their accuracy is crucial to obtain meaningful results. The first FFs for the representation of complex molecules were developed by Lifson's and Allinger's groups [4][5] focusing in the first case on biomacromolecules and in the second in small organic molecules. While Allinger's MMx FF reproduced accurately hydrocarbon-based compounds the flexibility of Lifson's functional make them preferred by the community and in fact, most of current force-fields are based on it; eq. 1:

$$V_{FF} = V_{bond} + V_{angle} + V_{dihedral} + V_{Coulomb} + V_{Lennard-Jones}. \qquad (1)$$

where V_{bond} and V_{angle}stand for stretching and bending (typically represented by harmonic terms), $V_{dihedral}$ is a torsional energy related to the rotation of chemical bonds (captured by a Fourier expansion), $V_{Coulomb}$ stands for electrostatic interactions (typically represented by an atom-centered Coulomb term) and $V_{Lennard-Jones}$ is a van der Waals term (typically represented by a r^{-12}- r^{-6} term).

FF developers quickly addressed their efforts in two different directions: i) the development of potentials able to reproduce small solvents such as water and organic liquids [6], [7] which were needed to perform realistic simulations of any chemical or biochemical system; ii) the development of operational FFs for describing proteins and nucleic acids [8]–[11]. In the first case, FF calibration focused in the non-bonded part and was carried out manually in a trail-and-error process until the refinement of the FF makes possible to obtain Monte Carlo (MC) or molecular dynamics (MD) ensembles reproducing well known experimental observables of the system (ex. vaporization heat, density, compressibility, internal energy, radial distribution functions, dielectric constant,…). In the second case, parametrization was less obvious as the number of parameters to fit for a macromolecule is enormous and the amount of experimental information available is rather limited. To reduce the parametrization problem macromolecular FFs assumed group transferability, but even with this strong assumption, parametrization was difficult due mainly to the lack of reference data. Consequently, a large part of terms in force-fields has been parametrized using low level gas phase quantum mechanical (QM) calculations, which introduces a non-negligible source of error in the FF parameters.

FFs for small liquids have remained unaltered for years [12], [13]; they are reliable and their caveats emerge only when they are used in conditions [14] very different to those considered in their original refinement. Macromolecule FFs do not have the same resiliency to the pass of time, and none of them has survived a decade as the *state of the art* in the field. The reasons are multiple: simplicity of the functional, problems of transferability, lack of reference experimental data and use of too simplistic QM calculations as fitted observables. All these practical problems make the parametrization prone to uncertainties, and after their publication all FFs are scrutinized by the community to detect errors that should be corrected in further revisions. Despite this inefficient evolutionary model, after more than 4 decades of refinements, the last generation of FFs for proteins and nucleic acids, provide results of high quality, in many cases not far from experiments [15]–[19].

Small molecules with unusual functional groups and potential therapeutic action are the ultimate challenge for FF development. Pharmaceutical companies have physical libraries with millions of such molecules and virtual datasets containing a much larger number of synthetically accessible compounds. Functional groups in these molecules can be extremely diverse, making transferability of parameters difficult. Experimental information on these compounds (many of them existing only *in silico*) is null, precluding empirically based parametrization. Finally, systematic QM parametrization exploring all degrees of freedom is simply impossible, as some ligands have a large number of correlated degrees of freedom (for some of our ligands in our BCE database (see previous companion paper) QM calculations on 10^{18} different conformers would be required to systematically explore their conformational space). This situation has forced the community to use of variety of approximated methods [20]–[29], which in general assign parameters based on risky group similarity criteria and very low level QM calculations.

We present here DFFR, a new FF refinement scheme aimed to improve automatic parameter assignment methods by recalibrating those torsions that have the largest impact in determining the conformational space of drug-like small molecule. The method is based on the maximum entropy principle [30] and introduces the minimum changes in a guess force-field needed to reproduce reference data, for example the relative population of representative conformations obtained from DFT/SCRF calculations of ensembles derived from QM/MM molecular dynamics simulation. The procedure is fast, flexible, expandable to incorporate other type of reference data and easy to automatize, i.e. suitable for the high throughput regime.

METHOD DESCRIPTION

DFFR starting point is a MD simulation (ideally an enhanced sampling one) obtained using a guess FF from which rough ensembles are obtained. The ensembles are analyzed to detect those torsions changing the most, i.e. represent the most important degrees of freedom in the molecule. Within the BCE procedure (see previous companion paper) this implies Hamiltonian Replica Exchange (HREX) [31] simulations performed with standard GAFF [21] FF and AM1-BCC charges [32], but the method can be coupled to any other sampling strategy and approximate FF. We present here examples of use of the method taking as reference DFT/SCRF and QM/MM data, but the method can be extended to any reference data including experimental one (details on how to incorporate NMR observables are provided).

Definition of the parametrization space. Even the approach outlined here can be extended to any FF term, we will focus here in the parametrization of the torsional term, as it has the largest impact in determining the conformational preferences of small molecules. In most FFs this implies the parametrization of an energy term defined as shown in eq. 2:

$$V_{dih}(\phi_1,..,\phi_N) = \sum_{j=1}^{N}\sum_{n=1}^{3}\frac{V_{j,n}}{2}\left(1+\cos\left(n\phi_j-\phi_{j,n}\right)\right) = \sum_{j=1}^{N} V_j(\phi_j)$$

where ϕ_j is the value of the dihedral j, $V_{j,n}$ are the amplitudes and $\phi_{j,n}$ are phase angles and the sum extends for all the chemical bonds.

A pure force approach would imply fitting amplitudes (barriers) and phases for all the dihedrals in the molecule, which is impractical even for medium sized molecules. Thus, as noted above, we ignore rigid torsions, i.e. those that are confined in a narrow region of the conformational space and fit only one dihedral per torsion. With these assumptions, we can define an optimal torsional ($\tilde{V}_{dih}$) applying only to a subset (M) of the original dihedral space (eq. 3).

$$\tilde{V}_{dih}(\phi_1,..,\phi_N) = \sum_{j=1}^{M}\sum_{n=1}^{3}\frac{\tilde{V}_{j,n}}{2}\left(1+\cos\left(n\phi_j-\tilde{\phi}_{j,n}\right)\right) + \sum_{j=M+1}^{N}\sum_{n=1}^{3}\frac{V_{j,n}}{2}\left(1+\cos\left(n\phi_j-\phi_{j,n}\right)\right) = \\ = \sum_{j=1}^{M}\tilde{V}_j(\phi_j) + \sum_{j=M+1}^{N} V_j(\phi_j)$$

Modification of the method to other Fourier expansions is straightforward. For a given dihedral, the best set of amplitudes and phases are those minimizing the error function with respect to a reference potential $V_{ref}\left(\overrightarrow{X'},\phi_j\right)$ as shown in eq 4:

$$\varepsilon = \sum_{\phi_j}\left(V_{ref}\left(\overrightarrow{X'},\phi_j\right) - \tilde{V}_{FF}\left(\overrightarrow{X'},\phi_j\right)\right)^2$$

where $\tilde{V}_{FF}(\vec{X},\phi_j) = V_{FF}(\vec{X},\phi_j) - V_j(\phi_j) + \tilde{V}_j(\phi_j)$ and $\overrightarrow{X'} = \{x_1,..,x_R;\phi_1,..,\phi_{j-1},\phi_{j+1},..,\phi_N\}$ refer to all coordinates different to the optimized dihedral. Notice that the sum extends over a certain 2π-periodic interval. As above, values marked with the super-index ˜ refers to optimized FF and those without it to the original set of parameters.

Minimizing eq. 4 for the entire torsional profile is the standard procedure used to refine force-field, but this is impractical for ligands with many rotatable bonds. Thus, as a first approach, we assume here that the general shape of the approximated potential energy function is not too different to the real one, and that the main discrepancies appear mainly in the relative population of the most stables regions. With these assumptions, we can simplify the configurational space as a set of families obtained by clustering of the original trajectory as shown in eq. 5:

$$\langle X \rangle = \sum_k \langle X_k \rangle \sim \sum_k P_k \chi_k$$

where the brackets mean Boltzmann's ensembles, X stands for configurational variables, k are the number of clusters in which the original ensemble is divided, χ_k is the representative structure of cluster k and P_k is the associated probability of the cluster. As shown below, in the unlikely case that we can have QM/MM data as reference simplifications implicit to the use of eq.5 are relaxed and we can use the entire density profile as reference. Note that by extending the sum to a large number of states ($k \to \infty$) the complete torsional density profile is obtained.

Maximum entropy principle and refinement strategy. The objective of the DFFR is to obtain a new set of parameters that make the associated state probabilities match the reference ones i.e. $\{\tilde{P}_k\} = \{P_k^{ref}\}$, perturbing as little as possible the general shape of the conformational space, i.e. introducing the minimum bias in the original FF. This refinement strategy follows the maximum entropy (ME) principle [30] which implies that the best fitting procedure is that reproducing the reference values introducing a minimum external information (i.e. reducing the loss of Shanon's entropy; [33]). As shown by others [34] ME solutions can be derived by imposing Lagrange constrains forcing the system to fulfill some reference values. Thus, it is straightforward to demonstrate that this can be done by minimizing the Lagrange function (eq. 6).

$$\Gamma(\lambda) = \ln\int d\phi \exp\big(-\lambda(\phi)\big)\, P_{FF}(\phi) + \int d\phi \lambda(\phi) P_{ref}(\phi)$$

where $-\lambda(\phi)$ is a continuum function describing the variation of Lagrange constraints with the torsion ϕ, $P_{FF}(\phi)$ is the probability distribution (in the dihedral space) obtained by the approximate FF and $P_{ref}(\phi)$ is a continuum function used as reference.

Imposing minimum in the Lagrange variable we derive eq.7:

$$\frac{\delta\Gamma[\lambda]}{\delta\lambda(\phi)} = \frac{-P_{FF}(\phi)\exp\big(\lambda(\phi)\big)}{A} + P_{ref}(\phi) = 0$$

where A is a constant, and we can conclude that (eq.8):

$$\exp(-\lambda(\phi)) \propto \frac{P_{ref}(\phi)}{P_{FF}(\phi)}$$

which assuming $P_{ref}(\phi) = P_{ME}(\phi)$ is equivalent to Bussi's equation (eq 16 in reference [34]) for a set of reference observables. Rearranging eq. 8 and using Boltzmann's relation we obtain (eq.9):

$$V_{ref}(\phi) = V_{FF}(\phi) + k_B T\lambda(\phi)$$

This means that the new FF, i.e. that reproducing the reference distribution ($P_{ref}(\phi)$) generated by ($V_{ref}(\phi)$) will be that obtained by adding to the original one a continuum biasing equation ($k_B T\lambda(\phi)$).

Biasing the dihedral distributions. In our case we define the continuum biasing function as a Gaussian Mixture (eq. 10)

$$P_{ref}(\phi_j) = \sum_{i=1}^{C} \Omega_i f_i(\phi_j, \mu_i, \sigma_i)$$

where Ω_i are the weights of each component and $f_i(\phi, \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(\phi-\mu_i)^2}{2\sigma_i^2}\right)$ are different Gaussian density functions with centers at μ_i and variances σ_i^2. Notice that $\sum_i \Omega_i = 1$.

Gaussian functions fit well population densities, but assume an energy function defined by the combination of harmonic potentials, which is not practical for MD codes. Thus, we should re-write the biasing potentials with other functions. To this we end, we first use again the Boltzmann's relationship (eq. 11).

$$V_{ref}(\vec{X}, \phi_j) - V_{FF}(\vec{X}, \phi_j) = -k_B T\left(\ln\frac{\sum_{i=1}^{C}\Omega_i f_i(\phi_j, \mu_i, \sigma_i)}{P_{FF}(\phi_j)}\right)$$

Note that here $\lambda(\phi)$ would be exactly $-\ln\frac{P_{ref}(\phi_j)}{P_{FF}(\phi_j)}$, as outlined in eq. 8.

We can now fit the torsional profiles to Fourier expansions (eq. 3) and use eq. 12 to obtain the new set of parameters.

$$\tilde{V}_j(\phi_j) \approx V_j(\phi_j) + V_{ref}(\vec{X},\phi_j) - V_{FF}(\vec{X},\phi_j) = V_j(\phi_j) - k_B T\left(\ln\frac{\sum_{i=1}^{C}\Omega_i f_i(\phi_j,\mu_i,\sigma_i)}{P_{FF}(\phi_j)}\right)$$

Iteration procedure for force-field refinement. The procedure described above provide a modified set of parameters $\{y\}_1$, which yields to a different potential and a different ensemble, eventually even to different clusters and representatives, i.e. $\{y\}_1 \rightarrow \{\tilde{V}_{j,n}, \tilde{\phi}_{j,n}\}_1$. This means that a further refinement might be required by computing new ensembles, new reference values and new biasing function, which in turn will lead to $\{y\}_2 \rightarrow \{\tilde{V}_{j,n}, \tilde{\phi}_{j,n}\}_2$, the process being repeated until convergence. This procedure will be slow and computationally expensive, as it will require many MD simulations and eventually more reference calculations. Fortunately, in most cases we can guess the corrected distribution of dihedrals from the previous one as shown in eq. 13:

$$P_{FF_s}(\phi_j) \propto \exp(-\beta V_{FF_s}) = \exp\left(-\beta V_{FF_{s-1}}\right)\exp(-\beta\Delta V) \propto P_{FF}(\phi_j)\exp(-\beta\Delta V)$$

where s is one iteration of the force-field refinement, and $\Delta V = V_{FF_s} - V_{FF_{s-1}}$. The Iteration is performed until convergence ($P_{FF_s} = P_{FF_{s-1}}$, or alternatively $\{y_s\}$= $\{y_{s-1}\}$) refining FF parametters without the need of re-computing the trajectories. In our experience (see Results below), the procedure fails only in a very few cases, where original parameters are extremely poor, and torsions show strong correlations. For general real applications the method is fast and quite accurate.

Definition of the reference probability ($P_{ref}(\phi_j)$). The method outlined here is in principle general, but practical implementation requires the definition of the reference probability function, which depends on the type of reference data available. We have explored three different sources of reference distributions:

- ***DFT/SCRF calculations***. This is the standard BCE procedure as described in a previous companion paper. It implies to compute the energy of each cluster representative at the DFT level using a continuum description of the solvent. We start by fitting the distributions around free energy minima found in MD simulations to a mixture of Gaussian functions (eq. 10) as shown in eq. 14:

  $$P_{FF}(\phi_j) = \sum_{i=1}^{C}\omega_i f_i(\mu_i,\sigma_i)$$

  where ω_i is the weight of the i-th component and $f_i(\mu_i,\sigma_i)$ is the i-th Gaussian density function with mean μ_i and standard deviation σ_i, as outlined above.

  We refine the geometry of the *k*-cluster representatives at the DFT/SCRF level [35] assuming that QM/SCRF optimization does not drive the molecule to a different cluster (which is a valid assumption in all tested systems). Thus, the weight of the different minima can be easily determined by using eq 15:

$$e_k = \frac{\exp(-\beta E_k)}{\sum_s \exp(-\beta E_s)}$$

where k refers to each cluster representative and energies are always referred to the most stable cluster representative. We then define for each cluster (eq. 16):

$$\Omega_i = \sum_{k \in C_i} e_k$$

where C_i is the set of QM-conformers classified in the i-th component (a structure is classified into the i-th component if μ_i is the closest value to the dihedral value of the structure). Assuming all free energy minima have equal curvature, we can use these weights to define the biasing Gaussians as shown in eq. 17:

$$P_{ref}(\phi_j) = \sum_{i=1}^{C} \Omega_i f_i(\phi_j, \mu_i, \sigma_i)$$

- ***QM/MM calculations***. Last generation MD codes [36]–[39] allow a reasonably fast implementation of QM/MM methods, where solute is represented at the quantum level while solvent is represented by means of classical force-fields. If the progression continues, we can expect a near future, where medium or low-level QM/MM simulations might be used to fit force-fields. In these cases, the Information extracted from QM/MM trajectories could be used in at least two different ways.

  i) Clustering QM/MM ensembles to obtain populations around minima [40], [41] which are fitted to Gaussian functions (as in eq. 14). We then reweight the Gaussian Mixture that represents $P_{FF}(\phi)$ using QM/MM cluster population, so $P_{ref}(\phi)$ can be written as

  $$P_{ref}(\phi) = \sum_{i=1}^{k} p_i \, f_i(\phi_j, \mu_i, \sigma_i)$$

  where

  $$p_i = \sum_{k \in C_i} \frac{P_k}{P_{TOT}}$$

  where C_i is the set of QM/MM-conformers classified in the i-th component, P_k is the cluster k population and P_{TOT} is the total number of snapshots in our trajectory. This procedure is equivalent to the DFT/SCRF one except for the substitution of probability derived by Boltzmann´s inversion of DFT/SCRF free energies to real distributions derived from sampling.

ii) The entire dihedral distribution of the QM/MM trajectory is used as the reference. In this case the Expectation-Minimization algorithm is used to fit the QM/MM data to a Gaussian Mixture in the entire dihedral space, from which a continuum biasing potential is derived and fitted to Fourier potentials as shown above.

- ***Experimental data***. In the daily practice of a drug design laboratory the availability of reference experimental data is rare. However, if available the method can easily incorporate such data into the refinement process. Particularly, ^{3}J-couplings and NOEs which can be translated into dihedral preferences and distance restraints which can be used to re-weight the states sampled in unbiassed simulations in a manner nearly equivalent to the refinement model used taking SCRF/DFT data as reference.

Technical details: Molecules considered here were prepared as described in a previous paper (see previous companion paper) and explored using HREX simulations [31] as implemented in Gromacs 4.6.7 [42] patched with Plumed 2.1 [43]. Simulations were performed in truncated octahedron boxes of TIP3P water molecules [44] (periodic boxes were selected to guarantee a minimum distance greater than 8 Å from the ligand to the closest face). The systems were neutralized by adding suitable ions, minimized, thermalized and equilibrated for 1 ns. We used 16 replicas with the scaling parameter λ going from 1 to 0.59 following a geometrical progression. Individual replicas were followed by 10 ns simulations, taking data every 1 ps. Replica exchanges were attempted every 500 steps and each of the individual trajectories was performed at the NVT ensemble. Particle Mesh Ewald (PME,[45]) and periodic boundary conditions were used to represent long-range electrostatic effects. All bonds linking hydrogens were frozen using SHAKE [46], which allowed us the use of 2 fs time scale for integration of Newton equations of motion. Trajectories were processed to define the most populated clusters using inhouse modified version of advanced clustering method based on dihedral matrix [40]. The geometries of cluster representatives were used as starting point for IEF-MST//B3LYP/6-31G(d) geometry refinement as implemented in Gaussian [47]. IEF-MST calculations were done considering water as solvent and the associated cavity and van der Waals parameters [35].

QM/MM-REMD (Quantum Mechanical/Molecular Mechanics-Replica-Exchange MD) simulations were done for five molecules using the DFTB3 Hamiltonian [48] (QM) and TIP3P water model [13] as implemented in the AMBER program [49]–[51]. 32 replicas were used for each of the molecules, with temperatures ranging from 298 to 498 K. Trajectories were extended for 10 ns for each replica. Results were subjected to the BCE procedure to define flexible bonds and the associated probability density maps.

Trajectories were stored for further analysis using MD database recommendations [52].

RESULTS AND DISCUSSION

Model systems: Small molecules are ideal systems to check for the ability of the method to correct errors in the guess force-field. We choose methanol, butane and alanine dipeptide as prototypic small molecules, for which torsions have been previously parametrized with high accuracy, allowing us to obtain reference distributions by running HREX simulations with "gold-standard" parameters, and in parallel "guess distributions" obtained by using different guess torsional parameters which deviates significantly from the real ones.

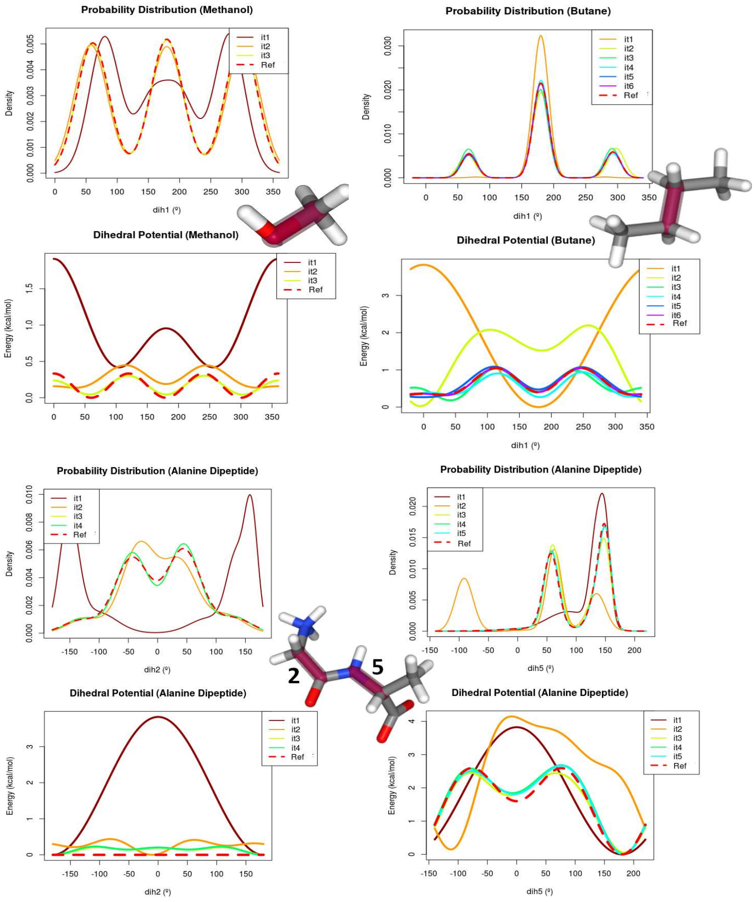


Figure 1. TOP: Density distributions of selected torsions for (left to right) methanol, butane and alanine dipeptide. BOTTOM: torsional potentials leading to the above distributions. Reference torsional potentials and associated distributions are marked as dashed lines. Lines marked as it1 correspond to guess torsional parameters, quite distant in general to the real ones. Lines marked as itn correspond to profiles obtained after n-virtual iterations of the refinement procedure (see above).

We found that, in general, convergence to the real FF parameters is very fast (see Figure 1) even in cases where very incorrect parameters were considered. For example, for methanol we defined a guess set of parameters (iteration_1) that underestimates the stability of the trans 180º minima and moves the position of the -60 and +60 minima, yielding to a very unrealistic distribution of torsion probability, but just three virtual iterations (see eq. 13 above) are enough to converge to the reference parameters (Figure 1). For butane our "guess" torsional parameters have only a V1 term which yields to a single maximum in the torsional probability distribution. As the +60 and -60 minima were not present in the original distribution convergence requires a few more virtual iterations to obtain parameters matching the reference ones (Figure 3). Finally, for alanine dipeptide we define as "guess" torsional parameters a single (and high) V1 term for Ψ and Φ torsions both centered at 0º, which yield to completely incorrect torsional distributions and a Ramachandran's plot where none of the secondary elements are present (Figure 1). The method requires only 4-5 iterations to converge to the reference values and provide correct mono and bi-dimensional probability maps. In summary, DFFR seems able to refine FF even in those cases where the guess values are inaccurate.

The DFT/SCRF refinement: We tested the power of the default refinement method using DFT/SCRF energy values as reference for 13 small drug-like molecules, all of them present in our BCE database and for which the bioactive conformation is known (see Suppl. Figure S1). Based on the default analysis of torsional space in BCE we refine from 1 to 3 torsions for each ligand. The refinement procedure leads to representation of the torsional space that agree much better with the reference QM data, as shown (for selected molecules) in Figure 2, even for those cases where original GAFF values leads to quite incorrect samplings (see for example dihedral 8 in panel B or dihedral 5 in panel F of Figure 2). A global analysis of the differences between QM torsional density (inferred from the minima free energy; see above) and the MM torsional density profiles demonstrates a massive improvement in the density profile errors (Figure 3) with respect to the results obtained using the default force-field GAFF. The difference in the stability of the best optimized geometry is however small (Figure 3), indicating that the QM optimization procedure is quite robust to the geometry of the cluster representative derived from classical simulations. In fact, inspection of individual profiles in Figure 2 strongly suggests that for most purposes DFT/SCRF and refined FF results are hardly distinguishable, allowing the user to perform classical calculations with a quality similar to that obtained at the DFT/SCRF -level of accuracy.

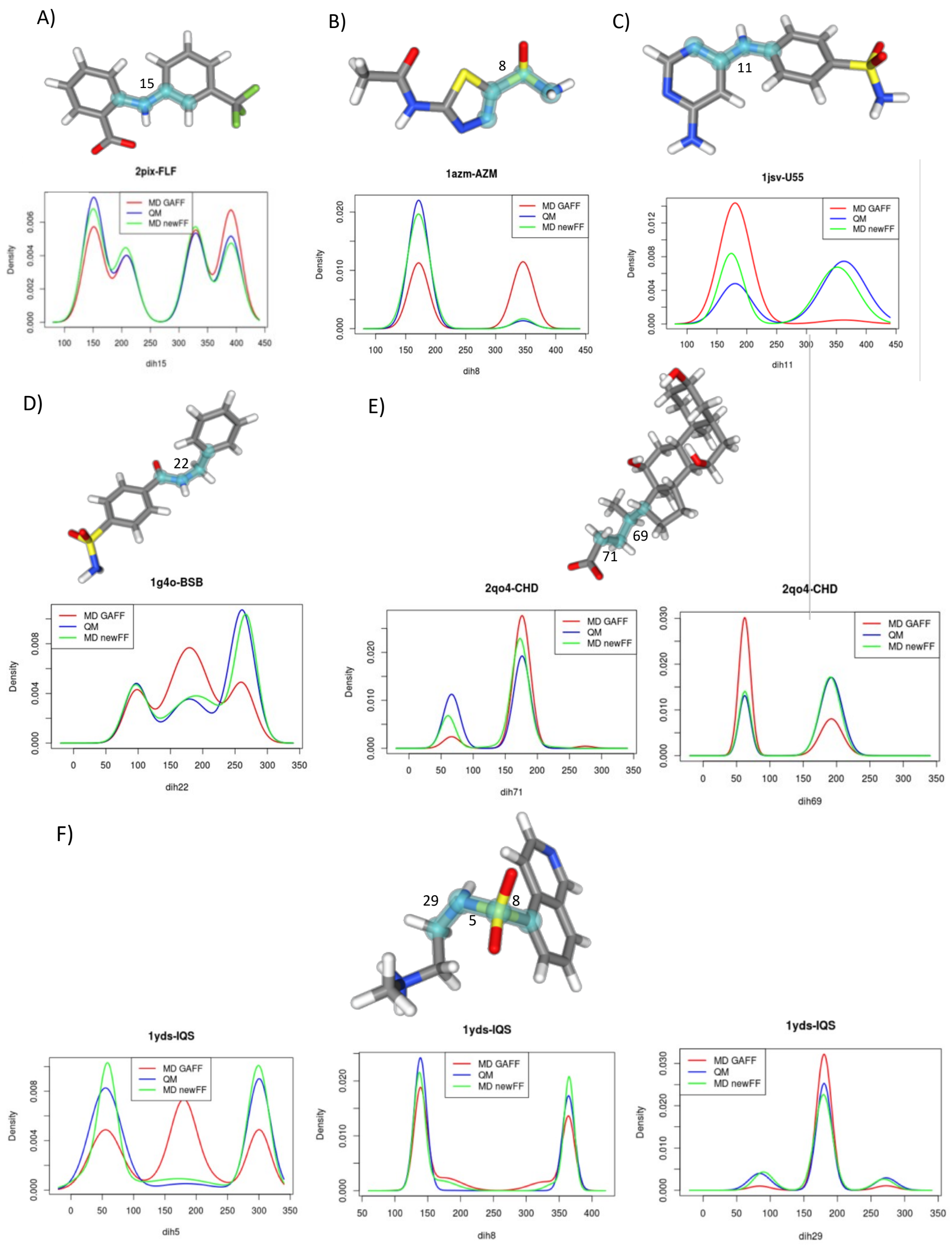


Figure 2. Selected density distributions obtained by the default FF (GAFF in red), the reference DFT/SCRF data (obtained by Gaussian fitting of the relative population of the fitting; in blue) and the refined FF (in green). Note in all cases the much closer similarity of blue and green profiles compared with the red one. Values shown here correspond typically to 3rd iteration refinement (in most cases functionals are already converged after 2 iterations). Representation of the ligands considered are shown in Figure S1.

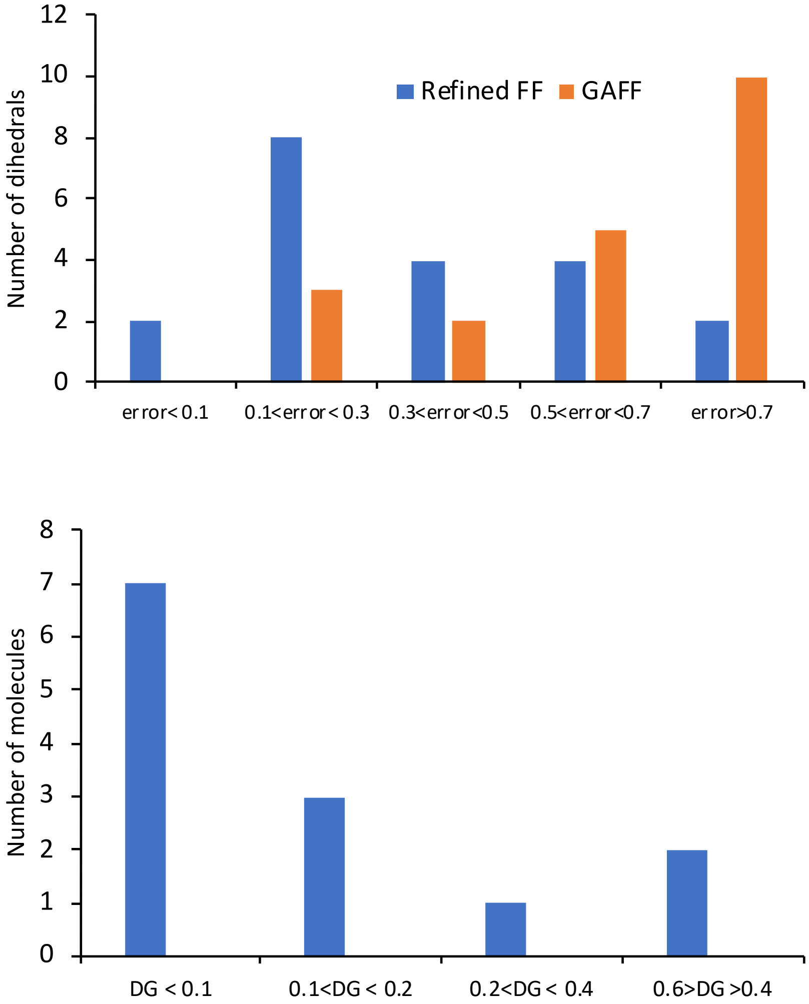
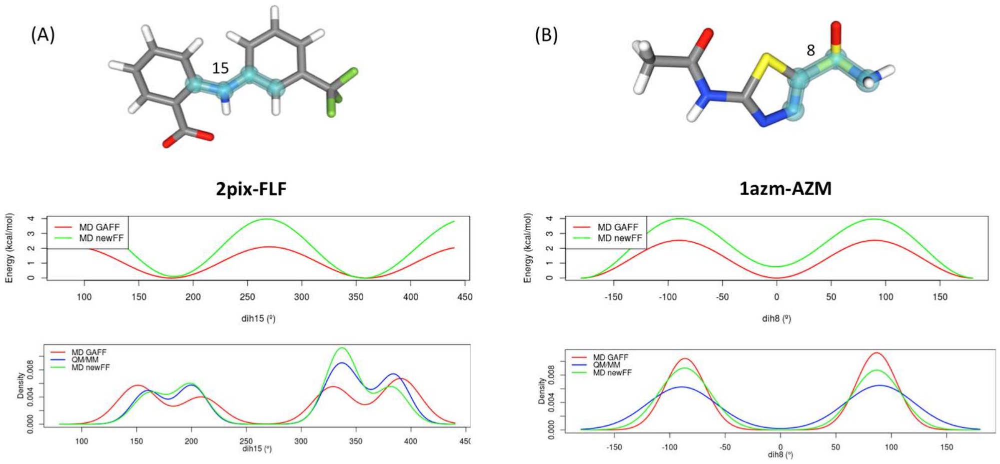


Figure 3. TOP panel: histogram of errors in the profile (refined FF or GAFF) with respect to DFT/SCRF values. Errors were determined as $error = \sqrt{\sum(\rho(DFT/SCRF) - DFT(FF))^2}$ computed for the entire 0→360 degrees every degree. BOTTOM: Difference in the free energy minima at the DFT/SCRF level obtained after optimization of GAFF and refined FF representatives.

The QM/MM refinement: The development of fast QM/MM codes taking advantage of GPUs will make possible in a future to refine some crucial dihedrals taking directly QM/MM data [53] eliminating then the intrinsic shortcomings of the incomplete sampling and the continuum-related artifacts implicit to DFT/SCRF calculations. The QM-based parametrization has the additional advantage that it can use direct torsional population profiles rather than Gaussian-fitted estimates obtained from the analysis of the relative stability of DFT/SCRF free energy minima. The obvious shortcoming of this approach is the large computational cost which limits today (perhaps not in a near future) its general application in the context of drug-design pipelines.

To test the applicability of QM/MM calculations within the DFT/SCRF framework we tested the method in six prototypical drugs for which QM/MM calculations were performed at the DFTB3 level of theory (see Methods). Results in Figures 4A and B were obtained for cases where GAFF default torsional parameters are reasonable and accordingly no dramatic changes in the torsional density profiles are found after recalibration of the FF. On the contrary, Figure 4F represent one of the cases where very significant changes in the torsional energy was required to reproduce the QM/MM torsional density profile. In all cases the improvement in the torsional profiles is evident, indicating the robustness of the process to the source of data used for recalibration of the FF. As computational power increases and QM/MM calculations can be done at higher levels of QM theory we can envision a full parametrization procedure relying on QM/MM simulations for hits appearing promising in drug design pipelines.



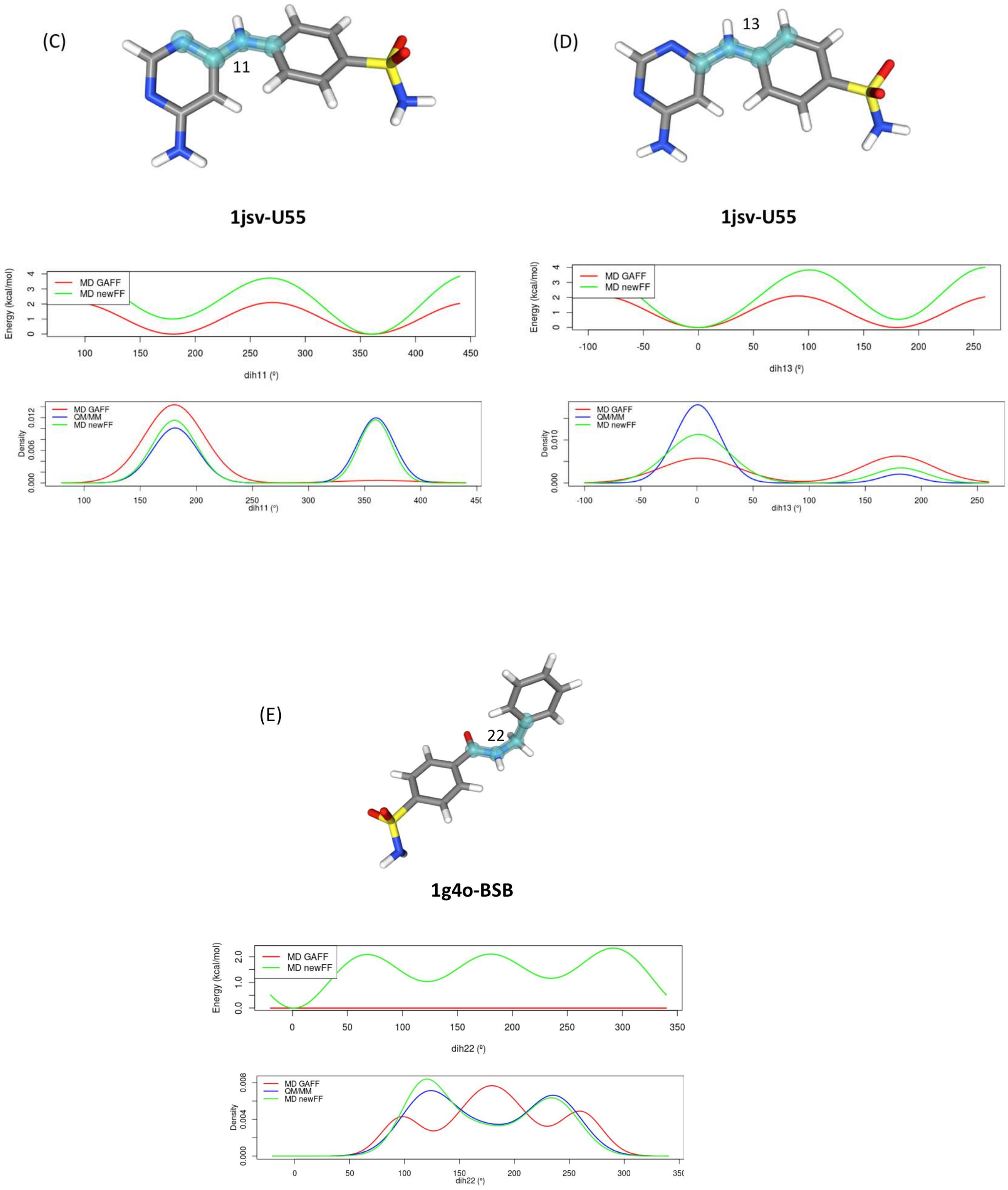


Figure 4. TOP: Potential energy profiles obtained from GAFF and our refined FF for selected dihedrals and molecules. BOTTOM: Density distributions around marked dihedral obtained from QM/MM, GAFF and the refined FF for the same set of dihedrals and molecules.

CONCLUSIONS

We present a fast, robust and automatized method for recalibration of FF based on reference distributions obtained from reference calculations obtained at the QM level. The method is general and can incorporate experimental data, for example relative state populations derived from J-coupling or NOEs, which are not difficult to obtain in a short time scale. The method is fast and can be implemented in the high throughput regime, which facilitates its application by pharma industry for the refinement of parameters assigned by means of low-level automatic FF annotators. DFFR is highly recommended for hits and lead compounds that can be used as starting point for an optimization process and that are likely be used in massive docking experiments or as a template in similarity-powered studies. The method will help to discard compounds unable to reach the bioactive conformation and will provide estimates of the distortion energy required to adopt the bioactive state, enriching then scoring functions used in most structure-based drug design protocols.

ACKNOWLEDGMENTS

We are indebted to the support of the Spanish Ministry of Science [RTI2018-096704-B-100], the Catalan SGR, the Instituto Nacional de Bioinformática; European Union's Horizon 2020 research and innovation program [BioExcel-2 project], Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) (all awarded to M.O.) as well as the MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona) and the CDTI (Neotec grant−EXP 00094141/SNEO-20161127) (awarded to Nostrum Biodiscovery). NDB is sponsored by the Fundación Botín (Mind the Gap Program). We are indebted to the support of BSC. MO is an ICREA Research Professor.

REFERENCES

[1] T. Schlick, *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*, vol. 21. Springer Science & Business Media, 2010.

[2] M. Levitt, "The birth of computational structural biology," *Nat. Struct. Biol.*, vol. 8, no. 5, pp. 392–393, 2001.

[3] A. R. Leach and A. R. Leach, *Molecular modelling: principles and applications*. Pearson education, 2001.

[4] N. L. Allinger, M. A. Miller, F. A. Van Catledge, and J. A. Hirsch, "Conformational analysis. LVII. The calculation of the conformational structures of hydrocarbons by the Westheimer-Hendrickson-Wiberg method," *J. Am. Chem. Soc.*, vol. 89, no. 17, pp. 4345–4357, 1967.

[5] S. Lifson and A. Warshel, "Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules," *J. Chem. Phys.*, vol. 49, no. 11, pp. 5116–5129, 1968.

[6] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, and R. W. Impey, "Michael L Klein," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[7] W. L. Jorgensen and J. Tirado-Rives, "Potential energy functions for atomic-level simulations of water and organic and biomolecular systems," *Proc. Natl. Acad. Sci.*, vol. 102, no. 19, pp. 6665–6670, 2005.

[8] A. Warshel, M. Levitt, and S. Lifson, "Consistent force field for calculation of vibrational spectra and conformations of some amides and lactam rings," *J. Mol. Spectrosc.*, vol. 33, no. 1, pp. 84–99, 1970.

[9] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a Swaminathan, and M. Karplus, "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.*, vol. 4, no. 2, pp. 187–217, 1983.

[10] P. K. Weiner and P. A. Kollman, "AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions," *J. Comput. Chem.*, vol. 2, no. 3, pp. 287–303, 1981.

[11] W. F. van Gunsteren and H. J. C. Berendsen, "Groningen molecular simulation (GROMOS) library manual," *Biomos, Groningen*, vol. 24, no. 682704, p. 13, 1987.

[12] W. L. Jorgensen, J. D. Madura, and C. J. Swenson, "Optimized intermolecular potential functions for liquid hydrocarbons," *J. Am. Chem. Soc.*, vol. 106, no. 22, pp. 6638–6646, 1984.

[13] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[14] A. Perez and M. Orozco, “Real-time atomistic description of DNA unfolding,” *Angew. Chemie Int. Ed.*, vol. 49, no. 28, pp. 4805–4808, 2010.

[15] M. Orozco, “A theoretical view of protein dynamics,” *Chem. Soc. Rev.*, vol. 43, no. 14, pp. 5051–5066, 2014.

[16] A. Pérez, F. J. Luque, and M. Orozco, “Frontiers in molecular dynamics simulations of DNA,” *Acc. Chem. Res.*, vol. 45, no. 2, pp. 196–205, 2012.

[17] P. D. Dans, J. Walther, H. Gómez, and M. Orozco, “Multiscale simulation of DNA,” *Curr. Opin. Struct. Biol.*, vol. 37, pp. 29–45, 2016.

[18] J. Sponer *et al.*, “RNA structural dynamics as captured by molecular simulations: a comprehensive overview,” *Chem. Rev.*, vol. 118, no. 8, pp. 4177–4338, 2018.

[19] P. D. Dans, D. Gallego, A. Balaceanu, L. Darré, H. Gómez, and M. Orozco, “Modeling, simulations, and bioinformatics at the service of rna structure,” *Chem*, vol. 5, no. 1, pp. 51–73, 2019.

[20] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, “Automatic atom type and bond type perception in molecular mechanical calculations,” *J. Mol. Graph. Model.*, vol. 25, no. 2, pp. 247–260, 2006.

[21] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, 2004.

[22] A. W. S. da Silva and W. F. Vranken, “ACPYPE-Antechamber python parser interface,” *BMC Res. Notes*, vol. 5, no. 1, p. 367, 2012.

[23] K. Vanommeslaeghe and A. D. MacKerell Jr, “Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing,” *J. Chem. Inf. Model.*, vol. 52, no. 12, pp. 3144–3154, 2012.

[24] K. Vanommeslaeghe, E. P. Raman, and A. D. MacKerell Jr, “Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges,” *J. Chem. Inf. Model.*, vol. 52, no. 12, pp. 3155–3168, 2012.

[25] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen, “LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands,” *Nucleic Acids Res.*, vol. 45, no. W1, pp. W331–W336, 2017.

[26] V. Zoete, M. A. Cuendet, A. Grosdidier, and O. Michielin, “SwissParam: a fast force field generation tool for small organic molecules,” *J. Comput. Chem.*, vol. 32, no. 11, pp. 2359–2368, 2011.

[27] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J. C. Gumbart, “Rapid parameterization of small molecules using the force field toolkit,” *J. Comput. Chem.*, vol. 34, no. 32, pp. 2757–2770, 2013.

[28] P. Procacci, “Primadorac: A free web interface for the assignment of partial charges,

chemical topology, and bonded parameters in organic or drug molecules," *J. Chem. Inf. Model.*, vol. 57, no. 6, pp. 1240–1245, 2017.

[29] C. Alemán and M. Orozco, "PAPQMD/AM1 Parametrization of the bonded term of aromatic biomolecules," *Biopolym. Orig. Res. Biomol.*, vol. 34, no. 7, pp. 941–955, 1994.

[30] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, p. 620, 1957.

[31] H. Fukunishi, O. Watanabe, and S. Takada, "On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction," *J. Chem. Phys.*, vol. 116, no. 20, pp. 9058–9067, 2002.

[32] A. Jakalian, D. B. Jack, and C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation," *J. Comput. Chem.*, vol. 23, no. 16, pp. 1623–1641, 2002.

[33] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[34] A. Cesari, S. Reißer, and G. Bussi, "Using the maximum entropy principle to combine simulations and solution experiments," *Computation*, vol. 6, no. 1, p. 15, 2018.

[35] I. Soteras, C. Curutchet, A. Bidon-Chanal, M. Orozco, and F. J. Luque, "Extension of the MST model to the IEF formalism: HF and B3LYP parametrizations," *J. Mol. Struct. THEOCHEM*, vol. 727, no. 1–3, pp. 29–40, 2005.

[36] S. Pronk *et al.*, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, no. 7, pp. 845–854, 2013.

[37] D. A. Case *et al.*, "AMBER 14, 2014," *Univ. California, San Fr.*, 2014.

[38] B. R. Brooks *et al.*, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, 2009.

[39] J. C. Phillips *et al.*, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005.

[40] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science (80- .).*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[41] M. Rosa, M. Micciarelli, A. Laio, and S. Baroni, "Sampling Molecular Conformers in Solution with Quantum Mechanical Accuracy at a Nearly Molecular-Mechanics Cost," *J. Chem. Theory Comput.*, vol. 12, no. 9, pp. 4385–4389, 2016.

[42] M. J. Abraham, D. Van Der Spoel, E. Lindahl, and B. Hess, "the GROMACS development team," *GROMACS user Man. version*, vol. 5, no. 2, pp. 1–298, 2014.

[43] G. Bussi, "Hamiltonian replica exchange in GROMACS: a flexible implementation," *Mol. Phys.*, vol. 112, no. 3–4, pp. 379–384, 2014.

[44] D. J. Price and C. L. Brooks III, "A modified TIP3P water potential for simulation with Ewald

summation," *J. Chem. Phys.*, vol. 121, no. 20, pp. 10096–10103, 2004.

[45] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[46] V. Kräutler, W. F. Van Gunsteren, and P. H. Hünenberger, "A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations," *J. Comput. Chem.*, vol. 22, no. 5, pp. 501–508, 2001.

[47] M. J. Frisch *et al.*, "Gaussian 16." Gaussian, Inc. Wallingford, CT, 2016.

[48] M. Gaus, Q. Cui, and M. Elstner, "DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB)," *J. Chem. Theory Comput.*, vol. 7, no. 4, pp. 931–948, 2011.

[49] A. W. Götz, M. A. Clark, and R. C. Walker, "An extensible interface for QM/MM molecular dynamics simulations with AMBER," *J. Comput. Chem.*, vol. 35, no. 2, pp. 95–108, 2014.

[50] R. C. Walker, M. F. Crowley, and D. A. Case, "The implementation of a fast and accurate QM/MM potential method in Amber," *J. Comput. Chem.*, vol. 29, no. 7, pp. 1019–1031, 2008.

[51] G. de M. Seabra, R. C. Walker, M. Elstner, D. A. Case, and A. E. Roitberg, "Implementation of the SCC-DFTB method for hybrid QM/MM simulations within the Amber molecular dynamics package," *J. Phys. Chem. A*, vol. 111, no. 26, pp. 5655–5664, 2007.

[52] A. Hospital, F. Battistini, R. Soliva, J. L. Gelpí, and M. Orozco, "Surviving the deluge of biosimulation data," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, p. e1449.

[53] L. Darré, I. Ivani, P. D. Dans, H. Gómez, A. Hospital, and M. Orozco, "Small details matter: the 2′-hydroxyl as a conformational switch in RNA," *J. Am. Chem. Soc.*, vol. 138, no. 50, pp. 16355–16363, 2016.