ParaMol: A Package for Parametrization of Molecular Mechanics Force Fields

João Morado,[†] Paul N. Mortenson,[‡] Marcel L. Verdonk,[‡] Richard A. Ward,[¶] Jonathan W. Essex,[*,†] and Chris-Kriton Skylaris[*,†]

[†]*School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom*

[‡]*Astex Pharmaceuticals, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, United Kingdom*

[¶]*Medicinal Chemistry, Oncology RD, AstraZeneca, Cambridge CB4 0WG, UK*

E-mail: j.w.essex@soton.ac.uk; c.skylaris@soton.ac.uk

Abstract

The ensemble of structures generated by molecular mechanics (MM) simulations is determined by the functional form of the force field employed and its parametrization. For a given functional form, the quality of the parametrization is crucial and will determine how accurately we can compute observable properties from simulations. Whilst accurate force field parametrizations are available for biomolecules, such as proteins or DNA, the parametrization of new molecules, such as drug candidates, is particularly challenging as these may involve functional groups and interactions for which accurate parameters may not be available. Here, in an effort to address this problem, we present ParaMol, a Python package that has a special focus on the parametrization of bonded and non-bonded terms of drug-like molecules by fitting to *ab initio* data. We demonstrate the software by deriving bonded terms' parameters of three widely-known drug molecules: aspirin, caffeine, and a norfloxacin analog, for which we show

that, within the constraints of the functional form, the methodologies implemented in ParaMol are able to derive near-ideal parameters. Additionally, we illustrate the best practices to follow when employing specific parametrization routes; the sensitivity of different fitting data sets, such as relaxed dihedral scans and configurational ensembles, to the parametrization procedure; and the features of the various weighting methods available to weight configurations. Owing to ParaMol's capabilities, we propose that this software can be introduced as a routine step in the protocol normally employed to parametrize drug-like molecules for MM simulations.

Introduction

Molecular mechanics-based (MM) simulation methods such as molecular dynamics (MD) and Monte Carlo (MC) are commonly employed to solve many problems in chemistry, physics, biochemistry, and condensed-matter.[1] The ability of these MM-based methodologies to correctly model systems of interest relies mainly on two aspects: their capacity to extensively sample the configurational space and the accuracy of the underlying force field (FF).

The sampling problem is still an area of intensive research, with many enhanced sampling methods being proposed in the past decades, *e.g.*, metadynamics,[2,3] Hamiltonian replica-exchange,[4–6] and umbrella sampling.[7] On the other hand, the accuracy of MM simulations relies on the underlying FF, which comprises a functional form and a set of parameters. The functional form consists of a function that defines the potential energy of the system and allows calculation of forces, which enables equations of motion to be numerically solved. Amongst the most commonly used fixed-charge FF functional forms are AMBER,[8,9] GROMOS,[10] CHARMM[11] and OPLS.[12–15] These FFs already contain extensive databases of parameters for different types of molecules. Even so, many applications require the parametrization of novel molecules or levels of accuracy in the conformations and energetics that are unattainable using default FF parameters. Of great importance is the parametrization of molecules for drug-design applications and for the calculation of quantum corrections to

classical free energies, which were shown to converge faster if MM descriptions more similar to the quantum level are employed.[16–18] Here, in an effort to address the problem of FF accuracy, we present ParaMol, a software package that is capable of deriving bespoke FF parameters in an automated fashion by fitting to *ab initio* data.

Different software packages have already been released for the purpose of automatic FF parametrization. Each has its own features, specific methods and design choices. For example, Paramfit[19] is capable of parametrizing the bonded parameters in the AMBER equation by fitting to *ab initio* forces and energies; ffTK[20] (VMD plugin) and GAAMP[21] were designed specifically to develop CHARMM-compatible parameters for small molecules and permit the parametrization of charges and bonded parameters; the CPMD software package[22] also contains a QM/MM force-matching implementation and can derive charges and bonded terms parameters for the AMBER and GROMOS96 equations; Schrödinger's proprietary software is capable of parametrizing the OPLS FF and systematically generating missing torsional parameters;[12,14,15,23] finally, ForceBalance,[24,25] stands out due to its generality - it is capable of parametrizing different FF functional forms to experimental data and has many optimization algorithms available.

ParaMol can be used both as a stand-alone package or as a Python package to create user-customized parametrization protocols. It differs from other parametrization software packages in some of its implementation choices and in its special focus on the parametrization of drug-like molecules from first-principles quantum mechanics. ParaMol aims to ease all steps in a standard parametrization workflow, *i.e.*, it automates configurational sampling, the calculation of reference data, and the procedure of obtaining the optimal FF parameters. Therefore, it can be easily introduced as a routine step in the standard workflow used to prepare drug-like ligands for MM simulations. It can also be extended to accommodate new objective functions, fitting properties and FF functional forms. It also has parallel capabilities that allow distributing the calculation of the objective function and *ab initio* training data amongst the available computational resources. Currently, the package

is able to derive parameters of class I additive potential energy functions, such as the ones used by AMBER, CHARMM, and OPLS FFs.[26] Class I FFs use harmonic functions to describe stretching and bending, no cross-terms, and Lennard-Jones for dispersion interactions. They are often adequate to describe the structures and non-bonded energies of (bio-)organic molecules.[8,27] Class II FFs add cubic or quartic terms to describe stretching and bending, as well as cross-terms to reproduce the coupling between degrees of freedom (DOFs) (*e.g.*, MM3[28–30] and CFF[31,32]). Finally, class III FFs include polarization and hyperconjugation (*e.g.* AMOEBA,[33,34] polarizable CHARMM[35,36] or fluc-q[37–39])

As an application example we assessed the limits of accuracy that can be attained by fitting bonded parameters of the GAFF functional form to QM calculations. For this purpose, we chose three widely-known drug molecules: aspirin, caffeine, and a norfloxacin analog. To illustrate the dihedral scan functionality that is available in ParaMol, we optimized the dihedral parameters associated with the main rotatable bond of a norfloxacin analog; furthermore, for aspirin, we explored the advantages and limitations of the use of dihedral scans against the generation of configurational ensembles through molecular dynamics, as ways of exploring the potential energy surface (PES), and optimized its intramolecular bond, angle, and dihedral parameters; finally, we employed adaptive parametrization to derive new bonded parameters for caffeine.

This paper is structured as follows: we first present the basic theory underlying the implementation of the ParaMol package, *viz.*, the generalization of the force-matching method,[40] the restrained electrostatic potential (RESP) model,[41–43] the optimization algorithms available, as well as some remarks about regularization and parameter preconditioning; then we describe the organization of the software package and its functionalities; finally, we conclude by presenting the application of different parametrization protocols to the previously mentioned test cases.

Theory and methods

Generalization of the force-matching method

A generalization of the original force-matching method[40] can be formulated in which, instead of only fitting forces, the aim is to fit the FF to reproduce - within the constraints of the functional form - any desired experimental or theoretical property of interest. In this context, the optimization procedure can be seen as a mathematical problem in the space of FF parameters, here denoted as $\boldsymbol{p}$, where $\boldsymbol{p}$ is a vector containing all optimizable parameters. The aim of the optimization is to determine the optimal set of parameters that minimize an objective function, here denoted as X. The objective function contains the squares of the residuals, and in its general form reads:

$$X(\boldsymbol{p}) = X_F(\boldsymbol{p}) + \sum_{\{A\}} X_A(\boldsymbol{p}) + \Theta(\boldsymbol{p}) \tag{1}$$

where X_F corresponds to the term of the objective function in which MM forces are fitted to reference values, X_A amounts for the fitting of any other property of interest A to reference data (*e.g.*, potential energy, electrostatic potential), and $\Theta(\boldsymbol{p})$ is a regularization term that can be optionally included in order to prevent over-fitting (discussed in detail in a subsequent section). Specifically, two different types of force-matching terms, X_F^I and X_F^{II}, are implemented in ParaMol. The type I force-matching term fits the norm of the atomic forces to reference data and it has the following form:[40]

$$X_F^I(\boldsymbol{p}) = \frac{1}{3N_a} \sum_i^{N_s} \omega_i \sum_j^{N_a} \frac{|\Delta \boldsymbol{F}_{i,j}|^2}{\mathrm{Var}(\boldsymbol{F}^{ref})} \tag{2}$$

where $\Delta \boldsymbol{F}_{i,j} = \boldsymbol{F}_{i,j}^{MM}(\boldsymbol{p}) - \boldsymbol{F}_{i,j}^{ref}$, $\boldsymbol{F}_{ij}^{ref}$ and $\boldsymbol{F}_{ij}^{MM}$ are the QM (reference) and MM force vectors, respectively, of atom j in conformation i, ω_i is the weight of the i-th conformation, N_s is the number of structures provided and N_a the number of atoms of the system. The type II force-matching term fits every component of the atomic forces to reference data and

it is given by:[25]

$$X_F^{II}(\boldsymbol{p}) = \frac{1}{3N_a}\sum_i^{N_s}\omega_i\sum_j^{N_a}\left[\Delta\boldsymbol{F}_{i,j}(\boldsymbol{p})^T\langle\boldsymbol{F}_{i,j}^{ref}\otimes\boldsymbol{F}_{i,j}^{ref}\rangle^{-1}\Delta\boldsymbol{F}_{i,j}(\boldsymbol{p})\right] \tag{3}$$

It is worth noting that the variance, $\mathrm{Var}(\boldsymbol{F}^{ref})$, and covariance, $\langle\boldsymbol{F}_{i,j}^{ref}\otimes\boldsymbol{F}_{i,j}^{ref}\rangle$, are used in X_F^I and X_F^{II}, respectively, so that the residuals in the objective function are dimensionless and maximally of unit magnitude. Furthermore, in equation (1), X_A is a general expression for the fitting of any property of interest A to reference data, which for the case of a global property is given by:

$$X_A(\boldsymbol{p}) = \sum_i^{N_s}\omega_i\frac{\left(A_i^{MM}(\boldsymbol{p}) - A_i^{ref}\right)^2}{\mathrm{Var}(A^{ref})} \tag{4}$$

Furthermore, similarly to what was done in equations (2) and (3), if A is an atom-based property the appropriate sum over all atoms and normalization constant have to be introduced. It is worth mentioning that a special case of equation (4) is considered when the property to be fitted is the energy, *i.e.*, when $A = E$. In this case, since different levels of theory have different energy references (*e.g.*, the QM and MM energies usually differ by several orders of magnitude) the expression used for X_E reads:

$$X_E(\boldsymbol{p}) = \sum_i^{N_s}\omega_i\frac{\left(E_i^{MM}(\boldsymbol{p}) - E_i^{ref} - \langle\Delta E\rangle\right)^2}{\mathrm{Var}(E^{ref})} \tag{5}$$

where E_i^{ref} and E_i^{MM} are the QM (reference) and MM potential energies, and $\langle\Delta E\rangle = \frac{1}{N_s}\sum_i\left(E_i^{ref} - E_i^{MM}\right)$ is a term that brings the two distributions together by subtracting the average difference between the reference and MM energies from the energy residuals.

Data set generation: dihedral scans and configurational ensembles

Regarding the schemes through which reference data sets of configurations and respective properties of interest may be generated for parametrization purposes, two methods are rou-

tinely employed to explore the PES of small organic molecules: dihedrals scans or configurational ensembles obtained through standard MM simulation methods.

The most common method to explore the PES is to perform 1-dimensional relaxed scans of each DOF of interest (*e.g.* dihedrals), wherein only the DOFs not explicitly being constrained are allowed to relax (by default all the DOFs not being scanned). There are mainly two disadvantages associated with this methodology: first, the energy can change dramatically if a substituent group falls into a different molecular configuration due to concerted motions, which causes discontinuities in the energy profiles; second, if there are non-negligible couplings between DOFs, *i.e.*, if the DOFs are not orthogonal, then the full potential will not be correctly described by a 1-dimensional surface, demanding higher-dimensional scans that quickly become prohibitive.[44,45]

Alternatively to the use of dihedral scans, it is also possible to use either MD or MC simulations to generate ensembles of configurations. Whilst the disadvantages of the relaxed scans are not present in this case, this procedure usually requires sufficiently long simulations that ensure exploration of the relevant parts of the PES, which may become computationally expensive to perform if high levels of theory are employed, or specific techniques to force sufficient coverage of sampling (*e.g.*, replica exchange algorithms[4–6]).

Dihedral fitting approaches

Even though the derivation of dihedral angle parameters may be performed using conformational ensemble data sets, computationally-speaking it is often less costly and more convenient to use data sets obtained from dihedral scans. We have implemented in ParaMol two different dihedral fitting approaches using dihedral scan data sets, which will be described next. For this purpose, we use the notation $E^{A,B}$, where A and B refer to the levels of theory used to calculate the single-point energies and to perform the geometry optimizations, respectively.

A commonly employed approach,[19,21,46] hereinafter referred to as QM-relaxed, is to derive

the dihedral angle parameters by determining the difference between the MM single-point energy ($E^{MM,ref}$) and QM single-point energy ($E^{ref,ref}$), obtained in vacuum and using the same QM geometry for both the MM and QM calculations. In this case, the objective function reads:

$$X_{dih}(\boldsymbol{p}) = \sum_{i}^{N_s} \omega_i \frac{\left(E_i^{MM,ref}(\boldsymbol{p}) - E_i^{ref,ref} - \langle \Delta E \rangle\right)^2}{\mathrm{Var}(E^{ref,ref})} \tag{6}$$

Nevertheless, as pointed out by other authors,[20,47,48] an approach that is often under-appreciated and that yields more adequate parameters, hereinafter referred to as MM-relaxed, may be obtained when a further MM optimization (with the proper constraints) is also carried out for every dihedral scan conformation. Therefore, since in this case the MM single point-energy ($E^{MM,MM}$) is calculated based on the MM-relaxed geometry rather than the QM-relaxed geometry, the objective function reads:

$$X_{dih,relaxed}(\boldsymbol{p}) = \sum_{i}^{N_s} \omega_i \frac{\left(E_i^{MM,MM}(\boldsymbol{p}) - E_i^{ref,ref} - \langle \Delta E \rangle\right)^2}{\mathrm{Var}(E^{ref,ref})} \tag{7}$$

The rationale underlying the MM-relaxed approach is that the MM energy is highly influenced by the intra-molecular energy terms associated with parameters that are not being optimized, which may come either from the non-bonded terms (van der Waals, Couloumb and 1-4 interaction terms) or the bonded terms (bond, angle and dihedral terms) of the FF. Therefore, QM optimizations may lead to geometries that are deformed from the point of view of the MM level of theory, which stems from the fact that the MM FF parameters may have been obtained by fitting to experimental data or QM levels of theory that are different from the one we are attempting to reproduce. As a consequence, the MM and QM dihedral profiles may present significant differences regarding the relative energies of the regions of the PES that are of primary interest for proper modeling, such as minima and transition states. Hence, by using QM geometries that acquire MM energies (QM-relaxed approach) instead of MM geometries that acquire MM energies (MM-relaxed approach), we may substantially

bias the parametrization procedure by attempting to correct for differences in the dihedral profiles that are unrelated to the FF terms associated with the dihedral(s) being scanned. Interestingly, the MM-relaxed approach can also be used to take into account more complex relaxation situations such as, *e.g.*, the environment-related effects that occur in solution or protein environment.[48]

Overall, it is advisable to employ the MM-relaxed approach as long as the resultant MM-optimized geometries do not significantly differ from the QM-optimized ones. As a rule of thumb, we may consider that if the global conformational preferences of the molecule have not changed after the MM optimization, then the MM-relaxed approach is preferred. Finally, it is also worth mentioning that the QM-relaxed approach is a good approximation whenever the DOFs not being scanned match in the QM and MM optimized geometries, or whenever the remaining FF terms do not contribute significantly to the dihedral profile. This concern is particularly important for hard DOFs, such as bonds and angles, which due to their large force constants, small differences in value lead to large changes in energy. Therefore, it is recommended to relax those before deriving dihedral parameters, as otherwise we may incur in biased optimizations.

Weighting methods

We have implemented a variety of weighting methods in ParaMol that give more importance to some conformations than others. These weighting methods must be applied with care so as to balance the effect of the increased weighting of some conformations on the energies of other conformations. Currently, the weighting methods available in ParaMol are the following:

- **Uniform weighting**: this is the simplest weighting method that is possible to apply. It attributes equal weight to all conformations, such that for any two conformations i and k we have:

$$\omega_i = \omega_k = \frac{1}{N_s} \tag{8}$$

This weighting method may be problematic if very high-energy conformations are present because, in order to minimize the errors in their description, the fitting procedure may adversely affect the description of highly-populated low energy conformations. This usually happens due to constraints of the functional form. A practical solution for this problem is to use ParaMol to prune out from the optimization conformations for which the reference energy is larger than a given value relative to the minimum energy conformation (*e.g.*, 10.0 kcal/mol).

- **Boltzmann weighting**: Boltzmann weighting based on the reference (QM) energies gives more importance to low-energy conformations than to high-energy ones. This non-uniform weighting is achieved by weighting each conformation by the factor:

$$\omega_i = \frac{\exp\left[-\beta(E_i^{QM} - < E^{QM} >)\right]}{\sum_j^N \exp\left[-\beta(E_j^{QM} - < E^{QM} >)\right]} \tag{9}$$

The disadvantage of Boltzmann weighting is that it usually leads to inaccurate energies for conformations located at or near high-energy barriers. This often compromises the dynamics of the model, preventing its use in standard simulation methods.[45]

- **Non-Boltzmann weighting**: The non-Boltzmann weighting method implemented in ParaMol is the one proposed by Wang *et al*,[49] for which the expression used for the i-th conformation's weight is given by:

$$\omega_i = \frac{\exp\left[-\beta(E_i^{MM} - E_i^{QM} - < \Delta E >)\right]}{\sum_j^N \exp\left[-\beta(E_j^{MM} - E_j^{QM} - < \Delta E >)\right]} \tag{10}$$

where $\Delta E = E^{MM} - E^{QM}$. This weighting method gives larger weights to conforma-

tions in which the MM energy is underestimated ($E^{MM} - E^{QM} < 0$) than to conformations in which the MM energy is overestimated ($E^{MM} - E^{QM} > 0$), with respect to the reference (QM) energy. Hence, as pointed out by its original authors,[50] configurations with negative $E^{MM} - E^{QM}$ have a spuriously large thermodynamic weight in the MM representation and are more likely to appear during MM sampling, which could lead to incorrect equilibrium averages due to incorrect equilibrium structures. On the other hand, configurations with positive $E^{MM} - E^{QM}$ have a spuriously small weight in the MM representation, which could result in overestimation of transition state energies and underestimation of fluctuations. Therefore, by heavily penalizing configurations with MM energies that are lower than QM energies, this weighting procedure avoids the creation of spurious MM minima and forces the fitting errors into the high-energy regions, which are, in a sense, higher-order error than the incorrect equilibrium averages.

- **Manual weighting**: This weighting method allows the user to choose the weights of each conformation, which will be constant throughout the whole optimization. This may be of special importance if the user knows which conformations should be given more/less importance. Other publications have suggested that weights of less than or equal to five are typically appropriate for the under-represented conformations, assuming weights of unity for the rest of the target data.[47]

Charge fitting to electrostatic potential: the RESP model

ParaMol can derive atom-centered point charges by fitting to a reference electrostatic potential (ESP).[41] Specifically, ParaMol contains an implementation of RESP model.[42,43] The objective function used in the multiconformational RESP fit reads:

$$X_{RESP}(\boldsymbol{q}) = \sum_{i}^{N_s} \omega_i \sum_{k}^{N_{grid}} \left(\sum_{j}^{N_{charges}} \frac{q_j}{r_{jk}} - V_k^{QM} \right)^2 + \lambda_1 \left(\sum_{j}^{N_{charges}} q_j - q_{tot} \right) + \sum_{m=2}^{N_{constraints}} \lambda_m f_m(\boldsymbol{q}) + \Theta(\boldsymbol{q}) \tag{11}$$

where ω_i is the weight of the i-th conformation, $\boldsymbol{q} = (q_1, ..., q_j)$ is the vector of charges allowed to vary during the fitting, V_k^{QM} is the value of the calculated ESP at the grid point k and r_{jk} is the distance between the atomic centre j and the grid point k. Furthermore, λ_1 corresponds to the Lagrange multiplier used to constraint the sum of the charges to the total molecular charge and λ_m (with $m > 1$) to the Langrange multipliers used do impose other types of constraints such as, for instance, symmetry constraints.[43] Finally, as in equation (1), $\Theta(\boldsymbol{q})$ defines the penalty function optionally applied so that the fit becomes restrained.

ParaMol is able to perform the charge fitting by using SciPy's[51] implementation of the COBYLA,[52] SLSQP[53] or Trust Region[54] algorithms. Moreover, we have also implemented an analytical solution of the the system of equations that arises from taking the derivatives of equation (11) with respect to the charges and Lagrange multipliers. More information about the implementation of the analytical solution can be found in refs. 42 and 41.

Preconditioning of the optimizable parameters and regularization

In order to avoid over-fitting, which may occur whenever the amount of reference data used in the optimization is not extensive enough, regularization has to be applied so that, during the optimization, the parameters remain within a range of values that makes physical sense. This is done through the inclusion of the penalty functions Θ in eqs. (1) and (11). In Bayesian statistics, penalty functions correspond to the negative logarithm of a *prior* distribution, and the regularized objective function corresponds to the *posterior* distribution.[24] Hence, it is possible to design penalty functions by making assumptions regarding the *prior* distribution

of the parameters. ParaMol has implemented different regularization methods. For instance, if the user wants to apply L1 regularization, *i.e.*, if the prior distribution of a parameter p is assumed to be given by $P(p) = \exp(-\frac{|p-p^0|}{\gamma})$, where γ controls the width of the distribution and p^0 is the parameter initial guess, then the penalty function reads:

$$\Theta_{L1}(\boldsymbol{p}) = \alpha \sum_{m}^{N_p} \frac{|p_m - p_m^0|}{\gamma_m} \tag{12}$$

where α is an adjustable parameter that controls the strength of the regularization. Similarly, if the user wants to apply L2 regularization, *i.e.*, if the *prior* distribution of the parameters is assumed to be a Gaussian, a harmonic penalty function is then employed, which reads:

$$\Theta_{L2}(\boldsymbol{p}) = \alpha \sum_{m}^{N_p} \frac{(p_m - p_m^0)^2}{\gamma_m^2} \tag{13}$$

The widths of the *prior* distributions may be automatically generated or manually chosen by the user using physical knowledge. Regarding the automatic generation of these hyper-parameters, ParaMol uses a procedure wherein either the arithmetic or geometric mean is calculated for classes of FF parameters (*e.g.*, bond force constants, dihedral phases, etc.). All parameters within the same class will then use this mean value as the width of their *prior* distributions. This is similar to the approach followed by ForceBalance.[25] Moreover, the procedure used to automatically generate the *prior* widths may be also used to construct the Jacobi preconditioner, which scales the parameters so that that they are all treated on the same footing by the optimization algorithm. Specifically, the Jacobi (diagonal) preconditioner used in ParaMol is given by $\boldsymbol{P} = \gamma_m \delta_{mm}$.

Finally, if charges are being fitted, it is also possible to apply a hyperbolic regularization term that prevents the charges from deviating too much from a target charge of zero.[42] This hyperbolic penalty function is given by:

$$\Theta_{HB}(\boldsymbol{q}) = \alpha \sum_{m}^{N_{charges}} \left((q_m^2 + \beta^2)^{1/2} - \beta \right) \tag{14}$$

where α and β are adjustable hyper-parameters that define the asymptotic limits of the strength of the restraint and the tightness of the hyperbola around its minimum, respectively.

Optimization algorithms

We have implemented in ParaMol global and local optimization algorithms that perform non-linear minimization of the objective function, *viz.*, non-reversible Monte Carlo,[55] gradient descent,[56] stochastic gradient descent[57] and simulated annealing.[58] Furthermore, ParaMol also interfaces with the Python SciPy package, from which several minimization algorithms can be used (*e.g.*, Nelder-Mead, Powell, BFGS, L-BFGS-B, SLSQP, COBYLA, Trust Region). Since ParaMol has no implementation of analytical derivatives of the objective function with respect to the set of parameters being optimized, whenever necessary the Jacobian matrix is calculated using numerical derivatives, and the Hessian matrix is approximated using BFGS or SR1 updates.[51,59]

In addition to the non-linear, iterative optimizers previously described, ParaMol also offers analytical linear least square (LLS) solutions to the parametrization of the bonded part (bond, angle and dihedral terms) of class I FFs.[45,46] This fitting approach can be employed alongside any of the available regularization schemes, though currently only to find the minimum of the squared deviations of the energies, as shown in equation (5). It also does not support the use of the non-Boltzmann weighting of equation (10), as the dependence of this weighting method on the MM energies makes it suited to be solved only through non-linear optimization. The main disadvantage of the LLS fitting approach is that it provides a single, deterministic answer, whereas a scatter of possible solutions with nearly the same quality concerning the objective function usually exist. On the other hand, iterative methods such as the stochastic Monte Carlo or gradient-based optimizations can find other nearby

solutions, which may have value if they produce different simulation outcomes that may be preferred in specific cases (*e.g.*, produce the right helical propensity or orientation of a drug molecule in the protein binding site). Nevertheless, it should be stressed that these solutions should only be fielded once the absolute optimum obtained by the LLS fitting has been attempted.

ParaMol package structure

ParaMol is mainly designed to be used as a Python package that can be easily extended by the user to include extra functionalities or to develop parametrization protocols that are not included by default. The Python (sub)subpackages and modules that comprise ParaMol's top-level package, as well as the main interactions between them are depicted in figure 1. ParaMol uses OpenMM[60] as its MM engine and has implemented wrappers of AMBER, DFTB+[61,62] and ASE.[63] The latter allows single-point or geometry optimization calculations to be performed using any of the calculators or optimizers available in it. Moreover, ParaMol has also implemented symmetrizers that allow subjecting re-parametrizations to the symmetries defined in the topology files used by MM packages such as AMBER, CHARMM or GROMACS. Interfaces to read and write input files for these packages are also available.

In order to set up a custom parametrization protocol using ParaMol, firstly we create the ParaMol's representation of our systems of interest resorting to the *ParaMolSystem* object defined in the *system* module of the *System* subpackage. An instance of this object stores the reference data, contains the MM (modules in *MM_engines* subpackage) and QM engines (wrappers defined in the modules of the *QM_engines* subpackage) used by the system, and stores ParaMol's representation of the system's force field (modules in *ForceField* subpackage). Furthermore, we have to create an instance of the *ObjectiveFunction* object, which is defined in the *objective_function* module of the *ObjectiveFunction* subpackage. The *ObjectiveFunction* object requires some properties (objects defined in the modules of the *ObjectiveFunction.Properties* subsubpackage) to be fitted to the reference data, and an instance

of the *ParameterSpace* object defined in the *parameter_space* module of the *ParameterSpace* subpackage, which stores the vector space of optimizable parameters. We also need to use one of the optimizers available in the modules of the *Optimizers* subpackage to perform the minimization of the objective function.

Alternatively, it is also possible to use one of the tasks already implemented in the modules of the *Tasks* subpackage to perform specific parametrization protocols (described in detail in the next subsection). Tasks greatly simplify the use of ParaMol because they usually only require *a priori* instancing of *ParaMolSystem* objects and of the desired task. More information about the ParaMol Python package and examples of how to use it can be found at ParaMol's website - `https://paramol.readthedocs.io`.

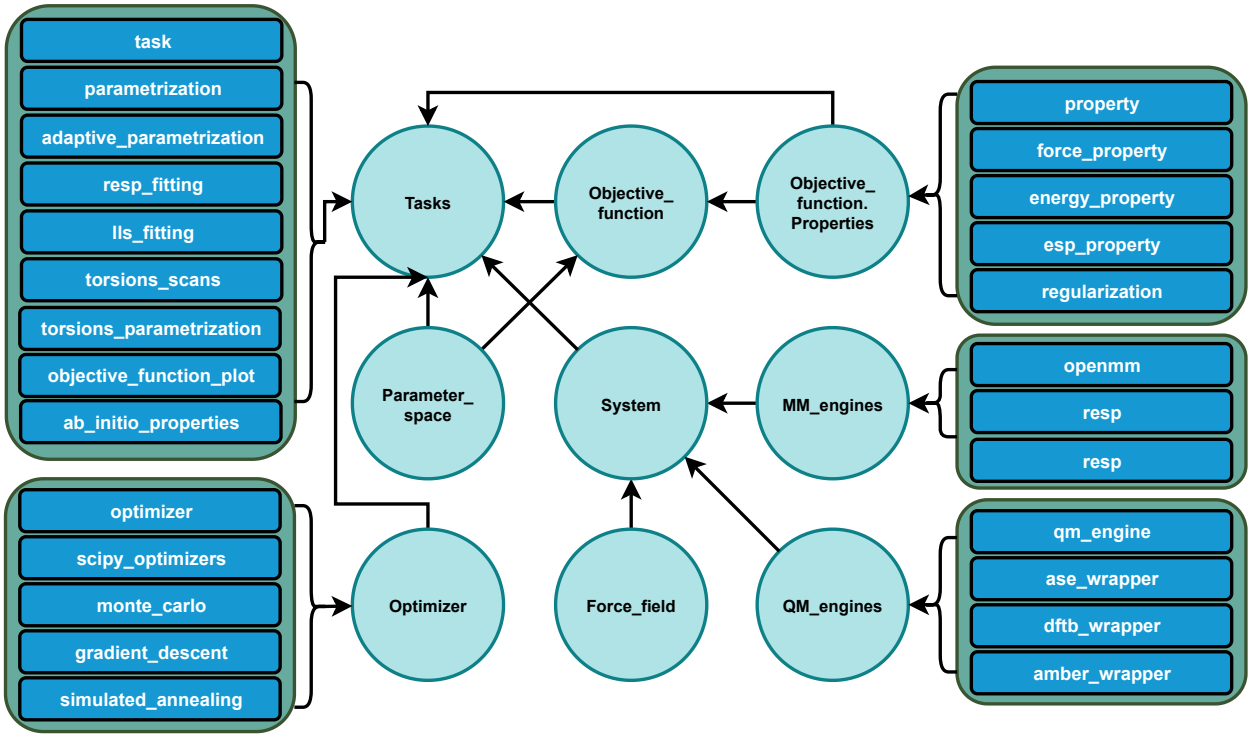


Figure 1: Overview of the structure of the ParaMol Python top-level package. Paramol's (sub)subpackages are represented as cyan circles and the respective modules as blue rectangles. The most relevant interactions between (sub)subpackages are represented with arrows. The direction of the arrows indicates that modules from the destination (sub)subpackage require modules from the source (sub)subpackages (*e.g.* the modules of the *System* subpackage require the modules of the *Force_field*, *MM_packages* and *QM_packages* subpackages). Some (sub)subpackages and modules are not shown for the sake of conciseness.

ParaMol tasks

ParaMol includes built-in tasks that perform specific parametrization protocols and also routines that aid the parametrization protocols themselves as, *e.g.*, utilities that assess convergence of the optimization procedure and utilities to calculate *ab initio* reference data. Currently, the parametrization tasks available in ParaMol are the following:

- **Parametrization** (*ParaMol.Tasks.parametrization*): This tasks performs ParaMol's standard parametrization protocol. Specifically, the Parametrization task creates the parameter space, the objective function, and the optimizer that will be used in the derivation of the FF parameters. It also preconditions the optimizable parameters and it defines the constraints to which the optimization will be subjected, *e.g.*, total charge or symmetry constraints. Regarding symmetry-constraints, these have to be defined manually by the user so that physical-based symmetries are retained. Alternatively, it is also possible to apply ParaMol's AMBER, CHARMM or GROMACS symmetrizer that subject the optimization to the symmetries defined in the respective topology files. Finally, during the optimization procedure itself, the objective function typically has to be evaluated hundreds to thousands of times. This evaluation can be done either in serial or in parallel. For the latter case, it is possible to use OpenMM support of different platforms and distribute the computation of the MM properties amongst the available CPUs or GPUs by using Python's multiprocessing package.

- **LLS fitting** (*ParaMol.Tasks.lls_fitting*): This task is very similar to the Parametrization task, except that it does not support the non-Boltzmann weighting scheme and can only parametrize the bond, angle and dihedral terms of class I FFs by minimizing the squared deviations of the energies. The LLS solution is calculated by resorting to Numpy's[64] numpy.linalg.lstsq function.

- **Adaptive parametrization** (*ParaMol.Tasks.adaptive_parametrization*): This task performs adaptive parametrization, which consists in a self-consistent loop wherein

in each iteration configurational sampling and parameter optimization are carried out. The workflow of this task is described as follows: given an initial guess of FF parameters, a set of configurations is generated using any integrator available in OpenMM, and the reference *ab initio* data for this set is calculated; then, a new set of optimal parameters is determined by resorting to the *Parametrization* task; finally, convergence of the self-consistent procedure is assessed, which is assumed to occur if the RMSD of the current parameters with respect to the previous iteration parameters is less than a user-defined threshold; if convergence was not achieved, another iteration of the loop is performed. It is also worth mentioning that the correction to the weights of the conformations described in Ref. 25 may be optionally applied in every iteration. This correction removes the bias introduced by the fact that conformations at different iterations are generated using different classical potentials, since in each iteration different FF parameters are used.

- **Dihedral Scans** (*ParaMol.Tasks.torsions_scans*): This task performs 1D or 2D relaxed dihedral scans. Specifically, for the 1D case (a 2D scan would follow the same approach), the task requires specification of the quartet of atoms a-b-c-d for which the potential energy scan will be performed by rotation about the b-c bond. By default, at a given step of the scan, only the dihedral angle being scanned is fixed, allowing the remaining DOFs to relax during the geometry optimization. However, it is also possible to further constrain other DOFs during a scan as, *e.g.*, bonds, angles, or other dihedrals, a feature that resorts to the capabilities of the ASE package.[63] Nevertheless, it is worth mentioning that these extra constraints come with the caveat that the PES being explored is now an adiabatic constrained PES.

- **Automatic Soft Dihedral Parametrization** (*ParaMol.Tasks.torsions_parametrization*): This task allows the automatic parametrization of soft (rotatable) dihedrals in a special fashion, inspired by the protocol used by GAAMP.[21] This approach is of particular

importance because soft dihedrals, which have small energy barriers, are the ones that control the conformational preferences of a molecule. Therefore, their accurate description is of paramount importance as they crucially shape the topology of the PES. The first step of the task is the identification of the soft bonds - here defined as bonds that contain soft dihedrals -, which is done resorting to the RDKit package.[65] ParaMol then iterates over all soft bonds to perform relaxed scans of their soft dihedrals. At this point, it is worth mentioning some specifics of this algorithm, *viz.*: if a soft bond has more than two soft dihedrals of the same type, *i.e.*, soft dihedrals that share exactly the same atom types, a relaxed dihedral scan is only performed at the first encounter with this soft dihedral type; furthermore, if two or more soft bonds have exactly the same soft dihedrals types, they are considered to belong to the same soft bond type and, thus, scans are only performed at the first encounter with a soft bond of this type. Another aspect of this algorithm is that, every time a new soft dihedral type is scanned, an optimization of the parameters of that soft dihedral type is performed. This step has proved to be important to generate smoother energy profiles because, by default, ParaMol performs MM geometry optimizations before the QM geometry optimization, a preconditioning that substantially decreases the computational cost of the high-level calculation. Hence, by having a gradually better MM representation, we "bias" the MM optimization algorithm to determine more QM-like energy minima that will, eventually, re-direct the QM optimizations towards relevant conformers. Finally, once ParaMol has iterated over all soft bonds containing soft dihedrals, a concomitant parametrization of all soft dihedral parameters is performed using the calculated relaxed dihedral scans as a data set. Worthy of note is that, in the final optimization, the optimized parameters generated in the intermediate re-parametrizations are forgotten, as ParaMol performs the final re-parametrization starting from the originally provided MM parameters. Further information about this task can be found in Supporting Information (SI).

- **RESP charge fitting** (*ParaMol.Tasks.resp_fitting*): This task performs the derivation of charges by fitting to a reference ESP that can be obtained from quantum chemistry packages, as it was previously described in the subsection where the RESP model implementation was presented. ParaMol currently can extract the ESP directly from a Gaussian output. The output of other software has to be converted by the user to the format read by ParaMol.

- **Calculation of *ab initio* reference data** (*ParaMol.Tasks.ab_initio_properties*): This task calculates *ab initio* reference data by using either any QM calculator available in the ASE package, or one of the wrappers of quantum chemistry packages implemented in ParaMol. These calculations can be either performed in serial or in parallel by distributing the workload amongst the available CPUs by using Python's multiprocessing package.

Application examples

In what follows, we present examples of re-parametrization of drug molecules. For this purpose, we used the general Amber force field form (GAFF), for which the functional form reads:

$$V = \sum_{bonds} K_b(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}\left[1 + \cos\left(n\phi - \gamma\right)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \tag{15}$$

where r_{eq} and θ_{eq} are equilibrium structural parameters; K_b, K_θ and V_n are the bond, angle and dihedral force constants, respectively; n is the dihedral multiplicity and γ is the phase angle. The re-parametrizations were performed using SciPy's SLSQP optimizer and they were deemed to be converged whenever the objective function between two successive iterations did not change by more than 10^{-6}, *i.e.*, $X_{n+1} - X_n < 10^{-6}$ (10^{-8} for the norfloxacin analog example). Furthermore, GAFF parameters were used as the initial guess for the optimizations, except when stated otherwise. L2 (harmonic) regularization was applied, and the *prior* widths used throughout this study were inspired by the values reported in Ref. 50 (see Supporting Information). The objective function included as targets either forces - equation (3) - , energies - equation (5) - , or both at the same time. The parametrization of the drug molecules used in this paper was performed using Antechamber packages, which are part of AmberTools. AM1-BCC charges were calculated after the geometry was optimized at the target level of theory, which for the present work was either the DFTB+[61,62] implementation of SCC-DFTB including the D3 dispersion correction[66] with Becke-Johnson damping,[67] the nonlocal van der Waals DFT functional VV10,[68,69] as implemented in the linear-scaling DFT package ONETEP,[70–72] or the Psi4[73] implementation of the long-range corrected hybrid DFT functional wB97X-D[74] with the 6-31G* basis set. The choice of these QM levels of theory relies on the evidence that they perform quite well in determining conformations and

respective energies.[75–79] The topology and coordinates files used as inputs to ParaMol were created using LEaP. All atom type symmetries were preserved during the re-parametrization, unless otherwise stated.

Dihedral scans and fitting: norfloxacin analog

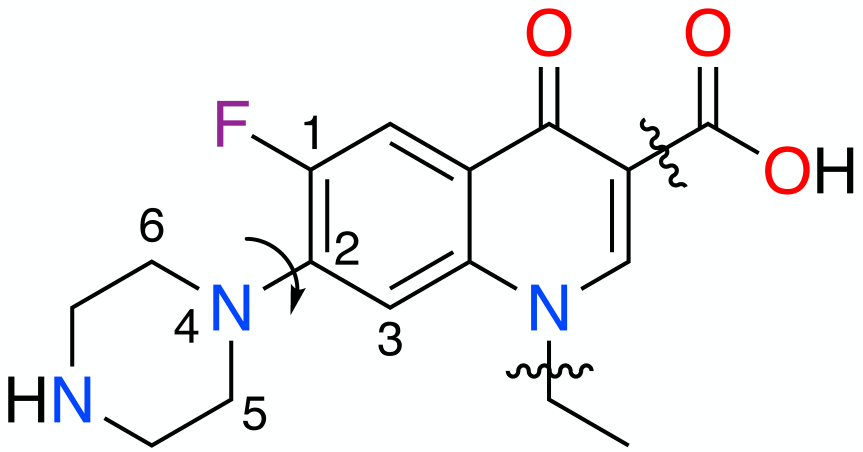


Figure 2: Molecular structure of norfloxacin. Owing to unavailability of fluorine parameters in the mio-1-1 set,[80] the molecule used in this example is a norfloxacfin analog, wherein we substituted the fluorine attached to the C1 carbon by a hydrogen atom. The differences in the torsional preferences introduced by this change are negligible, as it is shown in SI using data from the Cambridge Structural Database (CSD).[81] Additionally, as a typical example of a fragment-based approach, we cut the molecule at the positions indicated by the wavy lines. All dangling bonds were capped with hydrogen atoms.

The first example of this paper concerns the dihedral scan functionality implemented in ParaMol. In order to illustrate the procedure and the issues that may arise when re-parametrizing dihedrals of a drug molecule, we optimized the parameters (force constants and phase constants) of the dihedrals associated with the main rotatable bond of a norfloxacin analog (C2-N4) (see figure 2). As it is an achiral molecule, to increase the transferability of the parameters we have constrained the phase constants to be fixed to the symmetric position of 180° (0° would be equivalently valid).[11,46] Topologically-speaking, two dihedrals contain C2 and N4 as their inner atoms: C5-N4-C2-C1 and C5-N4-C2-C3. Both have the same atom types and, therefore, share the same set of parameters. GAFF models this dihedral type

(c3-nh-ca-ca) by including only one term with periodicity $n = 2$. Nevertheless, to increase the flexibility of the FF, we included all terms in the Fourier expansion with peridocities from $n = 1$ to $n = 6$. The aim of the numerical experiments performed here was to assess the performance of the different weighting methods implemented in ParaMol, *viz.*, uniform, Boltzmann and non-Boltzmann weighting, when attempting to reproduce a target dihedral energy profile. Furthermore, we also illustrate the differences between the MM-relaxed and QM-relaxed approach. The results obtained for the wB97X-D/6-31G* MM-relaxed dihedral fittings are presented in figure 4, and the final optimized parameters are show in table 1 (the results of the SCC-DFTB-D3 re-parametrizations are shown in SI). The final parameters obtained using the QM-relaxed approach of equation (6) are shown in table 1, and the wB97X-D/6-31G* dihedral energy profiles are shown in figure 5. All fittings were performed using the objective function of equation (7) with an additional L2 regularization term with a scaling constant of $\alpha = 0.1$. The weighting temperature employed was 500 K, and both the SLSQP SciPy optimizer and the LLS fitting approach were employed to derive the optimized dihedral parameters.

Before analysing the results obtained, it is worth discussing some considerations about the the QM dihedral energy profiles. Owing to substantial discontinuities that were obtained in the one-dimensional dihedral scans, we opted to perform fittings to a two-dimensional PES. The observed discontinuities were related to conformational changes that the piperazinyl ring underwent as the N4-C2 bond was rotated, caused by a flipping of the pyramidal geometry of the N4 center, which led to sudden energy variations. This phenomenon is clearly seen by following the profile defined by the red stars in figure 3, which indicate the minimum energy structure for a given ϕ angle. Hence, to avoid the jumps in energy that would be seen in the one-dimensional dihedral scan, we have opted to calculate a two-dimensional PES wherein we varied the C5-N4-C2-C1 (ϕ) dihedral angle from to -180° to 170° in steps of 10°, whilst concomitantly varying C2-C6-N4-C5 (ψ) improper dihedral angle from 120° to 180°, and -178° to -120° in steps of 2°, performing a total of 2196 geometry optimizations that resulted

in the 2D PES represented in figure 3.

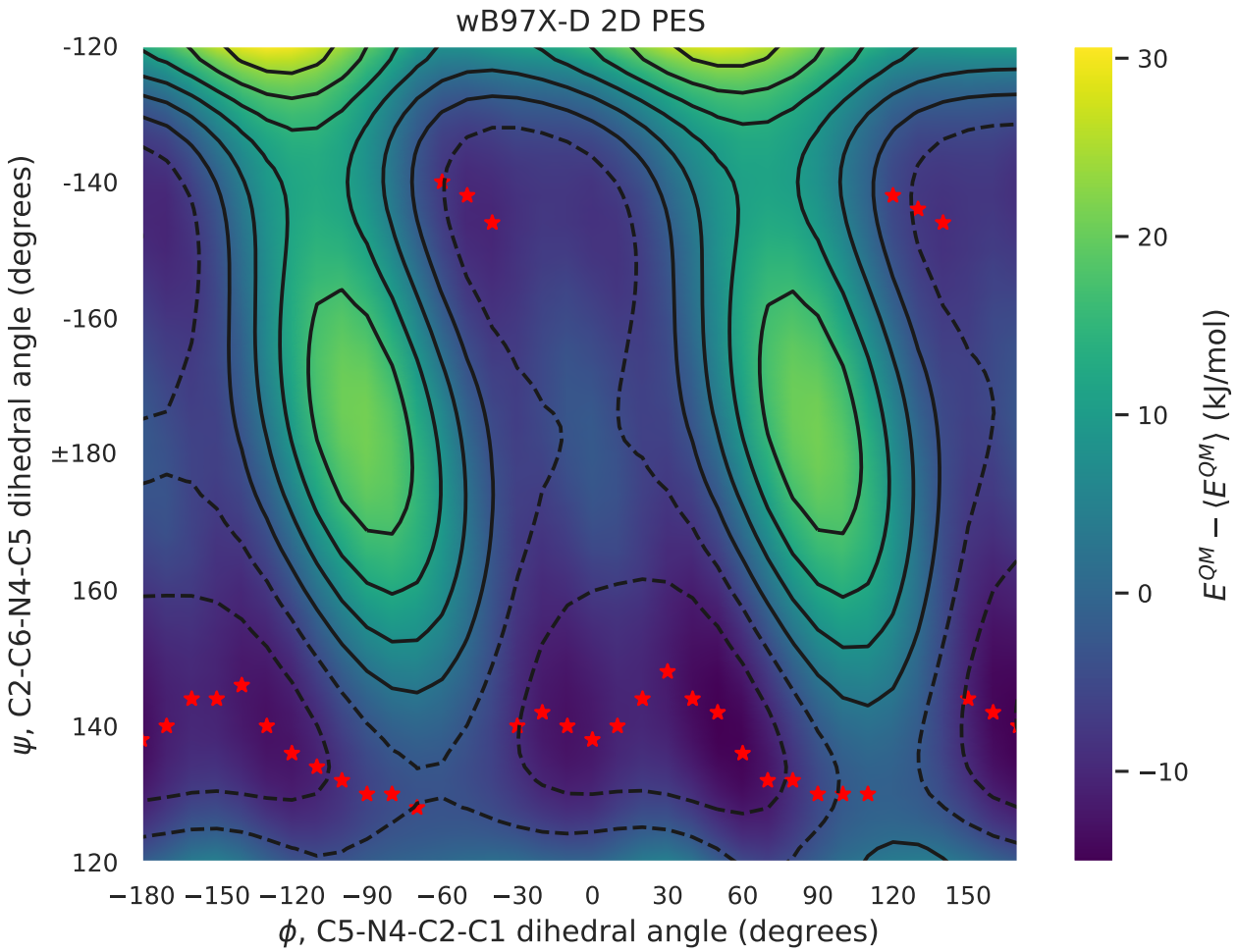


Figure 3: wB97X-D/6-31G* 2D PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) dihedrals scan for the norfloxacin analog fragment. The red stars correspond to the minimum energy structure for a given ϕ dihedral angle value.

Besides the spontaneous conformational changes that may occur when rotating a chosen dihedral, another issue that commonly produces discontinuities in relaxed dihedral scans is hysteresis in the energy associated with the dihedral being relaxed.[45] This is prone to occur when optimizing one data point after another, since non-orthogonal DOFs may be put under strain, therefore accumulating potential energy that is released once the molecule crosses a given threshold, causing the strained DOFs to relax. A commonly employed solution to identify and correct this issue is to perform scans in both directions and to pick the one that yields the more physically-sensible profile. This practice is of particular importance

because sudden physically-based changes in energy as the ones seen in figure 3 may be easily mistaken by path-related hysteresis, and if the latter are artificial, the former are desirable to be captured (ideally exhaustively scanned by performing higher-dimensional scans) if the aim is to extensively describe the conformational preferences of a molecule. Additionally, in many applications it is usually possible to impose a certain degree of symmetry in one-dimensional dihedral profiles by constraining specific DOFs (in this example, by freezing the C2-C6-N4-C5 improper dihedral), so that conformational changes and energy jumps are avoided. Nevertheless, it is important to bear in mind that by doing so we are exploring an adiabatic constrained PES that misrepresents the (at least the local) minimum energy for a given dihedral angle value. Consequently, due to this reason, for this application example we have decided to proceed with two-dimensional relaxed scans, as they are a better representation of the PES of the molecule.

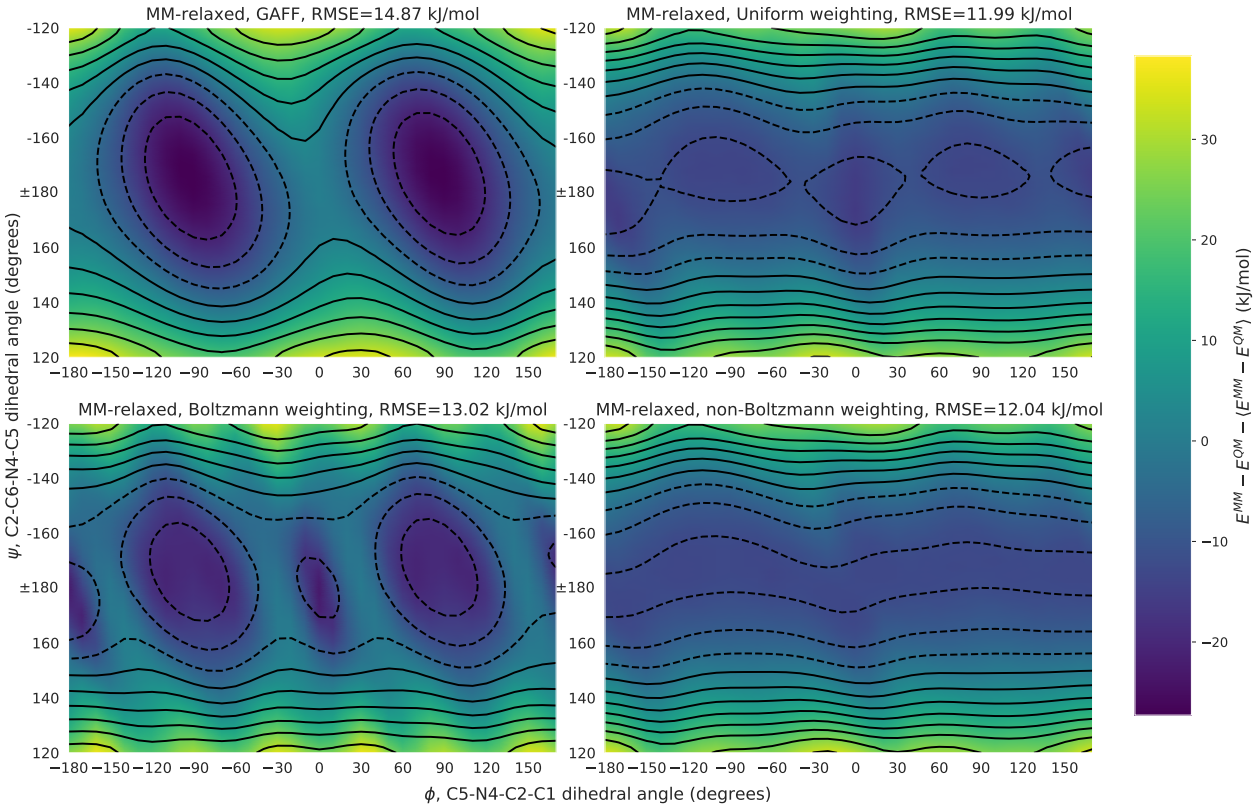


Figure 4: Relative errors of the MM FFs (GAFF, uniform, Boltzmann and non-Boltzmann weightings) with respect to the target (wB97X-D/6-31G*) 2D PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) dihedrals scan. The MM-relaxed approach was employed to optimize the FFs, which means that the MM PES was calculated based on the MM-relaxed rather than the QM-relaxed geometry used in the fittings.

Through the analysis of the fitting curves of figure 4 and of the SCC-DFTB-D3 MM-relaxed fittings shown in SI, we conclude that a significant improvement is observed for the three sets of parameters with respect to GAFF. Specifically, uniform and non-Boltzmann weighting performed the best in both wB97X-D/6-31G* and SCC-DFTB-D3 levels of theory, leading to fittings with root mean square errors (RMSE) values of 11.99/11.47 kJ/mol and 12.04/11.49 kJ/mol, respectively, for wB97X-D/SCC-DFTB-D3, while Boltzmann weighting performed slightly worse in terms of RMSE (13.02/12.52 kJ/mol). Furthermore, non-Boltzmann weighting - which emphasizes the correction of regions of the scan where the MM energy is lower than the QM energy - led to an overall robust description of the QM minima and, more importantly, it showed a tendency to skew the distribution of the errors

towards positive values, being overall the weighting scheme with less negative relative errors (see figure 4). On the other hand, since Boltzmann weighting emphasizes having a good description of the QM minima, it was the scheme that performed the worst for conformations located near high-energy barriers (see, *e.g.*, regions located at $\phi = [-120°, -90°]$ and $\phi = [90°, 120°]$): it underestimated the energies of these transition state conformations by as much as *ca.* 20 kJ/mol for both wB97X-D/6-31G* and SCC-DFTB-D3. Hence, overall, even though uniform weighting led to the best RMSE values, the residuals of this scheme tend to be symmetrically-distributed around zero (see figure 4 and SCC-DFTB-D3 fittings in SI), which makes it prone to the creation of artifacts in the PES, such as spurious minima; conversely, as non-Boltzmann weighting emphasizes correcting regions of the PES for which the MM energy is lower than the QM energy, the creation of spurious MM minima is substantially mitigated, a feature that leads us to advocate for its use.

Table 1: wB97X-D/6-31G* dihedral force constants (kJ/mol) derived using the MM-relaxed/QM-relaxed approach.

	GAFF	**Uniform**	**Boltzmann**	**non-Boltzmann**
	SciPy SLSQP solution			
V_1	0.00	-2.19 / -3.88	0.16 / -1.73	-1.67 / -4.21
V_2	17.57	7.98 / 8.26	6.86 / 7.22	7.87 / 8.62
V_3	0.00	-7.64 / -4.18	-3.52 / -2.12	-6.97 / 0.60
V_4	0.00	0.15 / 0.92	1.67 / 2.03	-0.29 / 0.61
V_5	0.00	0.50 / 0.66	-1.94 / 1.34	0.56 / 1.37
V_6	0.00	0.27 / 0.13	1.60 / 1.64	0.10 / 0.32
	LLS solution			
V_1	0.00	-2.18 / -3.88	0.21 / -1.74	-
V_2	17.57	7.98 / 8.26	6.86 / 7.22	-
V_3	0.00	-7.64 / -4.18	-3.50 / -2.07	-
V_4	0.00	0.15 / 0.92	1.67 / 2.03	-
V_5	0.00	0.50 / 0.66	-1.94 / 1.34	-
V_6	0.00	0.27 / 0.13	1.60 / 1.64	-

Through the analysis of the values of the final optimized parameters that are shown in table 1, we conclude that the SLSQP SciPy optimizer and the LLS fitting gave identical results in terms of accuracy. Moreover, the parameters did not stray far away from physically-

sensible values, which robustly indicates that the regularization applied was strong enough to avoid non-physical force constants. Therefore, even though not applying regularization may allow the optimization procedure to reproduce small details in the dihedral profile correctly, we advocate for the use of regularization as it normally allows parameters to be obtained that that are more suited to be used in standard MM simulation methods. It is also worth noting that by imposing L2 regularization, the optimized parameters depend on their initial guesses and, therefore, the results here presented might be potentially improved by starting the optimizations from better initial guesses. Nevertheless, since in many cases it is not straightforward to postulate good initial parameters, we decided to test the limiting case in which $V_i = 0.0$ kJ/mol, $\forall i$, for which we have shown than even a blind guess led to improved FFs. Our experience suggests $V = 0.0$ kJ/mol is usually a good initial guess and, consequently, it is the one we recommend using by default in the absence of better ones.

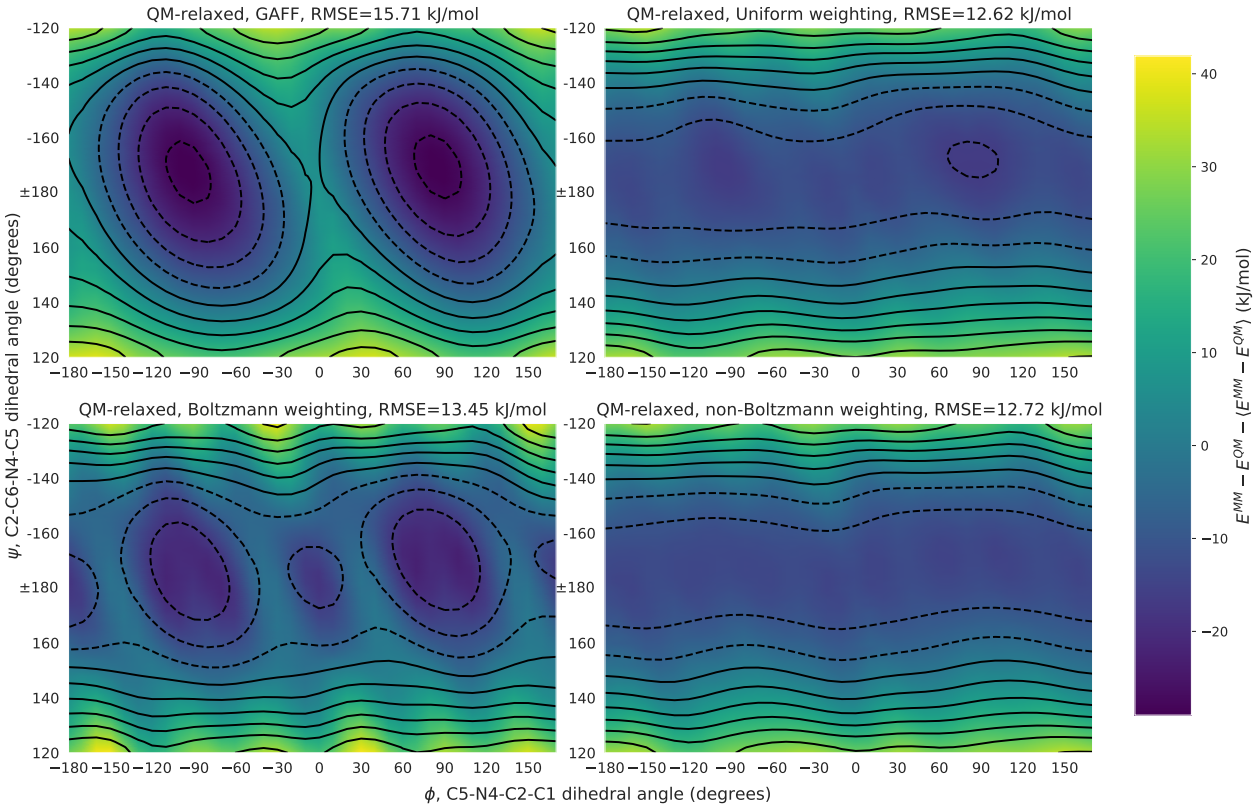


Figure 5: Relative errors of the MM FFs (GAFF, uniform, Boltzmann and non-Boltzmann weightings) with respect to the target (wB97X-D/6-31G*) 2D PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) dihedrals scan. The QM-relaxed approach was employed to optimize the FFs, which means that the optimized QM geometry was to used to calculate both the MM and QM PES used in the fittings.

Finally, with regards to the QM-relaxed approach, it is worth discussing the bias that, in most cases, this methodology introduces in the final derived parameters. A naive analysis of the obtained QM-relaxed energy profiles might lead us to consider them as putatively correct: the fittings shown in figure 5 exhibit a similar agreement with the target wB97X-D/6-31G* PES as the ones obtained for the MM-relaxed approach, and the derived FF parameters are within physically-sensible ranges (table 1). Despite this, the artifacts introduced by this approach manifest themselves when MM-relaxed energy profiles are calculated from the QM-relaxed-derived FFs. Hence, we proceeded to perform this extra MM-relaxation of the QM-relaxed-derived FFs, for which the results obtained are shown in figure 6. Through the analysis of these plots, it can be seen that the QM-relaxed approach led to the creation of

non-negligible artifacts in the PES as, *e.g.*, the spurious minima observed at *ca.* 0° and ±180°. Hence, since the QM-relaxed approach is critically dependent on the other intramolecular FF parameters,[48] it substantially biased the derived FF parameters and, therefore, we advocate against its use. The artifacts arising from using the QM-relaxed approach are normally more serious the lower the dimensionality of the PES we are fitting to, since the chances of not covering substantial mismatches between the MM and QM levels of theory increase.

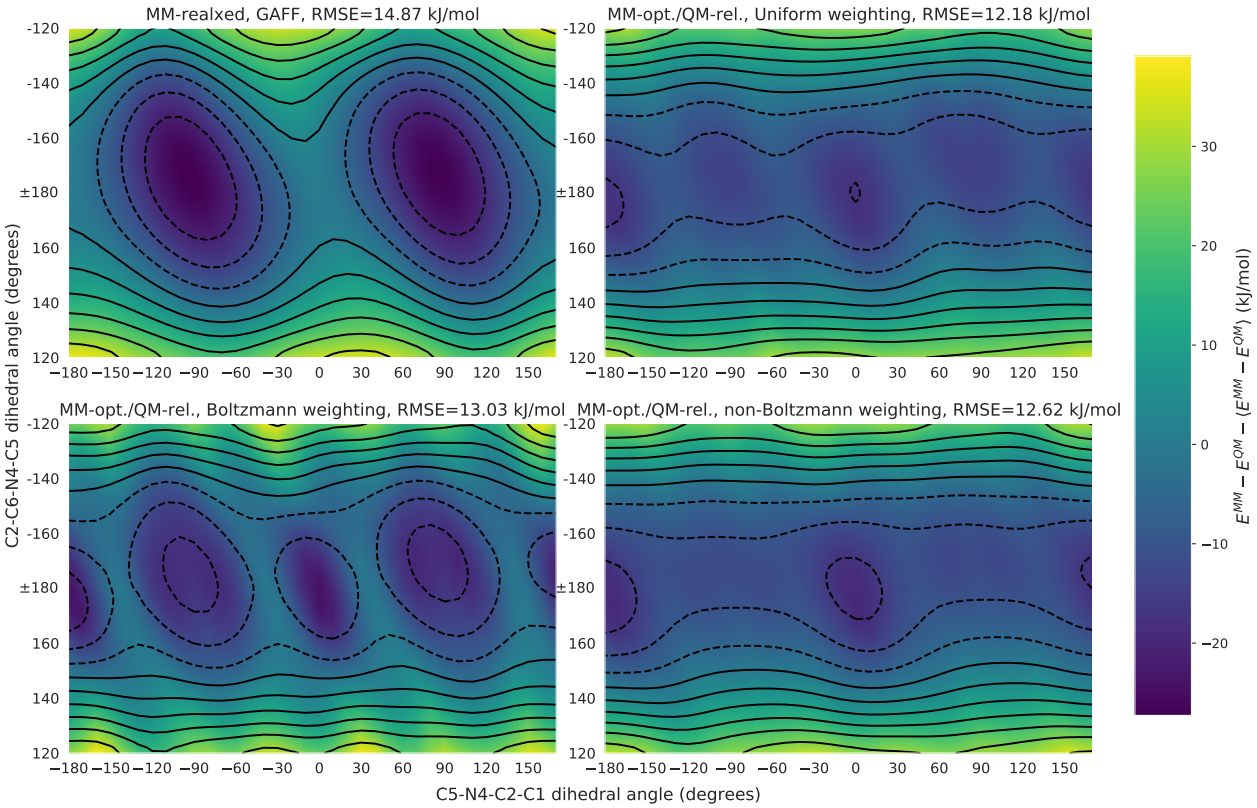


Figure 6: Relative errors of the MM FFs (GAFF, uniform, Boltzmann and non-Boltzmann weightings) with respect to the target (wB97X-D/6-31G*) 2D PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) dihedrals scan. The MM PES used to calculate the relative errors were obtained by MM optimization of the QM-relaxed PES of figure 5.

Parametrization of aspirin

As a second example of the parametrization methodologies implemented in ParaMol, we explore and discuss the results obtained in the parametrization of aspirin when it was performed using data sets either obtained through relaxed dihedral scans or by generating an ensemble of configurations through a MD simulation. The main aim of these parametrization experiments was to reproduce the conformational preferences of aspirin , *i.e.*, its configurational distribution, at the SCC-DFTB-D3 level of theory.

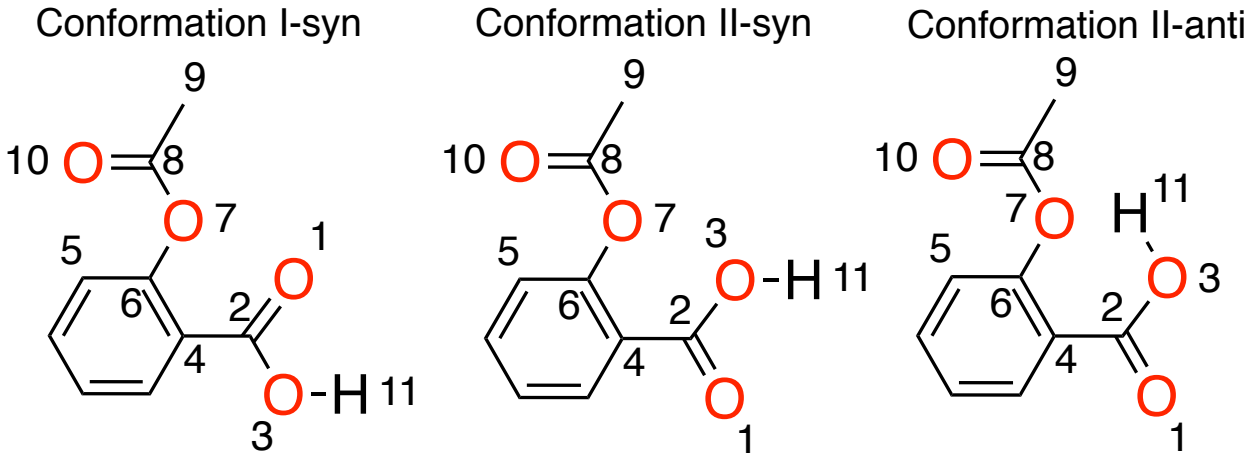


Figure 7: Molecular structures of aspirin. SCC-DFTB-D3 MD simulations sample mostly configurations that occur through rotation of the C6-O7 bond in conformations I-syn and II-syn. Moreover, even though conformation II-anti is not sampled in the SCC-DFTB-D3 MD simulation, it is shown here because it is sampled in some re-parametrized FFs. Representative crystal structures of these conformations obtained from the CSD are shown in SI. It is important to mention that, in solution, the carboxylic acid of aspirin assumes, predominantly, its deprotonated state (pKa=3.49-3.6 at 25 °C[82]). Therefore, the results and discussions in this section concerning the II-syn and II-anti conformations are simply illustrations of the features of the parametrization protocols employed, as these are mainly relevant in gas phase.

All dihedral relaxed scans were performed using the tasks available in ParaMol. The 1-dimensional energy profiles of aspirin were created from 36-point relaxed scans wherein each point was spaced by 10°. These were obtained for all soft dihedrals excluding the ones associated with the trivial rotation of the methyl group, as the QM energy profile of such dihedrals generally can be reproduced reasonably by the GAFF,[21] and the dihedrals

involved in the rotation of the O7-C8 bond, as it is fairly rigid (see figure 7). All scans' energy minimizations were performed while fixing only the dihedral being scanned and they were deemed to be converged when the force on all individual atoms was less 1×10^{-2} eV Å^{-1}. On the other hand, to generate an ensemble of aspirin configurations at the SCC-DFTB-D3 level of theory, we performed a gas-phase MD simulation using the DFTB+ package during 10 ns, wherein snapshots were collected every 1 ps, resulting in a total of 10000 configurations. In this simulation, the Nosé-Hoover thermostat was applied to maintain the temperature at 350 K with a coupling strength of 3400 cm^{-1}, a value close to the calculated highest vibrational frequency of the molecule (3670.75 cm^{-1}).

As seen in figure 8, the most striking difference between the populations of the soft dihedral predicted by GAFF and SCC-DFTB-D3 is observed for the populations of the dihedrals involved in the rotation of the C6-O7 soft bond (C4-C6-O7-C8 and C5-C6-O7-C8), which have the same atom types and, therefore, they share the same set of parameters. They are originally modelled by the GAFF through only one dihedral term with periodicity $n = 2$, leading to the two minima observed in its dihedral distribution, in contrast with the four minima predicted by SCC-DFTB-D3. Hence, as the number of minima of our target distribution does not match the number of minima predicted by the potentials of these dihedrals, we increased the flexibility of GAFF in our re-parametrization by including all terms in the Fourier expansion with periodicities from $n = 1$ up to $n = 4$. Furthermore, even though dihedrals C4-C6-O7-C8 and C5-C6-O7-C8 share the same set of FF parameters, their SCC-DFTB-D3 dihedral populations are quite different to each other, which implies that if we were to use a single potential to model both dihedrals, we would have to rely on the non-bonded terms to implicitly break their symmetry. However, this is a clear example of the inability of the non-bonded terms of equation (15) to correctly model the intra-molecular interactions that occur in aspirin, especially given the weak hydrogen bond that can be established between O10 and H11, which is predominantly of electrostatic character since it occurs at distances of ca. 4.0 Å (see SCC-DFTB-D3 distribution in figure 9).[83] This can be

concluded by examining the populations of the C5-C6-O7-C8 and C4-C6-O7-C8 dihedrals at *ca.* $\pm 130^{\circ}$ and $\pm 60^{\circ}$, respectively, for which the GAFF populations decay smoothly, not skewing towards the weakly hydrogen-bonded conformations that are present in the SCC-DFTB-D3 ensemble. Moreover, the configurational distributions generated by simple re-parametrization of the original GAFF potential energy function predicted wrong global minima and, in general, gave poor agreement with respect to the SCC-DFTB-D3 distribution (see SI). Hence, as a work-around for this issue, we decided to break the symmetry of these dihedrals during the parametrizations performed in ParaMol, which artificially compensates for the limitations of the non-bonded (especially electrostatic) terms of the FF. In practical terms, this is equivalent to introducing a new atom type at, *e.g.*, position C4, since this carbon is linked to a carboxylic group and, consequently, its nature is very different from the C5 carbon atom. This makes it possible to independently optimize the parameters for each of these dihedrals, a step which proved to be essential for the reproduction of the SCC-DFTB-D3 configurational distribution, as the simple augmentation of the number of dihedral terms was insufficient to do so. Lastly, for the two dihedral types involved in the rotation around the C8-C9 bond, GAFF assigns three terms with periodicites $n = 1, 2, 3$ to the O7-C8-C9-H dihedrals (o-c-c3-hc type), and one term with periodicity $n = 2$ to the O10-C8-C9-H dihedrals (os-c-c3-hc type). Despite this, we decided to only assign terms with $n = 3$ since this is a multiplicity not forbidden by symmetry. This is because these dihedrals have a sp^3 carbon as one of the inner atoms (C1), which has three identical hydrogen substituents and, therefore, all terms with multiplicity that is not a multiple of 3 will cancel.[45]

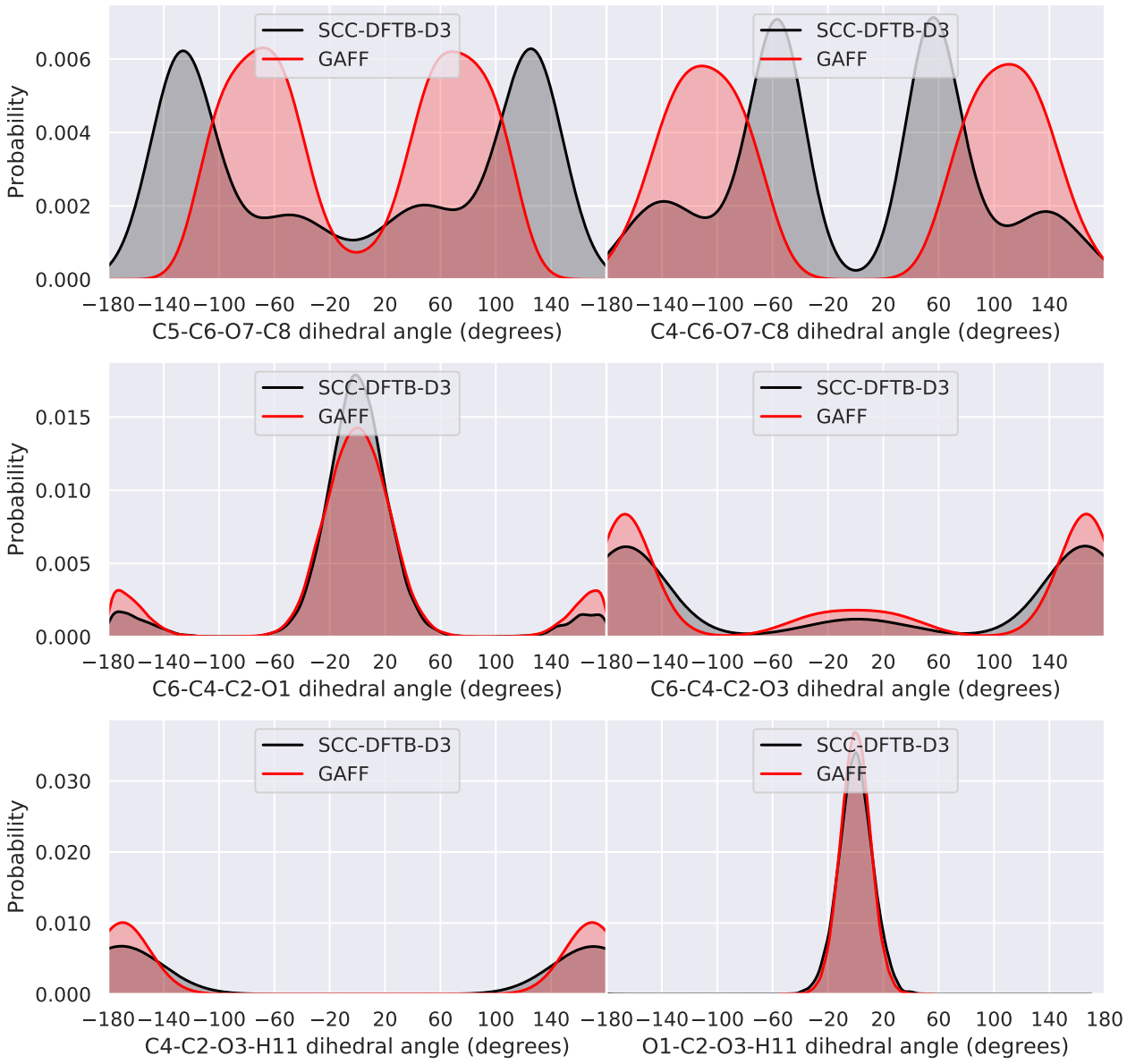


Figure 8: Kernel density estimations of the populations of the soft dihedrals of aspirin obtained by MD simulations (SCC-DFTB-D3 and original GAFF). The soft dihedrals here presented - C5-C6-O7-C8, C4-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11 and O1-C2-O3-H11 - are the ones for which parameters were generated using the data set consisting of relaxed dihedral scans.

Regarding the re-parametrizations performed using the ensemble of 10000 configurations generated through a SCC-DFTB-D3 MD simulation, these were designed to assess the performance of the different weighting methods, here employed at various regularization strengths. Specifically, we re-parametrized all intra-molecular parameters of aspirin, such that the vec-

tor of parameters that was optimized was $\boldsymbol{p} = \{\boldsymbol{K_b}, \boldsymbol{r_{eq}}, \boldsymbol{K_\theta}, \boldsymbol{\theta_{eq}}, \boldsymbol{V_n}, \boldsymbol{\gamma}\}$. In order to do this, we employed an objective function that fits both energies and forces through the use of the equations (5) and (3), respectively, plus the additional L2 regularization term of equation (13). The total number of parameters concomitantly optimized was 108 and the original GAFF parameters were used as initial guesses. After new FF parameters were derived, the optimized FFs were used to perform 100 ns of MD simulations, which were carried out using a Langevin integrator with a friction coefficient of 2 ps^{-1}, a time-step of 1 fs and at a temperature of 350 K. Snapshots of the simulations were collected every 10 picoseconds, amounting for a total of 10000 snapshots.

The configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle obtained when using the non-Boltzmann weighting - equation (10) - at temperature $T = 500$ K is shown in figure 9, whilst the distributions obtained for $T = 300$ K, 1000 K, 2000 K are shown SI. Examining the cases for which $T = 300$ K and $T = 500$ K, it can be concluded that for strong regularization strengths, *i.e.*, $\alpha = 1$ and $\alpha = 0.1$, the optimized FFs reproduced quite well the general features of the target distribution, as they were able to sample the 4 minima that occur through rotation of the C6-O7 bond of conformation I-syn (outer edge of the distribution), as well as the states for which aspirin assumes the conformation II-syn (inner edge of the distribution). On the other hand, when an intermediate regularization scaling factor ($\alpha = 0.01$) was employed, configurations in which aspirin assumes the II-anti conformation were sampled for $T = 500$ K, even though these are not observed in the SCC-DFTB-D3 distribution. This spurious sampling is further aggravated when the weakest regularization strength ($\alpha = 0.001$) was employed, for which the simulations became kinetically trapped in these conformations, despite being started from the global minimum geometry (conformation I-syn with a C5-C6-O7-C8 dihedral angle value of *ca.* 130°). The sampling of spurious conformations is a frequent issue when using re-parametrized FFs in MM simulations methods and it may occur whenever these geometries are absent in the data set that was used to perform the fitting. Hence, in this case, since the fitting procedure

had no information about the SCC-DFTB-D3 forces and energies of the II-anti conformations and the transition states that lead to them, the barrier heights for conversion of the carboxylic group from syn to anti conformation were underestimated. A possible solution for this issue is to further re-optimize the FF, this time also including the sampled spurious conformations, so that the optimization procedure also takes them into account. In this way - and since, if anything, non-Boltzmann weighting tends to overestimate barrier heights and/or equilibrium energies -, we would prevent the over-sampling of undesired geometries during the dynamics. Nevertheless, it is worth noting that this problem was not present for the strongly-regularized FFs - a clear indication of the importance of regularization that, by not allowing the FF parameters to stray away too much from physically-sensible values, helps in preventing the creation of spurious minima.

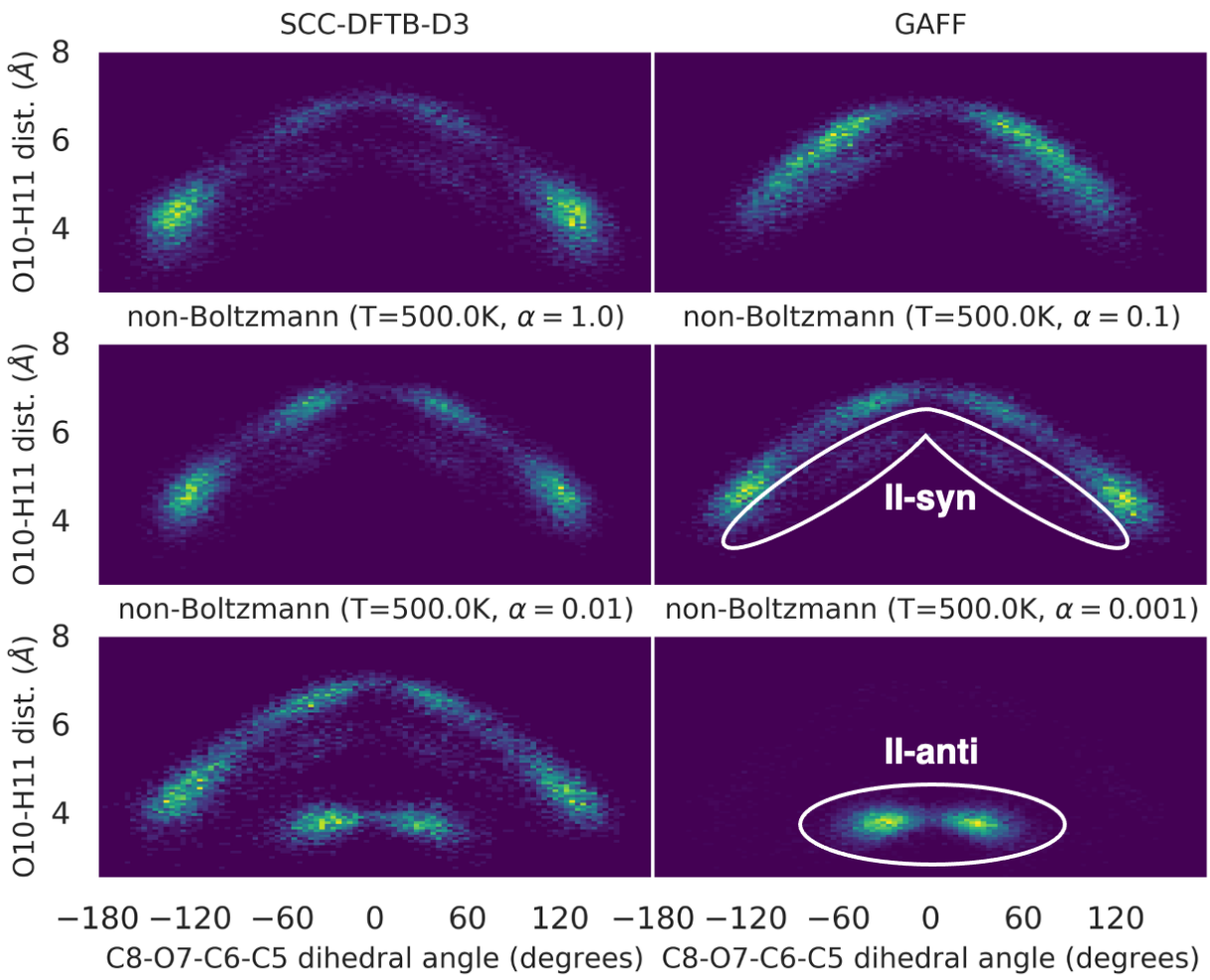


Figure 9: Aspirin configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle DOFss for SCC-DFTB-D3, GAFF, and the GAFF.MOD (re-parametrized) FFs derived employing non-Boltzmann weighting at a temperature of 500 K and different regularization strengths ($\alpha = \{1.0, 0.1, 0.01, 0.001\}$). The data set used in the re-parametrization was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations generated through MD simulations.

The configurational distributions that were obtained when using Boltzmann weighting - equation (9) - with strong regularization ($\alpha = 1.0$) and at different temperatures $T = 300$ K are shown in SI. Through their analysis, we can conclude that the overall agreement to the target distribution was poor, albeit the conformation I-syn and conformation II-syn minima were sampled (except for $T = 500$ K), yet with incorrect frequency. Furthermore, it can also be observed that the distributions are highly asymmetric and show sampling of the

spurious II-anti conformation, suggesting an overestimation of the barrier heights between the different minima and an underestimation of the syn to anti energy barrier. It is also worth mentioning that the configurational distributions for the FFs generated with other regularization strengths are not shown here because they sampled unphysical configurations. This observation strongly indicates that Boltzmann weighting requires strong regularization to produce FFs that can be potentially used in MM modeling. The reason for this may be attributed to the fact that, as Boltzmann weighting emphasizes the description of the QM minima, if these regions of the PES are over-fitted at the cost of poorly describing the remainder of the energy landscape, as soon as the molecule moves away from the minima, the PES is completely unphysical, which, ultimately, leads to distorted geometries and wrong dynamics.

Finally, the configurational distributions that were obtained when uniform weighting was employed - equation (8) - with different regularization levels ($\alpha = \{1.0, 0.1, 0.01, 0.001\}$) are shown in SI. As for the Boltzmann scheme, the uniform weighting-derived FFs gave poor agreement with the target distribution and all but the strongest regularization case were kinetically trapped at the global minimum conformation from which the simulations were started. This is a clear indication either of the over-stabilization of this minimum or of the overestimation of the transition states to which it is connected. The former argument is justified by the fact that unlike, for example, the non-Boltzmann scheme, uniform weighting allows equally for positive and negative errors in the fit, which may lead to negative $E^{MM} - E^{QM}$, *i.e.*, spuriously large thermodynamics weights. Incidentally, the asymmetries that might be imposed on the PES by equally allowing for positive and negative errors are an issue that was already reported and discussed by other authors.[50] On the other hand, the latter reason may occur because the global minimum geometries are the most populated in the SCC-DFTD-D3 distribution and, consequently, the configurations that appear with higher frequency in the data set are used in the fittings. Therefore, the optimization procedure implicitly captures this configurational over-representation and, thus, it primarily attempts

to minimize the objective function at this region, a bias that, eventually, may lead to poor description of under-represented configurations, such as transition states.

Table 2: RMSE of the energies (kJ mol^{-1}) / Average RMSE of the atomic force (kJ mol^{-1} $Å^{-1}$ $atom^{-1}$). These RMSEs were calculated for the configurations of the SCC-DFTB-D3 configurational ensemble data set, and they represent the energies and forces errors between the SCC-DFTB-D3 level of theory and the re-parametrized FFs models. The formulae used to compute these RMSEs are reported in SI.

	$\alpha = 1.0$	$\alpha = 1.0$	$\alpha = 0.01$	$\alpha = 0.001$
GAFF	53.24 / 140.95	53.24 / 140.95	53.24 / 140.95	53.24 / 140.95
non-Boltzmann (T=300 K)	44.03 / 94.68	41.81 / 87.29	40.61 / 83.25	39.75 / 84.81
non-Boltzmann (T=500 K)	43.85 / 93.87	41.73 / 87.15	40.38 / 82.62	39.60 / 83.12
non-Boltzmann (T=1000 K)	43.63 / 93.97	41.05 / 87.87	39.59 / 82.55	38.81 / 81.57
non-Boltzmann (T=2000 K)	43.61 / 95.13	40.39 / 88.91	38.61 / 83.38	37.75 / 81.20
Boltzmann (T=300 K)	49.78 / 124.33	47.27 / 110.25	46.88 / 109.13	48.66 / 123.05
Boltzmann (T=500 K)	44.73 / 124.62	38.97 / 110.61	37.15 / 113.44	36.81 / 123.93
Boltzmann (T=1000 K)	42.35 / 111.84	37.02 / 103.67	34.12 / 102.17	33.53 / 103.05
Boltzmann (T=2000 K)	42.29 / 107.81	37.12 / 100.01	33.95 / 95.15	32.92 / 95.45
Uniform	42.28 / 104.50	37.11 / 96.75	33.92 / 91.54	32.84 / 91.44

Overall, as general guidelines when using configurational ensembles as the fitting data, we advocate for the use of non-Boltzmann weighting as this weighting scheme is generally less sensitive to the regularization strength and it yields best performance when attempting to reproduce a target configurational distribution. Furthermore, strong regularization scaling factors ($\alpha = 1.0$ and $\alpha = 0.1$) result in more reliable parameters than intermediate ($\alpha = 0.01$) and weak ($\alpha = 0.001$) regularization scaling factors, as a better agreement to the target distribution was observed for the former cases. Additionally, employment of strong regularization also comes with the advantageous feature that it does not allow the parameters to deviate much from their original values, a highly desirable property as we aim to keep the FF parameters within a range of physically-sensible values. Intermediate and weak regularization strengths lead to FFs that have, in general, lower energy RMSEs and lower average atomic force RMSEs (see table 2) than their strongly-regularized counterparts (except when Boltzmann weighting is employed at low temperatures, as, in these situations, there is a tendency to over-fit the QM minima). However, these putative better fittings come at the

cost of creation of artifacts in the PES, such as, *e.g.*, spurious minima. Finally, as expected, progressively employing higher weighting temperatures ($T = 1000$ K and $T = 2000$ K) in the non-Boltzmann and Boltzmann schemes leads to results that become gradually similar to the ones that are obtained upon use of the uniform weighting scheme. Hence, since uniform weighting does not perform particularly well when using a configurational ensemble as the fitting data set, we advocate against the use of high weighting temperatures. Therefore, unless one has a specific reason to do so, temperatures in a range between 300 K to 500 K are preferable.

Let us now turn our discussion to the results obtained through re-parametrization of the soft dihedrals of aspirin when these were performed using relaxed dihedral scans as the fitting data set. The soft dihedrals for which dihedral were scanned were C5-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11 and O1-C2-O3-H11. Furthermore, for optimization purposes, the C4-C6-O7-C8 dihedral was additionally included in this set due to the symmetry breaking previously mentioned. The optimizations *per se* were performed using the MM-relaxed approach of equation (7) with strong regularization ($\alpha = 1.0$) and at a temperature of 500 K. The parameters that entered in the optimization are given by the vector $\boldsymbol{p} = \{\boldsymbol{V_n}, \boldsymbol{\gamma}\}$, wherein any pair of V_n and γ belongs to a term of the previously mentioned dihedrals. These were all optimized concomitantly, amounting to a total of 26 parameters. The fittings obtained when employing Boltzmann weighting are shown in figure 10, while the fittings for the non-Boltzmann and uniform weighting methods are shown in SI. Through the analysis of these figures, it is possible to conclude that most of the improvement in the fittings occurred for the C5-C6-O7-C8 dihedral, an observation that supports the argument that the main source of the mismatch seen between the GAFF and the SCC-DFTB-D3 distributions comes from the soft dihedrals that model the rotation around the C6-O7 bond. For the remaining dihedrals, modest improvements or even slight worsening is obtained. The latter effect occurs because the optimization procedure may sacrifice some accuracy in specific dihedrals to obtain a better global agreement. Furthermore, through the analysis of the

configurational distribution represented in figure 11, we can conclude that, independently of the weighting scheme applied, the agreement obtained to the target distribution was quite good. It is also interesting to notice the sampling, even though very rarely, of the II-anti conformation and, surprisingly, of the I-anti conformation, which was visited even less often, when using non-Boltzmann and uniform weighting. Therefore, since there is nothing in the energy profiles that potentially indicates an underestimation of the barrier height for the syn to anti conversion, which, as can be seen in the O1-C2-O3-H11 dihedral energy profiles, is of *ca.* 40 kJ/mol, we do not exclude that this anti conformations could be rarely visited during a SCC-DFTB-D3 dynamics, and that we simply did not sampled them because we only performed 10 ns of MD, in comparison to the 100 ns of the classical MD. Incidentally, a careful analysis of the SCC-DFTB-D3 conformations reveals the presence of 5 frames that would correspond to the anti-I conformation, which are simply not seen in the distribution plot due do its extreme rareness.

Overall, all weighting schemes performed similarly well. Nevertheless, as a general guideline to follow when employing this re-parametrization approach, we also advocate for the use of non-Boltzmann weighting due to its features, although this recommendation is less strict than for the configurational ensemble data set re-parametrizations. Furthermore, with regards to the regularization strength, our experience indicates that, in most cases, this re-parametrization approach requires a strong-to-intermediate regularization scaling factor ($\alpha = 1.0$ or $\alpha = 0.1$), as attempts to use weaker regularization strengths resulted, in general, in unstable FFs that tend to be unsuitable for dynamics.

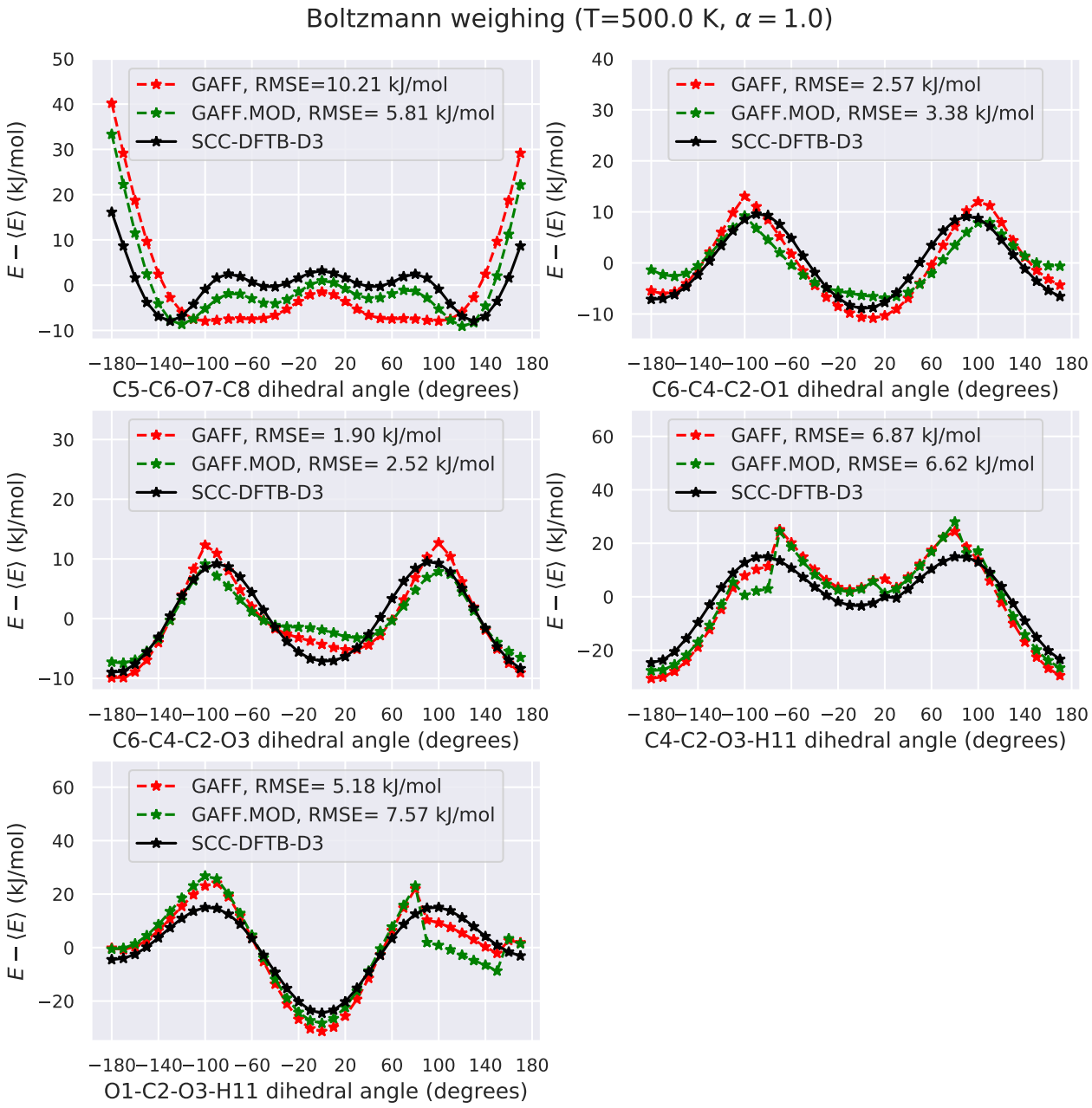


Figure 10: Comparison of SCC-DFTB-D3, GAFF and GAFF.MOD (re-parametrized FF) dihedral energy profiles for the C5-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11, O1-C2-O3-H11 dihedrals. The GAFF curves correspond to MM-relaxed energy profiles. The GAFF.MOD FF was obtained by employing the MM-relaxed approach with Boltzmann weighting (T=500.0 K, $\alpha = 1.0$) in the concomitant optimization of the parameters of the dihedrals above represented plus the C4-C6-O7-C8 dihedral using the automatic soft fihedral parametrization task.

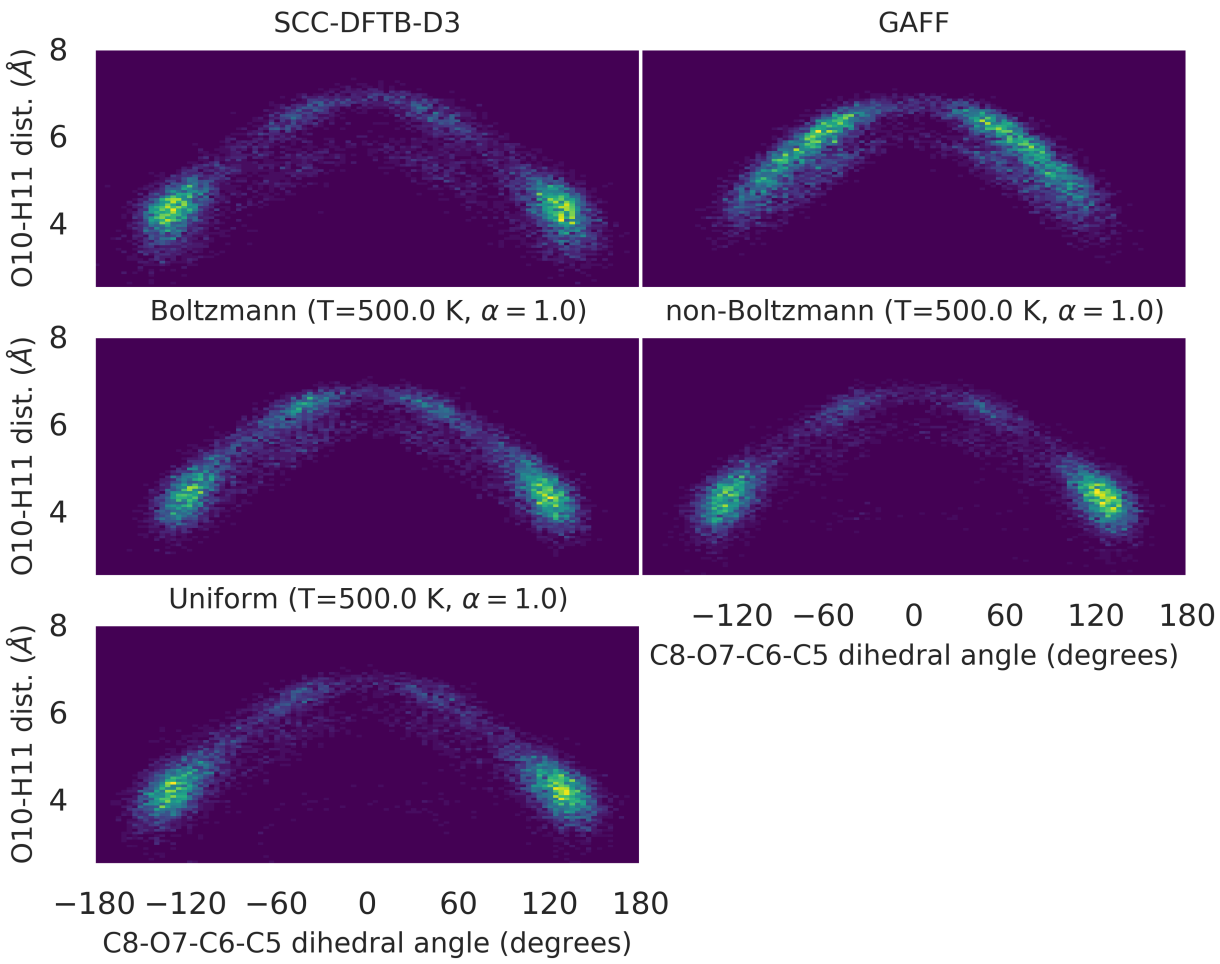


Figure 11: Aspirin configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle DOFss for SCC-DFTB-D3, GAFF, and the FFs derived through re-parametrization of the soft dihedrals of aspirin when employing Boltzmann (figure 10), non-Boltzmann and uniform weighting methods (T = 500 K, α = 1.0, figures in SI). All represented distributions contain 10000 configurations generated through MD simulations.

Lastly, we also attempted to reproduce the SCC-DFTB-D3 distribution by simply parametrizing the C5-C6-O7-C8 and C4-C6-O7-C8 dihedrals, *i.e.*, the soft dihedrals that model the rotation around the C6-O7 bond, as these are the main source of the mismatch seen between the GAFF and the SCC-DFTB-D3 distributions. These re-parametrizations were performed using the MM-relaxed approach of equation (7) with strong regularization ($\alpha = 1.0$) and at a temperature of 500 K. As seen in the dihedral energy profiles of figure 12, the re-parametrized FFs (GAFF.MOD) dihedral energy profiles are in excellent agreement with respect to the

SCC-DFTB-D3 ones, leading to a decrease in the energy RMSE from 10.21 kJ/mol (GAFF) to 3.19 kJ/mol (Boltzmann weighting), 3.24 kJ/mol (non-Boltzmann weighting) and 1.29 kJ/mol (uniform weighting). Furthermore, through the analysis of the configurational distribution represented in figure 13, we can conclude that regardless of the weighting scheme applied, the agreement to the target distribution is quite good. Despite this, a slight underrepresentation of the configurations of the II-syn conformation can also be noted, as well as very rare sampling of the II-anti conformation for the Boltzmann-weighted FF. Nevertheless, all in all, parametrization of the dihedrals associated with the main faulty soft bond proved to be an efficient route to correctly model the conformational dynamics of aspirin, especially given that it is computationally cheap compared with the previously presented methods, and that all weighting schemes performed similarly well.

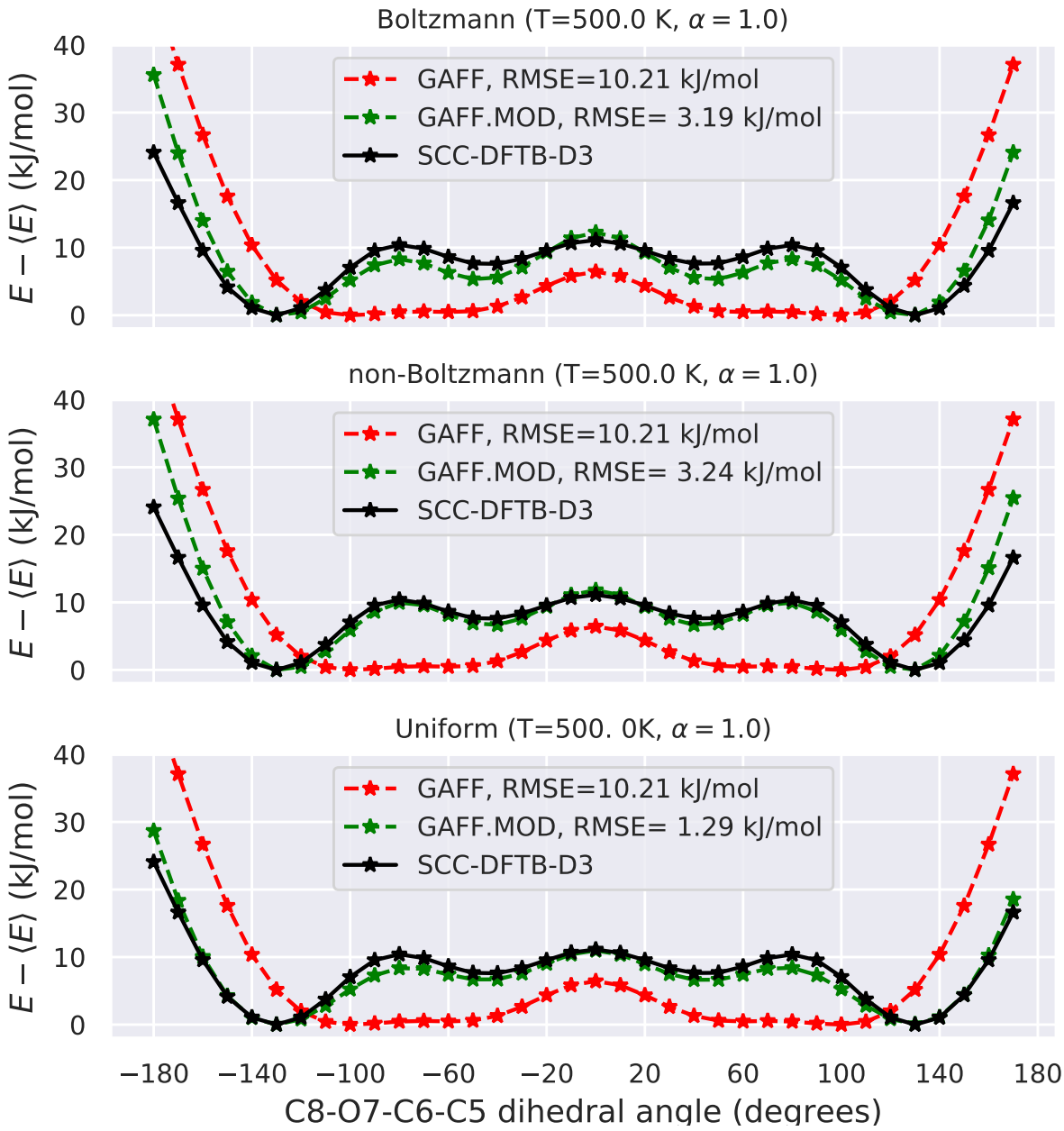


Figure 12: Comparison of SCC-DFTB-D3, GAFF and GAFF.MOD (re-parametrized FF) dihedral energy profiles for the C5-C6-O7-C8 dihedral. The GAFF curves correspond to MM-relaxed energy profiles. The GAFF.MOD FF was obtained by employing the MM-relaxed approach with the indicated weighting method ($T = 500$ K, $\alpha = 1.0$) in the optimization of the parameters of the C5-C6-O7-C8 dihedral.

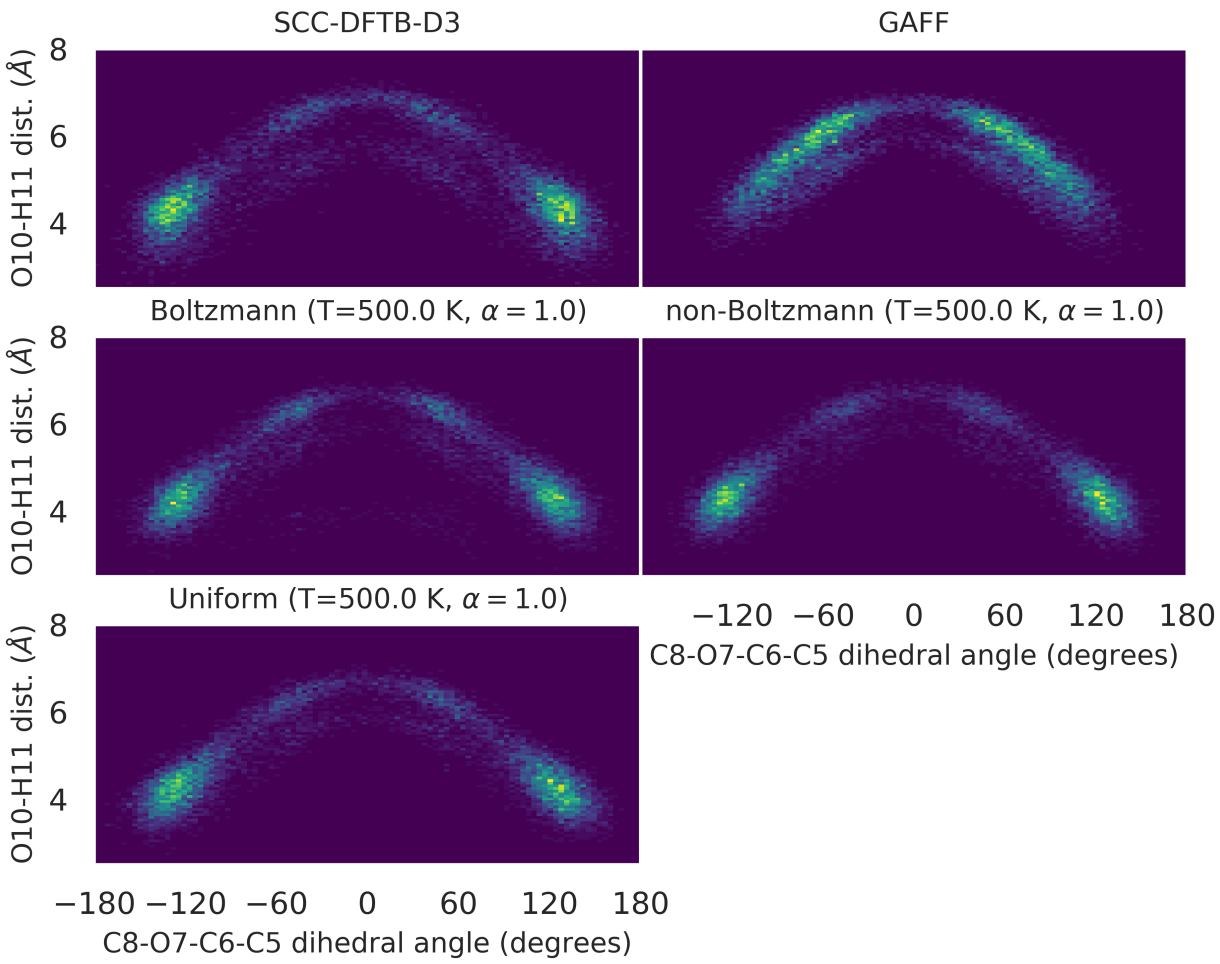


Figure 13: Aspirin configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle DOFss for SCC-DFTB-D3, GAFF, and the GAFF.MOD FFs of figure 12 (obtained by re-parametrization of the dihedrals associated with the main faulty soft bond), which were derived employing different weighting methods ($T = 500$ K, $\alpha = 1.0$). All represented distributions contain 10000 configurations generated through MD simulations.

Adaptive parametrization of caffeine

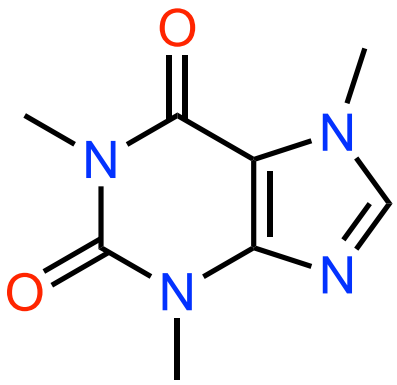


Figure 14: Molecular structure of caffeine.

As a last illustrative example of ParaMol's parametrization capabilities, we re-parametrized all intra-molecular parameters of caffeine to the VV10 level of theory using adaptive parametrization (the SCC-DFTB-D3 results are shown in SI). Specifically, the vector of parameters that entered in the optimization was $\boldsymbol{p} = \{\boldsymbol{K_b}, \boldsymbol{r_{eq}}, \boldsymbol{K_\theta}, \boldsymbol{\theta_{eq}}, \boldsymbol{V_n}, \boldsymbol{\gamma}\}$. The total number of optimizable parameters was 156. The minimized objective function included an energy, force and regularization terms, as given by equations (3), (5) and (13).

In every iteration of the adaptive parametrization procedure, 100 new configurations separated 0.5 ps from each other were generated and added to the previous ones. These configurations were obtained using Langevin dynamics with a friction coefficient of 2 ps^{-1}, a time-step of 1 fs, and a temperature of 300 K. No special sampling technique was employed to explore the PES of caffeine as it is mostly planar and does not have much conformational variability. The adaptive parametrization procedure was deemed to be converged when the root-mean square deviation of the parameters between two successive iterations was less than 10^{-4}. The adaptive parametrization performed in total 23 iterations until convergence (2300 structures in total in the last iteration).

The plots of the RMSD of the parameters and the components of the objective function as a function of the iteration number are shown in figure 15. Through the analysis of this figure, we can see that most of the improvement in the objective function occurred in the first

2-3 iterations, as shown by the θ_{L2} regularization term, which remained practically steady afterward, indicating that only small adjustments to the FF parameters occurred. After this substantial initial refinement, the steady increase of the X_E energy term may be attributed simply to the convergence of the relative populations of the configurational ensemble used in the parametrization, as the initial and final X_E values were practically the same. Interestingly, this steady increase is not seen in the X_F term, suggesting that for caffeine it was less sensitive to the completeness of the configurational ensemble. Furthermore, not much variation in any term occurred after the 15th iteration, as can be seen through the stabilization of the objective function terms, which is a robust indication that, at this point, both the sampling and the parameter optimization were practically converged.

Finally, to evaluate the improvement of the energies and forces, we generated two different data sets using either the FF before (GAFF) and after re-parametrization (GAFF.MOD), wherein we sampled 1000 configurations using each one of these FFs. This was done by performing short MD simulations with the same settings as before (snapshots were collected every picosecond). Since each data set was generated by sampling from its respective FF, this analysis evaluated how close to the target level of theory each FF samples. The plot that shows the correlation between the QM energies and the MM energies is represented in figure 16 and the atomic forces errors are shown in the molecular structures of figure 17. The root mean square error of the energies before and after re-parametrization was 12.82 kJ/mol and 6.73 kJ/mol, respectively, which reveals that GAFF.MOD samples conformations that are closer to VV10 level of theory, energetically-speaking. On other hand, the average RMSE of the atomic forces improved from 83.61 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ to 50.95 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ after re-parametrization, a clear indication that GAFF.MOD is an improved FF in relation to GAFF since it predicts better energies and forces.

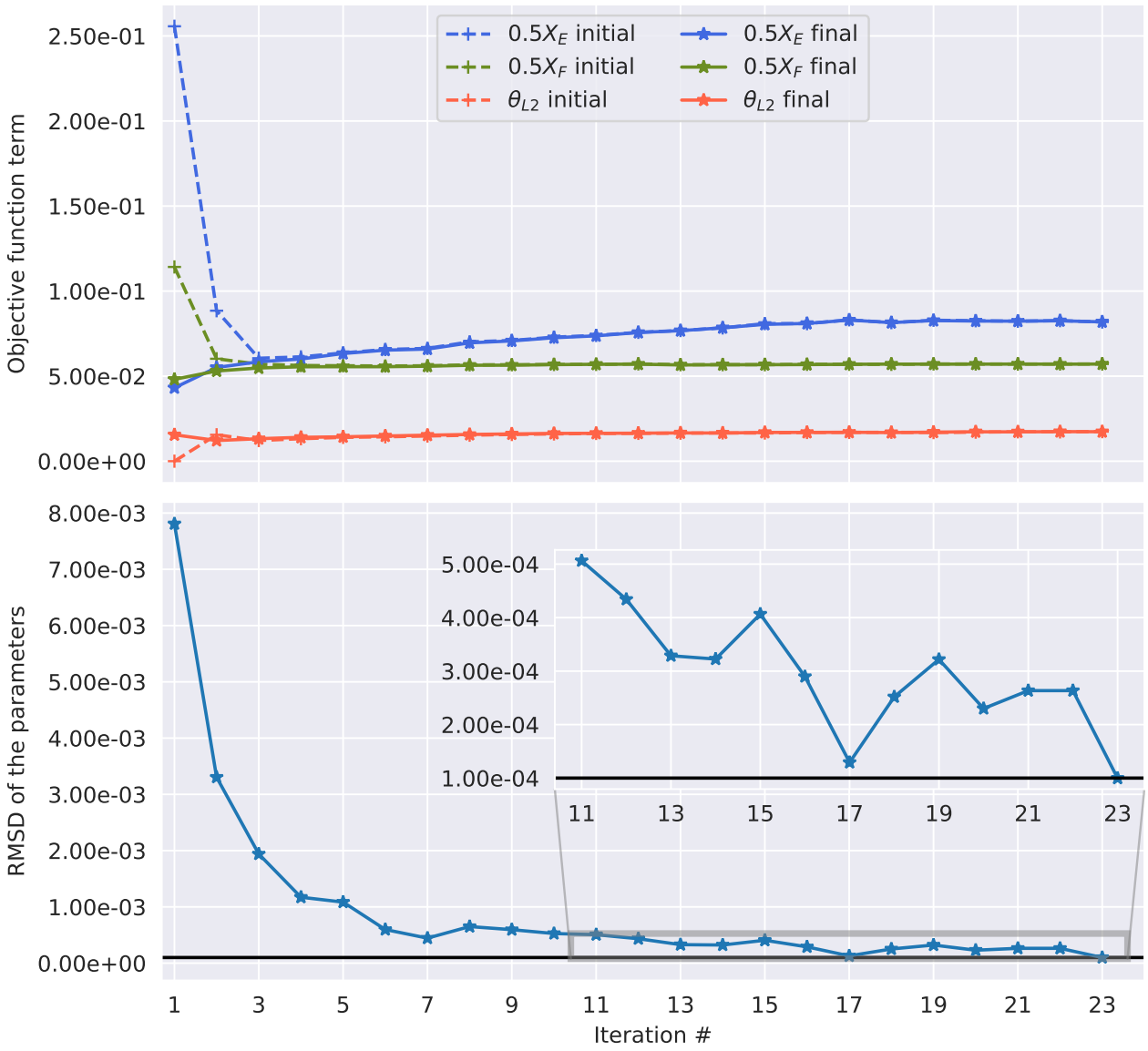


Figure 15: Top panel: Plot of the values of each term included in the objective function at the beginning (dashed lines) and end (solid lines) of each iteration. X_E corresponds to the energy term, X_F to the forces term, and θ_{L2} to the regularization term. Bottom panel: Plot of the RMSD of the parameters as a function of the number of the iteration number.

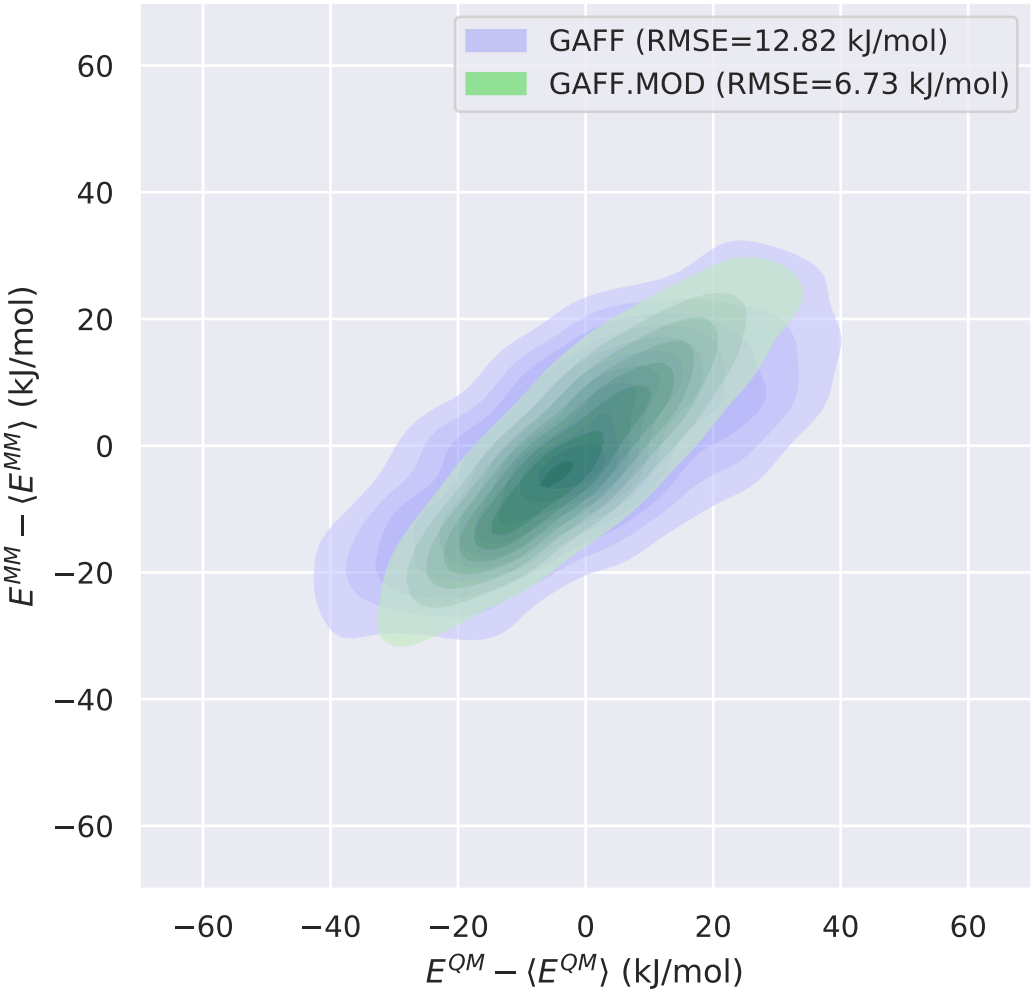


Figure 16: Correlation between the QM energies and the MM energies of caffeine before (GAFF) and after (GAFF.MOD) the adaptive re-parametrization procedure. Each data sets consists of 1000 configurations generated through a short MD simulation that used the respective FF.

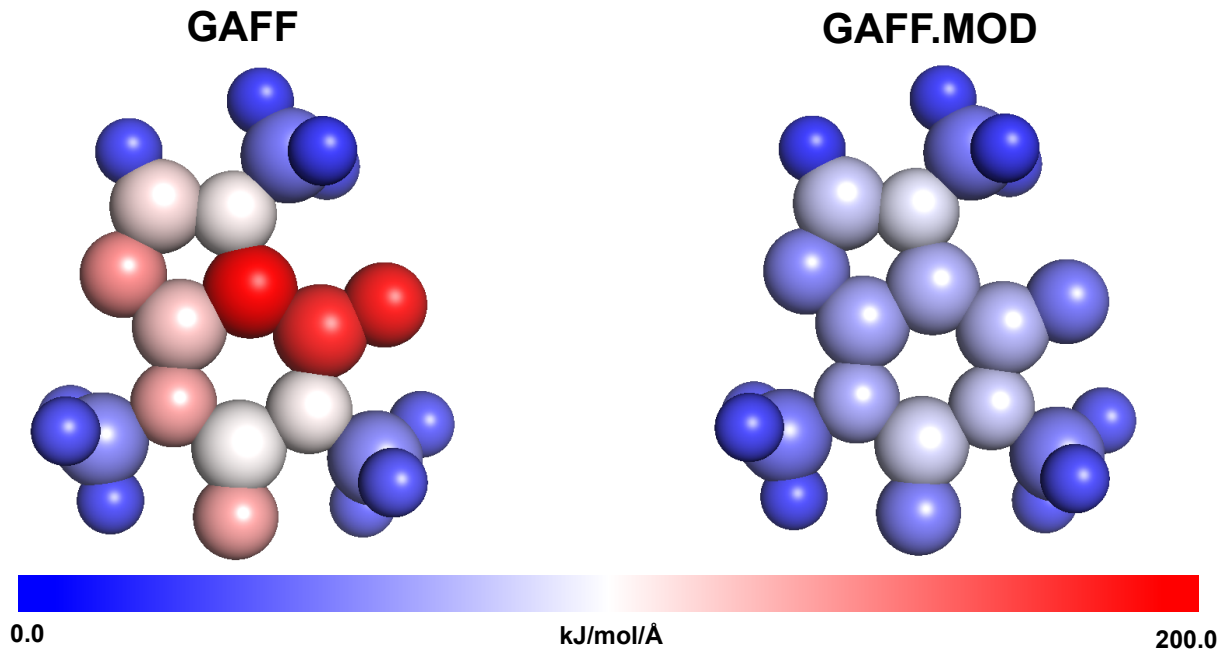


Figure 17: Atomic force errors before (GAFF, left) and after (GAFF.MOD, right) re-parametrization. The average RMSE of the atomic forces improved from 83.61 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ (GAFF) to 50.95 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ after re-parametrization (GAFF.MOD).

Conclusions

We presented ParaMol, software that has the capability of re-parametrizing class I force-field with a special focus on drug-like molecules. As explained and demonstrated in the examples provided, ParaMol has many automated capabilities that allow the re-parametrization of molecules through the use of different protocols. Its application may have implications in different areas of chemistry with biological relevance that require FFs with high levels of accuracy. The results obtained demonstrate that, within the constraints of the functional form, the methodologies implemented in ParaMol were able to derive near-ideal parameters for small organic molecules.

We have shown that the use of MM-relaxed dihedral scans is a robust way to re-parametrize the parameters of dihedrals and that this methodology is not very sensitive to the weighting method, yet it requires strong-to-intermediate regularization strengths. On

the other hand, since fittings to QM-relaxed dihedrals scans are critically dependent on the intra-molecular FF parameters, they substantially bias the derived FF parameters and, therefore, their use should be avoided. Furthermore, configurational ensembles generated through standard MM simulation methods may also be used as parametrization data sets, even though they make the optimizations more sensitive to the weighting method. In this context, the best results were obtained when using non-Boltzmann weighting, which proved to be the most reliable weighting scheme, despite its tendency to overestimate transition state energies and underestimate fluctuations. Moreover, Boltzmann weighting, which emphasizes the description of QM minima, tends to over-fit low energy regions of the PES at the cost of poorly describing the remainder of the energy landscape. Hence, it requires strong regularization to produce FFs that can be potentially used in MM modelling. Finally, since uniform weighting allows for positive and negative $E^{MM} - E^{QM}$ values, it is prone to the creation of asymmetries in the PES, which often lead to spurious minima due to artificially large thermodynamics weights and poor description of under-represented configurations (*e.g.*, transition states). Similarly to Boltzmann weighting, uniform weighting also requires strong regularization to mitigate some of these undesirable features.

It is also worthwhile mentioning that the sampling of spurious conformations is a common issue that arises when re-parametrizing FFs, which may occur whenever not strong enough regularization is employed or if the undesired sampled geometries are absent in the data set used to perform the fitting. A possible solution for this issue is to further re-optimize the FFs including the spurious conformations, such that the optimization procedure has information about them. Owing to the features of non-Boltzmann weighting, it is the indicated method to apply in these situations since, if anything, it tends to overestimate barrier heights and/or equilibrium energies - attributes that, ultimately, prevent the over-sampling of "artificial" geometries.

When using configurational ensembles as parametrization data sets, temperatures in a range between 300 and 500 K should be applied if using Boltzmann or non-Boltzmann weight-

ing, as progressively employing higher temperatures leads to results that become gradually similar to the ones that are obtained when using uniform weighting (which does not perform particularly well in this setting). Alternatively, it is also possible to resort to the ParaMol's soft dihedral parametrization task, which identifies and concomitantly parametrizes all dihedrals associated with a molecule's rotatable bonds. This method has a computational cost significantly lower than the configurational ensemble approach, whilst inheriting all features implicit to dihedral scans. Finally, adaptive parametrization is also an attractive and useful way to optimize parameters as it combines in one protocol self-consistent sampling and parametrization.

In general, most of the parametrization routines implemented in ParaMol can be performed automatically. However, care has to be taken when performing parametrizations using a non-linear, iterative optimizer at the expense of the LLS fitting approach, as the former may become trapped in local minima, whereas the latter is deterministic and ensures obtaining the global minimum. Consequently, whenever possible and suitable, the LLS solution is preferred. Moreover, manual quality checks may be required to identify poor data and outliers in the data set used in the parametrization, which is of particular importance since most of the FF optimization problems arise as a result of low-quality fitting data.

Owing to its potential, we suggest that ParaMol could be introduced as a routine step in the protocol normally employed to parametrize drug molecules for MM simulations. We hope that this software is useful for the drug-design community and any issue encountered during its use is encouraged to be reported to the GitHub page. The software is licensed under the MIT open source licence. The code is available at GitHub at `https://github.com/JMorado/ParaMol` and the documentation can be found at `https://paramol.readthedocs.io`.

Acknowledgement

The authors would like to thank Dr. Willem Nissink for the very helpful comments and suggestions for improving the paper. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. The authors also thank AstraZeneca for funding this study, and are grateful for the support from the EPSRC Centre for Doctoral Training, Theory and Modelling in Chemical Sciences, under Grant EP/L015722/1.

Supporting Information Available

- Formulae used to calculated RMSEs, diagrams of the workflow of the soft dihedral parametrization tasks, CSD data, SCC-DFTB-D3 results of the norfloxacin analog and caffeine, and configurational distributions of aspirin.
- Input and output files associated with the examples reported.

References

(1) Huggins, D. J.; Biggin, P. C.; Dämgen, M. A.; Essex, J. W.; Harris, S. A.; Henchman, R. H.; Khalid, S.; Kuzmanic, A.; Laughton, C. A.; Michel, J.; et al., Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2019**, *9*, e1393.

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

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

(4) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

(5) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

(6) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(7) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(8) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(9) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(10) Oostenbrink, C.; Villa, A.; Mark, A.; van Gunsteren, W. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656–1676, J CT.

(11) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2009**, NA–NA.

(12) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al., OPLS3: A Force Field Providing

Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

(13) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(14) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(15) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(16) Hudson, P. S.; Boresch, S.; Rogers, D. M.; Woodcock, H. L. Accelerating QM/MM Free Energy Computations via Intramolecular Force Matching. *J. Chem. Theory Comput.* **2018**, *14*, 6327–6335.

(17) Hudson, P. S.; Han, K.; Woodcock, H. L.; Brooks, B. R. Force matching as a stepping stone to QM/MM CB[8] host/guest binding free energies: a SAMPL6 cautionary tale. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 983–999.

(18) Giese, T. J.; York, D. M. Development of a Robust Indirect Approach for MM QM Free Energy Calculations That Combines Force-Matched Reference Potential and Bennett's Acceptance Ratio Methods. *J. Chem. Theory Comput.* **2019**, *15*, 5543–5562.

(19) Betz, R. M.; Walker, R. C. Paramfit: Automated optimization of force field parameters for molecular dynamics simulations. *J. Comput. Chem.* **2015**, *36*, 79–87.

(20) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid parameterization of small molecules using the force field toolkit. *J. Comput. Chem.* **2013**, *34*, 2757–2770.

(21) Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, *9*, 3543–3556.

(22) Doemer, M.; Maurer, P.; Campomanes, P.; Tavernelli, I.; Rothlisberger, U. Generalized QM/MM Force Matching Approach Applied to the 11-cis Protonated Schiff Base Chromophore of Rhodopsin. *J. Chem. Theory Comput.* **2014**, *10*, 412–422.

(23) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.

(24) 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.

(25) Wang, L.-P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *J. Chem. Theory Comput.* **2013**, *9*, 452–460.

(26) Xu, P.; Guidez, E. B.; Bertoni, C.; Gordon, M. S. Perspective: Ab initio force field methods derived from quantum mechanics. *J. Chem. Phys.* **2018**, *148*, 090901.

(27) Dubbeldam, D.; Walton, K. S.; Vlugt, T. J. H.; Calero, S. Design, Parameterization, and Implementation of Atomic Force Fields for Adsorption in Nanoporous Materials. *Adv. Theory Simul.* **2019**, *2*, 1900135.

(28) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.

(29) Lii, J. H.; Allinger, N. L. Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *J. Am. Chem. Soc.* **1989**, *111*, 8566–8575.

(30) Lii, J. H.; Allinger, N. L. Molecular mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111*, 8576–8582.

(31) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comput. Chem.* **1994**, *15*, 162–182.

(32) Ósk Jónsdóttir, S.; Rasmussen, K. The consistent force field. Part 6: an optimized set of potential energy functions for primary amines. *New J. Chem.* **2000**, *24*, 243–247.

(33) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7*, 3143–3161.

(34) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; et al., Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(35) Patel, S.; Brooks, C. L. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **2004**, *25*, 1–16.

(36) Vanommeslaeghe, K.; MacKerell, A. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*, 861–871.

(37) Rick, S. W.; Stuart, S. J.; Berne, B. J. Dynamical fluctuating charge force fields: Application to liquid water. *J. Chem. Phys.* **1994**, *101*, 6141–6156.

(38) Rick, S. W.; Berne, B. J. Dynamical Fluctuating Charge Force Fields: The Aqueous Solvation of Amides. *J. Am. Chem. Soc.* **1996**, *118*, 672–679.

(39) Ando, K. A stable fluctuating-charge polarizable model for molecular dynamics simulations: Application to aqueous electron transfers. *J. Chem. Phys.* **2001**, *115*, 5228–5237.

(40) Ercolessi, F.; Adams, J. B. Interatomic Potentials from First-Principles Calculations: The Force-Matching Method. *Europhys. Lett.* **1994**, *26*, 583–588.

(41) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **1990**, *11*, 431–439.

(42) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(43) Reynolds, C. A.; Essex, J. W.; Richards, W. G. Atomic charges for variable molecular conformations. *J. Am. Chem. Soc.* **1992**, *114*, 9075–9079.

(44) Burger, S. K.; Ayers, P. W.; Schofield, J. Efficient parameterization of torsional terms for force fields. *J. Comput. Chem.* **2014**, *35*, 1438–1445.

(45) Vanommeslaeghe, K.; Yang, M.; MacKerell, A. D. Robustness in the fitting of molecular mechanics parameters. *J. Comput. Chem.* **2015**, *36*, 1083–1101.

(46) Hopkins, C. W.; Roitberg, A. E. Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem. *J. Chem. Inf. Model.* **2014**, *54*, 1978–1986.

(47) Guvench, O.; MacKerell, A. D. Automated conformational energy fitting for force-field development. *J. Mol. Model.* **2008**, *14*, 667–679.

(48) 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 Refer-

ence Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

(49) Wang, L.-P.; Van Voorhis, T. Communication: Hybrid ensembles for improved force matching. *J. Chem. Phys.* **2010**, *133*, 231101.

(50) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J. Phys. Chem. B* **2017**, *121*, 4023–4039.

(51) Jones, E.; Oliphant, T.; Peterson, P., et al. SciPy: Open source scientific tools for Python. Accessed 1st of August, 2020; `http://www.scipy.org/`.

(52) Powell, M. J. A view of algorithms for optimization without derivatives. *Mathematics Today - Bulletin of the Institute of Mathematics and Its Applications* **2007**, *43*, 12.

(53) Kraft, D. *A Software Package for Sequential Quadratic Programming*; Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht; Wiss. Berichtswesen d. DFVLR, 1988.

(54) Conn, A. R.; Gould, N. I. M.; Toint, P. L. *Trust-region methods*; MPS-SIAM series on optimization; Society for Industrial and Applied Mathematics, 2000.

(55) Do, H.; Troisi, A. Developing accurate molecular mechanics force fields for conjugated molecular systems. *Phys. Chem. Chem. Phys.* **2015**, *17*, 25123–25132.

(56) Claridge, K.; Troisi, A. Developing Consistent Molecular Dynamics Force Fields for Biological Chromophores via Force Matching. *J. Phys. Chem. B* **2019**, *123*, 428–438.

(57) Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. Proceedings of COMPSTAT'2010. Heidelberg, 2010; pp 177–186.

(58) Bertsimas, D.; Tsitsiklis, J. Simulated Annealing. *Stat. Sci.* **1993**, *8*, 10–15.

(59) *Trust-region methods*; Springer Series in Operations Research and Financial Engineering; Springer New York, 2006.

(60) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, 1–17.

(61) Hourahine, B.; Aradi, B.; Blum, V.; Bonafé, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshaye, M. Y.; Dumitrică, T.; Dominguez, A.; et al., DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

(62) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method †. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(63) Larsen, A. H. et al. The atomic simulation environment - a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(64) Harris, C. R. et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362.

(65) Landrum, G. RDKit: Open-source cheminformatics. Accessed 1st of August, 2020; `http://www.rdkit.org`.

(66) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(67) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(68) Sabatini, R.; Gorni, T.; de Gironcoli, S. Nonlocal van der Waals density functional made simple and efficient. *Phys. Rev. B* **2013**, *87*, 041108.

(69) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *J. Chem. Phys.* **2010**, *133*, 244103.

(70) Prentice, J. C. A.; Aarons, J.; Womack, J. C.; Allen, A. E. A.; Andrinopoulos, L.; Anton, L.; Bell, R. A.; Bhandari, A.; Bramley, G. A.; Charlton, R. J.; et al., The ONETEP linear-scaling density functional theory program. *J. Chem. Phys.* **2020**, *152*, 174111.

(71) Womack, J. C.; Mardirossian, N.; Head-Gordon, M.; Skylaris, C.-K. Self-consistent implementation of meta-GGA functionals for the ONETEP linear-scaling electronic structure package. *J. Chem. Phys.* **2016**, *145*, 204114.

(72) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.* **2005**, *122*, 084119.

(73) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; et al., Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185–3197.

(74) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(75) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

(76) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. Energetics and structure of glycine and alanine based model peptides: Approximate SCC-DFTB, AM1 and PM3 methods in comparison with DFT, HF and MP2 calculations. *Chem. Phys.* **2001**, *263*, 203–219.

(77) Elstner, M. The SCC-DFTB method and its application to biological systems. *Theor. Chem. Acc.* **2006**, *116*, 316–325.

(78) Hujo, W.; Grimme, S. Performance of the van der Waals Density Functional VV10 and (hybrid)GGA Variants for Thermochemistry and Noncovalent Interactions. *J. Chem. Theory Comput.* **2011**, *7*, 3866–3871.

(79) Brémond, ; Savarese, M.; Su, N. Q.; Pérez-Jiménez, J.; Xu, X.; Sancho-García, J. C.; Adamo, C. Benchmarking Density Functionals on Structural Parameters of Small-/Medium-Sized Organic Molecules. *J. Chem. Theory Comput.* **2016**, *12*, 459–465.

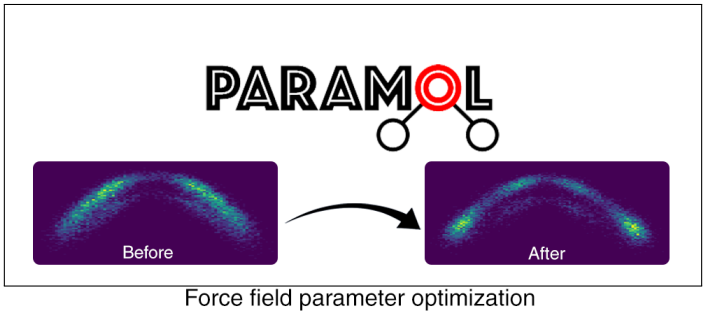
(80) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(81) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica, Section B: Structural Science* **2016**, *72*, 171–179.

(82) Dressman, J. B.; Nair, A.; Abrahamsson, B.; Barends, D. M.; Groot, D.; Kopp, S.; Langguth, P.; Polli, J. E.; Shah, V. P.; Zimmer, M. Biowaiver Monograph for Immediate-Release Solid Oral Dosage Forms: Acetylsalicylic Acid. *J. Pharm. Sci.* **2012**, *101*, 2653–2667.

(83) Jeffrey, G. A. *An introduction to hydrogen bonding*; Topics in physical chemistry; Oxford University Press, 1997.

Graphical TOC Entry



Force field parameter optimization

$$V = \sum_{bonds} K_b(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$