Published in final edited form as:
J Chem Inf Model. 2012 December 21; 52(12): 3155–3168. doi:10.1021/ci3003649.

Automation of the CHARMM General Force Field (CGenFF) II: Assignment of bonded parameters and partial atomic charges

K. Vanommeslaeghe, **E. Prabhu Raman**, and **A. D. MacKerell Jr**[*]
Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201

Abstract

Molecular mechanics force fields are widely used in computer-aided drug design for the study of drug candidates interacting with biological systems. In these simulations, the biological part is typically represented by a specialized biomolecular force field, while the drug is represented by a matching general (organic) force field. In order to apply these general force fields to an arbitrary drug-like molecule, functionality for assignment of atom types, parameters and partial atomic charges is required. In the present article, algorithms for the assignment of parameters and charges for the CHARMM General Force Field (CGenFF) are presented. These algorithms rely on the existing parameters and charges that were determined as part of the parametrization of the force field. Bonded parameters are assigned based on the similarity between the atom types that define said parameters, while charges are determined using an extended bond-charge increment scheme. Charge increments were optimized to reproduce the charges on model compounds that were part of the parametrization of the force field. A "penalty score" is returned for every bonded parameter and charge, allowing the user to quickly and conveniently assess the quality of the force field representation of different parts of the compound of interest. Case studies are presented to clarify the functioning of the algorithms and the significance of their output data.

Introduction

Part I of this series[1] described the CHARMM General Force Field (CGenFF)[2] atom typer, which deterministically assigns atom types to an arbitrary organic molecule using a programmable decision tree. In the present paper, a set of algorithms is presented to assign bonded parameters and charges to such a molecule. As discussed in part I, an essential component of a molecular mechanics force field is the parameter set, including the bonded parameters associated with combinations of atom types as well as Lennard-Jones (LJ) and charge parameters required to treat nonbonded interactions. Specifically in the case of Class I additive force fields (equation 1) such as CHARMM,[3] for every covalently bound pair of atom types, a reference distance and a force constant are defined (b_o and K_b in equation 1). Similarly, valence angle parameters K_Θ and Θ_0 are defined for every covalently bonded combination of three atom types and dihedral parameters K_φ, n and δ for every covalently bonded combination

[*] To whom correspondence should be addressed: alex@outerbanks.umaryland.edu.

Supporting Information

The following items are available as Supporting Information:

- A list of bond, angle and dihedral charge increments that were constrained during the charge increment optimization.
- Toppar stream files for the case studies as generated by the CGenFF program.
- A table with minimized bond length and angles for N-benzylacetamide (**3**), analogous to Table 2.

This material is available free of charge via the Internet at http://pubs.acs.org.

of four atom types. Also, where necessary, Urey-Bradley terms K_{UB} and $r_{1,3;0}$ and improper dihedral parameters K_φ and φ_0 are defined for select combinations of three and four atom types, respectively.

Intramolecular (internal, bonded terms)

$$\sum_{bonds} K_b(b-b_0)^2+\sum_{angles} K_\theta(\theta-\theta_0)^2+\sum_{dihedrals} K_\phi(1+\cos(n\phi-\delta))+\sum_{improper\ dihedrals} K_\varphi(\varphi-\varphi_0)^2+\sum_{Urey-Bradley} K_{UB}(r_{1,3}-r_{1,3;0})^2$$

Intermolecular (external, nonbonded terms)

$$\sum_{nonbonded} \frac{q_i q_j}{4\pi D r_{ij}}+\varepsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12}-2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] \quad (1)$$

As the number of atom types for which bonded parameter are required increases, it becomes increasingly unrealistic to optimize parameters for every possible combination of atom types; while CGenFF has a good coverage of individual chemical groups,[2] which continues to grow, organic molecules consist of virtually limitless combinations of chemical groups. Consequently, a typical drug-like compound will contain combinations of atom types that do not match existing bonded parameters in the force field. In the context of the program CHARMM,[4] these cases are called “missing parameters” because bonded parameters are required for every combination of atom types in the molecule. As for the nonbonded parameters, combining rules[5] are commonly used to derive the Lennard-Jones radius R_{ij} and well depth $\boldsymbol{\varepsilon}_{ij}$ (see equation 1) for each i,j atom pair in a chemical system from the per-atom quantities R_i, R_j, $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_j$, which in turn are associated with individual atom types. Therefore, the atom typing functionality described in Part I automatically takes care of the Lennard-Jones parameters. This leaves the charges q_i to be assigned in addition to the bonded parameters.

The problem of assigning bonded parameters and charges to an organic molecule is far from new, and a wide variety of approaches have been applied in organic force fields such as the Tripos force field,[6] CVFF,[7] UFF,[8] CFF93,[9] Allinger's MM2,[10] MM3[11–13] and MM4[14] force fields and Momany and Rone’s commercial CHARMm force field[15] (not to be confused with the academic CHARMM force field; note the different capitalization). Of particular note are the General Amber Force Field (GAFF)[16] and its associated antechamber toolkit,[17] which marked the first extension of a specialized biomolecular force field (ie. AMBER[18, 19]) to organic molecules, and MMFF94,[20–22] which provided inspiration for the present charge assignment scheme. Accordingly, the present charge assignment scheme is empirical, as are most of the aforementioned charge assignment approaches. Notable exceptions are AMBER and OPLS, where charge assignment schemes based on a QM electron density (typically AM1 or PM3) have been proposed for small organic molecules.[17, 23–27] While the resulting charges capture polarization and quantum effects and therefore are typically significantly more accurate than charges generated by more empirical assignment schemes,[28–30] they are also computationally significantly more expensive, which may become a bottleneck in high-throughput applications involving millions of molecules.

Approach and algorithms

To obtain missing parameters in CGenFF, existing parameters that were determined as part of the parametrization of CGenFF are identified by analogy based on the similarity between the atom types that define the parameters. Charges are assigned using a bond-charge increment scheme that is conceptually similar to the one implemented in MMFF94.[20–22] The charge increments, which function as parameters in the charge assignment algorithm, are optimized to reproduce the charges on a training set of model compounds that were part of the parametrization of the force field. A noteworthy innovation in the presented charge assignment scheme is the fact that apart from the single charge increment associated with each possible bond (as defined by 2 atom types), there are 2 charge increments associated with each possible angle and 3 charge increments with each possible dihedral angle (as defined by 3 and 4 atom types, respectively). The combined functionality in the CGenFF atom typer[1] and parameter assignment scheme makes it possible to apply the force field to an arbitrary molecule in an automated fashion. Notably, as a measure of the accuracy of the approximation, a "penalty score" is returned to the user for every bonded parameter and charge, which can be used to guide selective optimization of parameters and charges.

Assignment of parameters by analogy

The assignment of missing parameters is based on penalty scores quantifying the dissimilarity between atom types. The penalty scores are based on an $n \times n$ penalty matrix, where n is the number of atom types. It should be noted that this matrix is not necessarily symmetric; substituting a very specialized atom type with a more generic variant typically has a lower penalty score than vice versa because it carries a lower risk of catastrophically inconsistent behavior. Before discussing the construction of the penalty matrix, we give a brief overview of the computational methods for assigning parameters by analogy, which provides a foundation for assignment of the penalty scores.

Total penalty score (TPS) calculation—The total penalty score between two parameters (ie. a missing parameter and an existing parameter) can be calculated as the sum of the penalties for substituting the atom types defining the one parameter with the atom types in the second parameter. Thus, for every missing parameter, the program for assignment of parameters by analogy is tasked with finding the existing parameter in the parameter file (or parameter database) with the lowest TPS with respect to the missing parameter. This is accomplished by simply iterating through the parameter database, calculating the TPS for every parameter and returning the parameter with the lowest penalty. This process is sped up by the fact that both the penalty matrix and the data structure containing the parameters are collated by atom type, reducing the number and computational expense of lookup operations in the penalty matrix. The speed gain achieved by collating the parameters is a result of the fact that a parameter in a majority of cases will only differ from the previous parameter in the collated list by its last atom type(s). Thus, in the example of an angle parameter defined by atom types A-B-C, the algorithm can be made to keep track of the penalty of the first atom type (A) and the sum of the penalties of the first two atom types (A+B). If it is determined that only the last atom type is different (C'), only this atom type needs to be looked up in the penalty matrix and its penalty can be added to the latter sum. Similarly, if the two last atom types are different (B'–C"), only those two atom types must be looked up and their sum added to the previously determined penalty for A, thus substantially improving performance. Finally, for every existing parameter, the lowest of two TPSs should be used: the TPS between the "ordered" missing parameter A-B-C and the existing parameter, and the same TPS for the (physically equivalent) "reverse ordered" missing parameter C-B-A, as the former is not guaranteed to be always lower than the latter. All resulting parameters are stored in a binary search tree, which makes it easy to determine whether a missing parameter has been encountered and assigned before.

Penalty matrix definition: rule file—The penalty scores for atom substitutions ideally should reflect the chemical characteristics on which the definitions of the atom types were based. However, as the number of atom types in CGenFF is approximately 150 and is likely to increase with the parametrization of new chemical groups, directly populating the penalty matrix manually is not feasible, or would at the very least be extremely error-prone. Instead, the full matrix is constructed from a smaller number of submatrices and offsets, which are defined in a second section of the rule file used to assign atom types, as described in the preceding article.[1] This "parameter assignment" section has a formally similar layout as the atom typing rules, although it is processed quite differently. Atom types are divided in a tree-like structure, with categories and subcategories, delineated by a "cat" keyword defining the name of the category and an "end keyword". Every entry in a category can either be an atom type ("typ") or a reference to a subcategory ("sub"), the name of which is followed by a colon. The difference with the atom typing rules lies in the keywords after the colon, the first of which is "pri" (short for "priority"), followed by the penalty offset added to the atom type substitution's total penalty score when entering the subcategory through a "sub" keyword in a hierarchically higher category (this will be clarified by an example in the next paragraph). This is followed by one or more "alt" keywords detailing alternative atom types or subcategories within the same category. The number of "alt" entries is equal to the number of rules in the current category minus one; every "alt" keyword is followed by the name of another ("typ" or "sub") rule in the same category and the penalty score for substituting the current type with the named alternative type. Finally, the rule ends with an "up" keyword, followed by the penalty offset for going one level up in the hierarchy.

As an example, Table 1 contains a short extract of the rule file that handles *sp*^{3} nitrogen atom types. In the category NG3, two subcategories are defined: NG3N and NG3P for neutral and positively charged *sp*^{3} nitrogen atom types, respectively. Consider for example the protonated primary ammonium type NG3P3; substituting this atom type with NG3P2 (secondary ammonium), NG3P1 (tertiary ammonium) and NG3P0 (quaternary ammonium) types respectively incurs penalties of 1, 2 and 4. If these substitutions fail, the penalty for going up in the hierarchy is 8. Doing so brings us to the sub NG3P rule in the NG3 category, which specifies that substituting an atom in the NG3P subcategory with an atom in the NG3N subcategory carries a penalty of 2. The penalties are added, so the total penalty score so far is 8+2=10. After entering the category NG3N, the rule with the lowest "pri" penalty is picked, which by convention is the first rule in the category, in this case NG321 with a penalty of 0. Thus, the final penalty score for substituting NG3P3 with NG321 is 8+2+0=10. Similarly, the penalty score for substituting NG3P3 with NG311 is 8+2+0.5=10.5, and so forth. This scheme allows filling the full penalty matrix using a rule file of manageable size; it is left to the writer of the rule file to make a trade-off between specifying more specific (and thus potentially more accurate) values by making fewer larger categories on the one hand, and having a smaller rule file (and thus less potential for "programming" error) by creating more smaller categories on the other hand. However, as the number of penalties in a category goes with the square of the number of rules in that category, making large categories quickly becomes unwieldy. Thus, the largest category in the parameter section of the current version of the CGenFF rule file contains 25 hydrogen atom types, and the penalty scores in this category were machine-generated based on reference bond lengths in the CGenFF parameter file. The second largest and largest manually generated category contains 10 *sp*^{3} oxygen atom types.

Most of the penalties in the rule file are chosen manually; it is a relative rare occurrence that an objective measure for similarity could be found that performed satisfactory for the purpose of assigning parameters by analogy. During the process of "programming" the penalty matrix, certain constraints were respected; for instance, values at a certain level in the tree are all in the same order of magnitude. This is illustrated in Table 1: the NG3 category has an "up" penalty of 12 for both "sub" rules, while both of its subcategories have "up" penalties of 8 for all their atom types, and penalties within each category are lower than this "up" value. Also, the penalty for substituting atom A for atom B was usually kept similar to the penalty for substituting atom B for atom A. However, these rules were not always followed rigidly. For example, in cases where a very specific type is changed into a type that is clearly more general, the penalty was deliberately set lower than for the reverse substitution (ie. changing a general type into a very specific one). In the end, what mattered more than any constraint in the construction of the penalty matrix was its ability to assign sensible parameters by analogy, and a large part of the effort was focused on validating their ability to assign parameters by analogy. This validation also gave rise to the below recommendations to validate and re-optimize parameters with TPS values above certain thresholds. Most of this work was based on chemical intuition, as it is not straightforward to do this in a systematic or quantitative way. Indeed, CGenFF internal parameters are validated against MP2 target data, which is relatively expensive to compute, and it is not trivial to devise a measure to objectively compare their performance when applied on different molecules. The present paper contains two case studies; a more systematic validation study is the subject of future work.

The whole rule tree is read into a temporary data structure and checked for consistency by a tree-walking algorithm that also replaces all atom type and rule name strings with pointers. Then, a different tree-walking algorithm is started at the first “typ” rule and exhaustively visits all other rules in the tree in a similar fashion as outlined in the above example, filling in one line of the penalty matrix; the aforementioned pointers help keep the computational cost of this process well within acceptable limits. The same procedure is repeated for all “typ” rules to complete the penalty matrix. The rows and columns of the penalty matrix are then respectively sorted by atom type and by penalty.

Differential treatment of inner and outer atoms—To improve the algorithm’s capability to pick good analogies, two extra features were added to the basic algorithm described in the preceding paragraphs:

1. The penalties for substituting the central atom in an angle or improper dihedral and the two inner atoms in a dihedral are multiplied by 10 because changing these inner atoms has a much higher impact than changing the outer atoms. The same multiplication factor is applied to the penalties for substituting atoms in bond parameters because the impact of changing atoms in a bond parameter is typically comparable to changing the inner atoms in an angle or dihedral.

2. The penalty matrix used for these “inner atoms” (including the atoms in a bond parameter) will henceforward be called the “bonded matrix”, in contrast to the “nonbonded matrix” that is used for the two “outer atoms” in angle and dihedral parameters. Specifically, the bonded matrix focuses more on bonded properties (valence and hybridization state) while the nonbonded matrix is more aimed at describing discrepancies in nonbonded properties (electrostatic and Van der Waals). For example, in the definition of the bonded matrix in the rule file, the highest category in the hierarchy differentiates subcategories by hybridization state, which are subsequently divided by period (ie. row in the periodic table), while in the nonbonded matrix, the period has precedence over the hybridization state. The use of these two different matrices can be rationalized by considering that the potential associated with a dihedral defined by atoms 1–2–3–4 is typically the result of a combination of nonbonded 1–4 interactions and the bond order of the 2–3 bond and the hybridization states of atoms 2 and 3, which are bonded properties. A similar observation can be made about angle potentials. However, this reasoning does not apply to improper dihedrals, where better results were obtained by using the bonded matrix for all 4 atoms (although the aforementioned scaling factor only applies to the central atom).

Problem: strained rings—The algorithm as described above was made available for public beta testing as part of versions 0.9.0 and 0.9.1 of the CGenFF program. During subsequent testing, an important shortcoming was identified; see the 2-(dimethylcarbamoyl)indole (compound **2** in Figure 1) case study below. Consider a planar 5-membered ring as depicted in Figure 2; geometry dictates that the sum of the ring’s inner angles α must be 540° and that the sum of the angles α+β+β around each of the 5 trigonal planar centers must be 360°. Indeed, in practice, pyrrole and other 5-membered aromatic heterocycles have inner angles α that are relatively close to the ideal 540° / 5 = 108° and corresponding outer angles β close to (360° − 108°) / 2 = 126°. In certain cases where the parameter assignment algorithm is given a substituted 5-membered ring in which one of the outer angles corresponds to a “missing parameter”, such as the C10-C2-C3 angle in compound **2**, it will use an in-ring parameter as a source of analogy (or *vice versa*), which leads to an error of (126° − 108°) = 18° in the reference angle and an unacceptable deformation in the geometry. Indeed, in compound **2**, C10 is an *sp^2* carbon, which exhibits significant similarity to an *sp^2* carbon in a 5-membered ring from a nonbonded perspective. Therefore, the parameter assignment algorithm used the angle parameter defined by atom types CG2R51 CG2R51 CG2R51 as a source of analogy for CG2O1 CG2R51 CG2R51 (and CG2R51 CG2R51 NG2R51 for the CG2O1 CG2R51 NG2R51 parameter on the C10-C2-N1 angle), giving rise to the 19.7° deviation in Table 2 and to a pronounced deviation from planarity (φ(C1-C3-N1-C10) = 26.2°). Although the penalty score for this particular case was 26.5, which indicates that validation is required, similar cases were found with penalties lower than 10, the “recommended validation threshold”. As planar 5-membered rings are ubiquitous in drug-like molecules, and this phenomenon can be shown to be even more pronounced in 3- and 4-membered rings, it seriously jeopardized the applicability of the parameter assignment algorithm. We first attempted to remedy this shortcoming by increasing the penalties for substitutions between atoms types for rings with different sizes and non-cyclic atom types, but it was found that a very large increase in the penalties was required in order to eliminate the problem, and that this large increase considerably affected the quality of other parameters by analogy for ring substitutions. For example, before increasing the penalties, parameters with a 6-membered ring atom type for C2 were used as a source of analogy for the C2-C10-N11 and C2-C10-O10 angles and were given a justifiably low penalty. Penalties that are sufficiently high to cure the problem discussed above would make this impossible.

Solution: bond groups—As changing the penalty scores did not lead to an acceptable solution, an extra “bond group” term was introduced in the penalty function starting from version 0.9.6 of the CGenFF program. Table 3 shows the definitions of these bond groups as they occur in the rule file; here, a bond group consists of the keyword “bgrp” followed by a penalty and a list of one or more atom types. For the purpose of assigning bond group penalties, bond, angle and dihedral *parameters* are considered to contain 1, 2 and 3 *virtual* covalent bonds, respectively. For example, the angle parameter CG2O1 CG2R51 CG2R51 discussed above contains the 2 virtual bonds CG2O1–CG2R51 and CG2R51–CG2R51. For each virtual covalent bond in a missing parameter, the bond is member of a given bond group if both atom types are members of that bond group; this way, a virtual bond can be member of 0, 1, 2 or more bond groups at the same time. In the above example, the virtual bond CG2R51–CG2R51 is member of both the bond group for planar 5-membered rings and the bond group for all 5-membered rings in Table 3 because both atom types are members of both bond groups. Conversely, the virtual bond CG2O1–CG2R51 is not member of any bond groups because CG2O1 does not occur in any of these groups. When searching the parameter database for candidate parameters with optimal analogy to this missing parameter, bond group membership is compared pairwise between the virtual bonds in the missing and the candidate parameter. If the virtual bond under consideration in the missing parameter is a member of a bond group and the corresponding virtual bond in the candidate parameter is not a member of the same group, or *vice versa*, then that bond group’s penalty is added to the TPS. Such an approach severely penalizes, for example, the substitution of an endocyclic parameter for an exocylic term and vice versa as the bond group for a ring of a certain size includes all the endocyclic atom types that can occur in a ring of this size, and no other (exocyclic) atom types. Two extra features were added to the basic bond group functionality described in this paragraph:

1. Similar to the above, the bond group penalties for substituting the central bond in a dihedral parameter or any bond in an angle or bond parameter are multiplied by 10

because changing these bonds has a much higher impact than changing the outer bonds in a dihedral or the bonds in an improper.

2. The two first bond groups in the rule file are considered equivalent for the purpose of determining differences in bond group membership between two virtual bonds. This exception makes it possible to use the bond group feature to improve treatment of conjugated double bonds, as discussed below

It can now be seen how the bond group feature remedies the shortcoming in the treatment of strained rings; as mentioned above, the CG2R51 CG2R51 virtual bond is member of both 5-membered ring bond groups in Table 3, while the corresponding CG2O1 CG2R51 virtual bond in the parameter originally used as a source of analogy is member of neither, adding a bond group contribution of (20 + 20) × 10 = 400 to the TPS and thereby virtually making sure that a more appropriate parameter will be chosen as a source of analogy. At the same time, the usage of a 6-membered ring atom type for C2 as a source of analogy for the C2-C10-N11 and C2-C10-O10 angles does not incur any additional penalty because none of the virtual bonds involved are members of any bond group.

Additional applications of bond groups—In addition to fixing the original problem with strained rings, the bond group feature allows for a better treatment of biphenyls and conjugated double bonds. Specifically, as described in the preceding manuscript (part I),[1] there are 2 sets of atom types for conjugated double bonds, CG2DC1 CG2D1O CG25C1 CG251O and CG2DC2 CG2D2O CG25C2 CG252O, which are applied such that in a chain of conjugated double bonds, double bonded atoms are given atom types from the same set and single-bonded atoms get atom types from a different set. Initially,[*] it was necessary to apply very high penalties to substitutions involving these atom types in order to prevent single-bond parameters being assigned to double bonds and vice versa. As a side effect of these high penalties, favorable substitutions involving conjugated double bond atom types were also suppressed. For example, a missing parameter containing the virtual non-conjugated double bond CG2D1–CG2D1 could never be substituted with its "conjugated equivalents" CG2DC1–CG2DC1 and CG2DC2–CG2DC2, because penalties low enough to allow this substitution would make it equally likely to substitute the original virtual double bond with the virtual *single* bond CG2DC1–CG2DC2. The introduction of the first two bond groups in Table 3 (which, as mentioned above, are considered equivalent for the purpose of determining differences in bond group membership) effectively gives the parameter assignment algorithm explicit knowledge of the bond orders. Consequently, the penalties for substituting conjugated double bond atom types could be set very low, instead relying on the bond groups to penalize incorrect substitutions. Analogous observations were made for biphenyls.

Advantages and limitations—Finally, it should be noted that, compared to the "wildcards" used in a number of force fields, the current scheme offers a more exhaustive effort to identify the "best" parameter to be used when an exact match for a parameter is not present. It also offers the ability to generate penalty scores that may alert the user to possible limitations in the assigned parameters. Nevertheless, it should be emphasized that these penalty scores are based purely on the level of chemical similarity of the atom types in a parameter. Consequently, even in cases where there is a penalty of zero (ie. when the exact parameter is already available in CGenFF), the quality of that parameter for determining the physical properties of the molecule may be poor. For example, all the parameters for a given type of linker between two rings might be available in CGenFF. However, if the nature of the rings changes significantly without a corresponding change in the atom types across the ring (eg. the C-H carbon atoms in both 2-pyridone and the pyridinium ion carry the CG2R62 type, while the former is neutral and the

[*] In versions 0.9.0 and 0.9.1.

latter is charged), then the conformational properties of that linker are likely to change such that the associated dihedral parameters, with penalties of zero, may require optimization. This speaks to the limited transferability of empirical force field parameters and the need for the user to make decisions on the level of accuracy they need for system under study. Conversely, cases have been encountered with large penalties yet satisfactory accuracy, which can be understood by interpreting the penalty as a measure of the statistical imprecision or "spread" of the parameter assignment.

Charge assignment

Extended bond-charge increment scheme—Charges are assigned using a bond-charge increment scheme. From an atom-centric point of view, this can be formalized as $q_i = q_i^0 - \sum_j \beta_{ij}$ where q_i is the final partial charge on atom i, q_i^0 is the previously assigned formal charge,[1] and β_{ij} is the bond charge increment for the bond between atoms i and j, with $\beta_{ji} = -\beta_{ij}$. From a bond-centric point of view, this translates to the following pseudocode:

```
for each atom $i$ { $q_i$ = $q_i^0$}
for each bond $n$ between atoms $i$ and $j$ {
$q_i$ −= $\beta_{ij}(n)$ ;
$q_j$ += $\beta_{ij}(n)$ }
```

We extended this scheme to angle charge increments α and dihedral charge increments δ:

```
for each atom $i$ { $q_i$ = $q_i^0$ }
for each bond $n$ between atoms $i$ and $j$ {
$q_i$ −= $\beta_{ij}(n)$ ;
$q_j$ += $\beta_{ij}(n)$ }
for each angle $n$ between atoms $i$, $j$ and $k$ {
$q_i$ −= $\alpha_{ij}(n)$ ;
$q_j$ += $\alpha_{ij}(n)$ − $\alpha_{jk}(n)$ ;
$q_k$ += $\alpha_{jk}(n)$ }
for each dihedral $n$ between atoms $i$, $j$, $k$ and $l$ {
$q_i$ −= $\delta_{ij}(n)$ ;
$q_j$ += $\delta_{ij}(n)$ − $\delta_{jk}(n)$ ;
$q_k$ += $\delta_{jk}(n)$ − $\delta_{kl}(n)$ ;
$q_l$ += $\delta_{kl}(n)$ }
```

This is implemented by associating one charge increment β_{ij} with every bond in the parameter file, two charge increments α_{ij} and α_{jk} with every angle and three charge increments δ_{ij}, δ_{jk} and δ_{jk} with every dihedral. For every bond, angle or dihedral in the molecules, the charge increments associated with the corresponding parameter in the parameter file are used. If there exists no such parameter, charge increments from an analogous parameter are applied, using the same algorithm for finding the best analogy as described above, except that the "nonbonded matrix" is used for all atoms. As the algorithm for assigning parameters by analogy has the ability to match an existing parameter in which the atom types are in reverse order, care was taken to reverse the order and sign of the charge increments whenever such reversal occurs. It should be noted that the present algorithm only assigns correct symmetric charges to delocalized cases like amidinium, guanidinium and carboxylate when given symmetric formal charges q_i^0; therefore, it was necessary to include formal charge assignment in the atom typing rules, as discussed in the preceding manuscript (Part I).[1] Finally, when applying this charge assignment algorithm to a test set of molecules that were not part of the CGenFF parametrization,[1] it was found that charge increments associated with high-penalty dihedrals lead to a degradation in the quality of the predicted charges. Indeed, it is a known problem in force field development that transferability of parameters rapidly degrades with increasing number of atoms in the parameter's definition, and becomes problematic for dihedral parameters;[31, 32] accordingly, this non-transferabilty was shown to apply to the associated charge increments. This problem was solved by ignoring charge increments associated with dihedral parameters with a penalty score higher than an arbitrary limit, which is currently set to 50.

Charge increment optimization—The presence of charge increments associated with angles and dihedrals has the advantage of allowing charge effects to propagate over up to 3 bonds, which opens the possibility to capture inductive and (to a limited extent) mesomeric or resonance effects. Also, the presence of charge increments explicitly associated with transfered dihedral parameters may improve compatibility between these dihedral parameters and the 1–4 electrostatic interactions, which always has been one of the main factors limiting dihedral parameter transferability in general force fields. The disadvantage is that this leaves a very large number of charge increments to be optimized. As target data for this optimization, we used the 9681 charges on the 477 full model compounds currently in CGenFF. Since all partial charges in a molecule must sum up to the total charge of said molecule, this corresponds to 9681 − 477 = 9204 independent data points. For the actual optimization, three least-squares fits are performed:

- First, all bond charge increments for bonds between different atom types are optimized (the charge increment for a bond between two identical atom types is zero by definition). In total, 384 degrees of freedom were fit in this stage, and the final RMS charge deviation was 0.0394 *e*. The resulting bond charge increments are rounded to the third digit after the decimal point and a number of bond charge increments involving hydrogen atoms are set to CHARMM standard values in order to make the hydrogen charges consistent with the CHARMM charge assignment rules. Specifically, aliphatic hydrogen atoms are always assigned a charge of +0.09 *e*, except when they are located on a 5-membered ring aliphatic carbon atom directly adjacent to a positively charged nitrogen atom, where they are assigned a standard charge of +0.28 *e*. Similarly, hydrogen atoms bound to aromatic, methylene sp^2 (=CH_2) and methine sp^2 (=CH–) carbon atoms are given charges of +0.115, +0.21 and +0.15 *e*, respectively (see Table S1, Supporting Information). It should be noted that these changes amount to further rounding of the affected bond charge increments; since a vast majority of the molecules in the target data comply with the CHARMM charge assignment rules, the optimized bond charge increments were already at or very close to their ideal values before this adjustment, as testified by the 10^{-5} *e* increase in RMSD after adjustment.
- Next, charge increments for angles are optimized. As with the bonds, charge increments on symmetric angles are constrained at zero. The total number of degrees of freedom at this stage was 2112, and the final RMSD was 0.0174 *e*. Again, the charge increments are rounded to the third digit after the decimal point. Charge increments smaller than 0.0025 *e* that involve a hydrogen are further rounded to zero, as well as the charge increments on a number of select angles (see Table S1, Supporting Information), in order to enforce the aforementioned CHARMM hydrogen charge assignment rules. This raised the RMSD by 4×10^{-5} *e*.
- Finally, the optimization of the dihedral charge increments is performed, again constraining symmetric dihedrals to zero. Additionally, instead of setting a number

of selected charge increments at 0 *a posteriori* as was done for the angles, 7 select dihedrals were constrained *a priori* to 0 (see Table S1, Supporting Information). Indeed, as this is the final optimization, there are no "higher-order" charge increments that can be used to compensate for the non-perfect fit caused by zeroing charge increments *a posteriori*. This left 6705 degrees of freedom to be fit, resulting in an RMSD of 0.0081 *e*. After rounding the charge increments to the third digit after the decimal point and further rounding charge increments smaller than 0.0025 *e* that involve a hydrogen to zero, the RMSD increased by 1.2 × 10^{-4} *e* ; consequently, the RMSD calculated using the final charge increments was 0.0082 *e*.[†]

Fitting considerations and results—The number of degrees of freedom in the final fit (6705) is lower than the number of target data points (9204) so that the fitting problem is formally not underdefined. However, there are many redundancies (ie. chemical groups that occur in more than 1 molecule) in the target data. Additionally, bond charge increments in a ring are only defined plus or minus a constant; for example, in a 3-membered ring A-B-C, adding a constant fractional charge simultaneously to the increments β_{ab}, β_{bc} and β_{ca} does not change the resulting charge distribution. To counteract the resulting linear dependence in the fitting problem, all bond charge increments were subject to a restraining bias towards 0; in other words, the merit function χ^2 of the least-squares fit is given by

$$\chi^2 = \sum_{\text{atoms } i} \left(q_i - q_i^{t\arg et}\right)^2 + k \sum_{\text{increments } n} \beta_n^2$$

where q_i is calculated from the charge increments β_n as described above and *k* is a factor determining the strenght of the restraining bias, which we set to 10^{-3}. This bias has a minimal effect on properly determined charge increments, but effectively restrains in-ring increments and other poorly determined degrees of freedom to their lowest possible value. In this sense, the procedure is somewhat analogous to Bayly *et al.*'s RESP model.[23]

The restraining factor also effectively prevents overfitting of the aforementioned poorly determined degrees of freedom, and this preventive effect is further amplified by the incremental fitting and rounding scheme discussed in the subsection "Charge increment optimization" above. Indeed, of the 6705 degrees of freedom, 1536 fitted to absolute values of 0.002 *e* or lower, which usually do not have a significant effect on the charges. Thus, it can be argued that if the poorly determined degrees of freedom that were kept near zero by the restraining bias are not counted, 5169 well-defined degrees of freedom were fit to non-zero values using 9204 data points. This still does not fully eliminate the risk of overfitting in the sense that even a well-defined degree of freedom may reflect non-representative structures and charges in the target data. This risk is hard to quantify using the current data set because it is chemically very diverse[2] and contains too little redundancy for proper cross-validation. Thus, the only way to identify and cure possible cases of "well-defined overfitting" is to increase the size of the target data set, which is an ongoing effort. Indeed, as the CHARMM charge assignment methodology is laborious,[2] to perform a systematic validation study is a research project in its own right, and is the subject of future work.

Special treatment was required for conjugated double bond atom types. As mentioned under the subsection "Additional applications of bond groups" above, there are 2 sets of atom types

[†] All numbers presented in this paragraph are based on release 0.9.0 of the CGenFF program (force field version 2b6). As the charge increment fitting is based on explicitly parameterized compounds for which all parameters are available, the algorithmic improvements discussed in this section do not influence the results. The release of version 2b7 of the force field does, but not to a significant extent.

for conjugated double bonds, which alternate in a chain of conjugated double bonds. This implies that every parameter involving these conjugated double bonds has a counterpart in which the atom types are switched. For the purpose of optimizing charge increments, extra constraints were implemented so that the same increments are applied to both counterparts, in order to prevent charge assignment from arbitrarily depending on the assignment of the double bonded atom types. Similarly, charge assignments associated with parameters that remain identical (except for reversal of the *order* of the atom types) after swapping double bonded atom types were constrained to zero.

Figure 3 shows the agreement between the 9681 manually assigned and optimized partial charges in CGenFF, and their counterparts that were determined using the fitted charge increments. When using only the bond charge increments (green squares), the data exhibits a relatively large spread, especially considering that a 0.1 *e* deviation (solid diagonal lines) in the partial charge of an atom that participates in a hydrogen bond typically results in a 1–3 kcal/mol deviation in the strength of that hydrogen bond. The spread becomes gradually less as the angle and dihedral charge increments are included, in line with the halving of the RMSD for every additional term as shown above. Our final approach, including bonds, angles and dihedrals (blue diamonds), is in excellent agreement with the existing charges. A few minor outliers are visible in Figure 3 in the form of horizontal chains of data points that cross the diagonal. The chain at y=0.115 extending to the right of the diagonal is caused by the fact that the aromatic hydrogen charges on difluorobenzene, difluorotoluene and benzylphosphonate were manually optimized and hence deviate from the standard CHARMM charges to which the charge increment optimization was constrained. A similar chain representing the corresponding carbon atoms can be seen at y=−0.115 extending to the left when only taking the bond charge increments into account (green squares). Unlike the hydrogen atoms, this chain disappears when including “higher-order” increments because no constraint was applied to the charges on non-hydrogen atoms; this suggests that our charge assignment scheme would indeed be able to capture mesomeric charge effects in aromatic rings if given an appropriate training set. Another horizontal line, at y=+0.197, consists of aromatic hydrogen atoms in indole-like rings. Indeed, the CGenFF collection of model compound includes three different charge sets for indole and a number of indole-like heterocycles that were all optimized independently and have different charges as a consequense of the common excess of degrees of freedom in the charge optimization for these kind of compounds. The 0.197 charge (as well as other charges on the indole ring system) is a least-squares compromise between the aforementioned model compounds, as testified by the chain of data points being horizontally centered around the diagonal and by the acceptable reproduction of QM water interaction energies demonstrated in the 2-(dimethylcarbamoyl)indole case study below. Just like for the regular atomatic rings, there is a less clearly resolved cluster of outliers representing the corresponding carbon atoms at the negative side of the plot.

Per-charge penalties—Finally, a method was needed to derive a per-charge penalty σ_i from the per-parameter penalties of the parameters that contributed charge increments to a select atom. The first attempt was to use the maximum of all these contributions. This had the undesirable effect of producing high penalties for atoms that receive a small charge contribution from a high-penalty parameter (see for instance the $M_{+\infty}$ penalty for H2 and C2 in Table 4), making it clear that the magnitude of the contributing charge increments should influence the per-charge penalties. Next, we considered using averages of the per-parameter penalties σ_n, weighted by the absolute values of the corresponing charge increments β_n (equation 2 with p = 1). This approach was intially chosen in favor of sums (see below) because we wanted the magnitude and interpretation of the per-charge penalties to be on the same scale as the per-parameter penalties. However, it was observed that a large number of small penalties with large weights could strongly pull down the average. This problem, which would give rise to unintuitive results such as the two atoms in a highly polarized bond having very different penalties (compare for instance M_1 for C1 and C5 in Table 4), could be somewhat alleviated by using averages that are sensitive to high numbers, such as the quadratic, cubic, quartic and 8th-power means (ie. power means with exponent p = 2, 3, 4 and 8, respectively; see equation 2), but doing so comes at the cost of a decreased sensitivity to the weights, as can be rationalized by considering that in the limit for $p \rightarrow +\infty$, the power mean becomes the maximum.

$$\sigma_i = M_p(\sigma_n) = \sqrt[p]{\frac{\sum_{\text{increments } n} |\beta_n| \sigma_n^p}{\sum_{\text{increments } n} |\beta_n|}} \quad (2)$$

In the end, no acceptable trade-off between sensitivity to high penalties and sensitivity to the weights was found. Moreover, the very concept of averaging is not consistent with the cumulative nature of penalties; for a constant per-parameter penalty, applying a larger number of charge increments, or a charge increment that is larger in absolute value, should yield a larger per-charge penalty. In this sense, the nearly-identical M_2–M_8 penalties for N and H1 in Table 4 are undesirable. Therefore, we settled on using charge increment-weighted sums of the penalties. To retain the desired sensitivity to high penalties, we use the root of the sum of the squares, thus effectively treating the per-parameter penalties as uncorrelated relative errors of the charge increments that are summed to obtain partial charges (equation 3 with q = 1).

$$\sigma_i = \sqrt{\sum_{\text{increments } n} |\beta_n|^q \sigma_n^2} \quad (3)$$

Weighting this sum by the charge increments yielded results that were deemed too sensitive to the weights. Compare for instance C1 and C4 in Table 4; while C1 should have a higher penalty because it involves more high-value-high-penalty increments, the factor ~16 difference in the q=1 column is too high. To decrease this sensitivity, we instead tried weighting by the square root ($q = ½$) and the cube root ($q = \frac{1}{3}$) of the weight. Acceptable sensitivity was obtained with both options, but the cube root had the additional advantage of bringing the final sum in an acceptable range, eliminating the need to apply an arbitrary scaling factor to compensate for the fact that the sum of the absolute values of the charge increments is much lower than 1 in a vast majority of cases. Finally, the absolute values of all charge increments were increased by $\delta = (0.05)^6 = 1.5625 \times 10^{-8}$ before being used as weight factors. As can be seen in the final expression for σ_i, the per-charge penalty on atom i, this is equivalent to assigning a 5% weight factor to the per-parameter penalties σ_n associated with increments β_n=0 (equation 4). Doing so is important in the presence of dihedrals with a high penalty, for which the charge increment is set to 0 and the penalty to the maximum allowable value for dihedrals (currently 50, as mentioned earlier). Although these increments do not contribute to the partial charge, they should have some effect on the penalty to reflect the fact that a contribution to the partial charge was ignored.

$$\sigma_i = \sqrt{\sum_{\text{increments } n} \sqrt[3]{|\beta_n| + \delta}\, \sigma_n^2} \quad (4)$$

Benchmarking—As a simple benchmark for the performance of the methods described in the preceding and present paper, we downloaded a ready-to-dock version (reference pH = 7) of the Maybridge HitFinder™ Collection from the ZINC database[33] (http://zinc.docking.org/catalogs/maybhit/). The resulting mol2 file contained 18662 structures out of which 18429 could be successfully assigned atom types; the remaining 223 compounds contained chemical groups that are not yet supported by the CGenFF force field at the time of writing, most importantly 4-membered rings. Atom typing and assignment of charges and parameter by analogy on these 18429 structures took 619 seconds on a single core of a 3.1 MHz AMD Athlon II X2 255 processor, which can be considered a below-average desktop processor at the time of writing. This is equivalent to a rate of 29.8 molecules per second, demonstrating the suitability of the current method for high-throughput applications.

Case studies

Case study 1: 2-(dimethylcarbamoyl)indole (2)

As a case study, we first consider 2-(dimethylcarbamoyl)indole (compound **2** in Figure 1). The presented algorithm assigns the following bonded parameters by analogy (see Tables S2 and S3 in the Supporting Information for the original CHARMM toppar stream files generated without and with the bond group feature, respectively):

- 1 bond (penalty 40 both with and without the bond group feature discussed above under the heading “Assignment of parameters by analogy”[‡])
- 4 angles (maximum penalty 26.5 without bond groups or 71 with bond groups)
- 10 dihedrals (maximum penalty 88 without bond groups or 128 with bond groups)
- 1 improper dihedral (penalty 5)

All these parameters contain the C2–C10 bond, which is defined by a combination of atom types that was not part of the CGenFF parametrization. The highest penalty of the charges is 49.4 without bond groups or 75.4 with bond groups. The parameters and charges were validated using the original CGenFF parametrization procedure, the details of which are described elsewhere.[2] For the bonds and angles, the MP2/6-31G* conformation was compared with the CGenFF conformation, both minimized in vacuum. Measurements of relevant bond lengths and angles are compared in Table 2. All values are within acceptable limits (0.03Å for the bonds and 3° for the angles)[2] from their QM values, except the C2-C10 bond, which has a correspondingly high penalty, and the N1-C2-C10 and C3-C2-C10 angles, which exhibits serious deviations along with the C2-C3-N1-C10 improper dihedral. Specifically, without the bond group feature, the parameter for this angle was transferred from an in-ring parameter for 5-membered sp^2 rings, with a reference value near the ideal 108°, while the out-of-ring angles for these kind of rings should be near an ideal 126°; see “Assignment of parameters by analogy” above and Figure 2. As the sum of the 3 angle around C2 is only 336.9°, a catastrophic out-of-plane deviation of 26.2° arises in the C1-C3-N1-C10 improper dihedral. Although the penalty score for these angles is 26.5 and basic validation is recommended for penalties higher than 10, it would be more ideal if a penalty score higher than 50 would have been assigned, which would mandate extensive validation/optimization. Furthermore, similar cases were identified in a significant percentage of user-supplied molecules, some of which featured penalties lower than 10. The frequency and magnitude of this problem, taken together with the fact that it might potentially go undetected by a quick glance at the penalties, prompted us to implement the bond group feature. The inclusion of this extra contribution to the penalty function caused different angle parameters to be chosen by the assignment algorithm, which eliminated the out-of-plane deviation. It should be noted that although the sum of the angles is now close to 360°, there is still a large in-plane deviation. This is due to an unexpectedly large physical deviation from the idealized geometry; indeed, the MP2 values for N1-C2-C10 and C3-C2-C10 are far from 126°, such that the MP2 minimized geometry (Figure 4) is visibly different from the idealized geometry for a 5 membered ring (Figure 2). We speculate that this is caused by a

[‡] The assignment without and with the “bond group” feature was respectively performed with version 0.9.1 and 0.9.6 of the CGenFF program.

combination of favorable electrostatic interactions between the H1 and O10 atoms and steric repulsion between the dimethylamide group and the ring system, which cannot be relieved by tilting the amide group out of the plane of the ring because the system is kept planar by π-orbital overlap effects in addition to the electrostatic interaction mentioned above. As the force field's LJ parameters were optimized targeting bulk-phase properties, it appears that the steric repulsion seen in the vacuum QM calculation is not adequately reproduced, yielding a significantly different angular geometry. However, the new penalties for these angles are 64 and 71, respectively, providing the user a clear and unambiguous indication that their values should be verified and/or optimized. Note that although the bond group feature increased the penalties, the parameters by analogy became better, as the inappropriate parameters obtained without the bond group feature were given inappropriately small TPS values.

Manual parameter optimization based on geometric measurements of the QM and CGenFF minimized conformations readily allowed the N1-C2-C10 and C3-C2-C10 angles to be brought within 2° of the QM values (last columns in Table 2). Further properties reported in this paragraph include this correction so that the angular deviation does not interfere with the evaluation of the dihedrals. As for the interaction energies with water, which were calculated using the MP2 geometry (Table 5), C7, N1 and O10 are significantly outside the 0.2 kcal/mol difference criterion for CGenFF, and the interaction distance for C7 is longer than expected, in line with its interaction energy being too weak. This can be explained by the fact that the current training set for the charge increments is based on the assumption that substituted aromatic systems have the same charges as unsubstituted ones, thus negating the algorithm's ability to represent mesomeric effects. However, while allowing mesomeric substituent effects to influence the charges in the aromatic rings in the training set may improve N1, it will not necessarily improve C7 because the current implementation can only capture these electronic effects up to a distance of 3 bonds. Fortunately, the hydrogen bond donated by C7 is one of the weaker interactions in the system, and most relevant intermolecular properties are dominated by the stronger hydrogen bonds. In this sense, it is comforting that the hydrogen bond accepted by O10 is reproduced significantly better. Indeed, its interaction energy is only slightly outside of the CGenFF target range, which is an encouraging result given that C10 is involved in the formation of a novel bond, as indicated by its high penalty. Finally, although N1's penalty is in line with its deviation from the QM target data, the example of C7 illustrates that even a penalty of 0 does not guarantee perfect agreement. As discussed above, the penalty is a measure of the degree of analogy between chemical groups in the training set and the user's molecule, and neither factors in the quality of the training set, nor – more importantly – the fundamental appropriateness of applying charges and parameters by analogy for the given molecule. The directions of the HF and MP2 dipole moments differ by a relatively large 16°; the CGenFF dipole moment is in better agreement with the HF dipole moment (Table 6), which is a commonly observed consequence of the fact that the charges are primarily fitted to reproduce HF interaction energies. However, the agreement with both the HF and MP2 dipole moment is acceptable given the fact that the magnitude of the CGenFF dipole moment should be overestimated by 20 to 50%.[2] The potential energy scan (PES) around the C2-C10 rotatable bond is shown in Figure 5. The local minimum at 30° is 0.87 kcal/mol too low and shifted by 15° in CGenFF, while the barriers at 0° and 90° are 0.84 kcal/mol too high and 1.61 kcal/mol too low, respectively. While the agreement is not as good as in the CGenFF results following parameter optimization, these PES are expected to yield qualitatively similar results in MD simulations, which is very good considering that the penalties of the 4 dihedrals terms around this bond are 87, 88, 127 and 128, and our guidelines mandate extensive validation and/or re-optimization for any parameters with penalties higher than 50.

Case study 2: N-benzylacetamide (3)

A second case study is N-benzylacetamide (compound **3** in Figure 1, Table S4 of the supporting information). In this case, one angle with penalty 28.5 was assigned by analogy, as well as three dihedrals with penalties 47, 28.5 and 28.5. The highest penalty of the charges is 39.4; a comparison between CGenFF charges and standard CHARMM charges is given in Table 7. This table shows a fair correlation between the deviation from the standard CHARMM charges and the penalty; the three highest penalties correspond to the three charges that have significant (though less than 0.05 *e*) deviations. Given the lack of high-penalty bond and angle parameters, it is not surprise that all minimized bond lengths and angles are in excellent agreement (see Table S5 of the supporting information). Figure 6 depicts the relaxed potential energy surfaces around the C5-C6-C7-N8 and C6-C7-N8-C9 dihedrals. The parameters by analogy for C6-C7-N8-C9 produce good agreement with the QM from 0° to ~120°, but the PES diverge by 2 kcal/mol from 120° to 180°. Based on the geometry at 180° (Figure 7), we speculate that this is because the intramolecular electrostatic interaction of the amide NH group with the benzene ring is substantially different from its interaction with an ester or protonated carboxylic acid in a C-terminal amino acid, which is the origin of the parameter by analogy for this dihedral. Manual fitting of this dihedral leads to the introduction of a dihedral term with n=1, K_φ=1.7 and δ=180 and a second dihedral term with n=3, K_φ=1.7 and δ=0. The former term's relatively high K_φ might indicate that the intramolecular nonbonded interactions are not perfectly balanced for the purpose of reproducing this vacuum MP2 PES, and we speculate that the fitted parameters might be poorly transferable to other molecules. This is in line with the observation that the transferability of dihedral parameters in general is often problematic; see section "charge assignment" above. Although the penalty of 28.5 associated with this parameter mandates validation, it is possible in some cases to obtain a similarly mediocre agreement for parameters with even lower penalties; therefore, if accuracy is of the utmost importance in the molecule(s) under study, validation is recommended regardless of the penalties. Concerning the C5-C6-C7-N8 scan, the data points with poor agreement between CGenFF and MP2 are caused by the C6-C7-N8-C9 dihedral (which is relaxed at every scan point scan) being close to 180°. Indeed, the value for the the C5-C6-C7-N8 parameter did not change significantly during manual optimization, and the improvement in agreement is purely due to the optimization of the C6-C7-N8-C9 dihedral. This demonstrates that a higher penalty is not a guarantee for poor agreement; rather, the penalty scores should be thought of as measure for the approximation error on the parameter.

Summary

The present paper describes the algorithms for automatic assignment of bonded parameters and charges in CGenFF, thereby allowing for automatic application of CGenFF to arbitrary molecules. Bonded parameters are assigned by substituting atom types in the definition of the desired parameter. The assignment of bonded parameters also forms the basis for the charge assignment. A penalty is associated with every substitution and the existing parameter with the lowest sum of penalties is chosen as an approximation for the desired parameter; the aforementioned sum is returned to the user as a measure for the accuracy of the approximation. Charges are assigned using a variation on the classical bond-charge increment scheme, which essentially associates a bond charge increment with each bond parameter. We extended this scheme by associating two charge increments with each angle parameter and three charge increments with each dihedral parameter. To apply this bond-charge increment scheme to a novel molecule, the aforementioned algorithm for assigning parameters by analogy is used, albeit using slightly different criteria for analogy. This allows for per-charge penalty scores to be returned, which give the user an idea of the quality of the charges. Charge increment values are optimized to reproduce the charges on model compounds that were part of the parametrization of the force field as well as possible using a restrained least-squares fitting algorithm. A final charge RMSD of 0.0082 *e* was obtained. The validity of the final parameters and charges and the relevance of the associated penalty scores were illustrated by 2 case studies; a more thorough validation study will be the subject of a future paper. As the methods discussed in the preceding and present paper are entirely empirical, they are excellently suited for high-throughput applications, as testified by the fact that we processed a representative database of 18429 drug-like structures at a rate of 29.8 compounds per second on a single core of a below-average desktop processor. Finally, it should be reiterated that the penalties are indicative of the analogy of the generated parameters or charges with available terms in CGenFF, rather than the intrinsic fitness of these terms for application in the molecule of interest. Consequently, large penalties may yield satisfactory accuracy, while significant deviations from acceptable accuracy may occur when small or zero penalties are reported. Indeed, given the sensitivity of chemical properties to environment, in cases where parameters are available in CGenFF for a new molecule, there may still be significant limitations in the accuracy of the resulting model. It is important that the user is aware of such limitations and perform the appropriate validation and optimization as required.

Use of the CGenFF program for automatic atom typing and assignment of parameters and charges by analogy is provided at https://www.paramchem.org/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support from the NIH (GM51501, GM070855, CA107331), the NSF (CHE-0823198), the Waxman Foundation and the University of Maryland Computer-Aided Drug Design Center are acknowledged. Additionally, this research was supported in part by the NSF through TeraGrid[34] resources provided by NCSA and PSC.

References

1. Vanommeslaeghe K, MacKerell AD Jr. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. 2012 *in preparation*.
2. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, MacKerell AD Jr. CHARMM General Force Field (CGenFF): A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. J. Comput. Chem. 2010; 31(4):671–690. [PubMed: 19575467]
3. MacKerell, AD, Jr. Protein Force Fields. In: Schleyer, PvR; Allinger, NL.; Clark, T.; Gasteiger, J.; Kollman, PA.; H.F. Schaefer, I.; Schreiner, PR., editors. Encyclopedia of Computational Chemistry. Vol. Vol. 3. Chichester: John Wiley & Sons; 1998. p. 2191-2200.
4. Brooks BR, Brooks CL III, MacKerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodošček M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The Biomolecular Simulation Program. J. Comput. Chem. 2009; 30:1545–1614. [PubMed: 19444816]
5. Allen, MP.; Tildesley, DJ. Computer Simulation of Liquids. Oxford: Clarendon Press; 1987. p. 385
6. Clark M, Cramer R, van Opdenbosh N. Validation of the General Purpose Tripos 5.2 Force Field. J. Comput. Chem. 1989; 10:982–1012.
7. Dauber-Osguthorpe P, Roberts VA, Osguthorpe DJ, Wolff J, Genest M, Hagler AT. Structure and Energetics of Ligand-Binding to Proteins - Escherichia Coli Dihydrofolate Reductase Trimethoprim, a Drug-Receptor System. Proteins: Structure, Function and Genetics. 1988; 4(1):31–47.
8. Rappé AK, Casewit CJ, Colwell KS, Goddard WA III, Skiff WM. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. Journal of the American Chemical Society. 1992; 114:10024–10035.

9. Hwang MJ, Stockfisch TP, Hagler AT. Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. Journal of the American Chemical Society. 1994; 116:2515–2525.

10. Allinger NL. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. Journal of the American Chemical Society. 1977; 99(25):8127–8134.

11. Allinger NL, Yuh YH, Lii J-H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. Journal of the American Chemical Society. 1989; 111:8551–8566.

12. Lii J-H, Allinger NL. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 2. Vibrational Frequencies and Thermodynamics. Journal of the American Chemical Society. 1989; 111:8566–8575.

13. Lii J-H, Allinger NL. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 3. The van der Waals' Potentials and Crystal Data for Aliphatic and Aromatic Hydrocarbons. Journal of the American Chemical Society. 1989; 111:8576–8582.

14. Allinger NL, Chen KH, Lii JH, Durkin KA. Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. J. Comput. Chem. 2003; 24(12):1447–1472. [PubMed: 12868110]

15. Momany FA, Rone R. Validation of the General Purpose QUANTA 3.2/CHARMm Force Field. J. Comput. Chem. 1992; 13(7):888–900.

16. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general AMBER force field. J. Comput. Chem. 2004; 25:1157–1174. [PubMed: 15116359]

17. Wang JM, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. Journal of Molecular Graphics and Modelling. 2006; 25(2):247–260. [PubMed: 16458552]

18. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods R. The Amber biomolecular simulation programs. J. Comput. Chem. 2005; 26:1668–1688. [PubMed: 16200636]

19. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemical Society. 1995; 117(19):5179–5197.

20. Halgren TA. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. J. Comput. Chem. 1996; 17(5–6):490–519.

21. Halgren TA. Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. J. Comput. Chem. 1996; 17(5–6):520–552.

22. Halgren TA. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. J. Comput. Chem. 1996; 17(5–6):616–641.

23. Bayly CI, Cieplak P, Cornell WD, Kollman PA. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. J. Phys. Chem. 1993; 97:10269–10280.

24. Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. J. Comput. Chem. 2000; 21:132–146.

25. Jorgensen WL, Tirado-Rives J. Molecular Modeling of Organic and Biomolecular Systems Using BOSS and MCPRO. J. Comput. Chem. 2005; 26(16):1689–1700. [PubMed: 16200637]

26. Storer JW, Giesen DJ, Cramer CJ, Truhlar DG. Class IV charge models: a new semiempirical approach in quantum chemistry. J Comput Aided Mol Des. 1995; 9:87–110. [PubMed: 7751872]

27. Thompson JD, Cramer CJ, Truhlar DG. Parameterization of charge model 3 for AM1, PM3, BLYP, and B3LYP. J Comput Chem. 2003; 24:1291–1304. [PubMed: 12827670]

28. Cornell WD, Cieplak P, Bayly CE, Kollman PA. Application of RESP Charges to Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. Journal of the American Chemical Society. 1993; 115:9620–9631.

29. Jakalian A, Jack DB, Bayly CI. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. J. Comput. Chem. 2002; 23(16):1623–1641. [PubMed: 12395429]

30. Udier-Blagović M, Morales de Tirado P, Pearlman SA, Jorgensen WL. Accuracy of Free Energies of Hydration Using CM1 and CM3 Atomic Charges. J. Comput. Chem. 2004; 25(11):1322–1332. [PubMed: 15185325]

31. Halgren TA. COMP 100 - Grand challenge force fields and beyond. Abstracts of Papers of the American Chemical Society. 2008; 236 100-COMP.

32. Shim J, MacKerell AD Jr. Computational ligand-based rational design: Role of conformational sampling and force fields in model development. MedChemComm. 2011; 2(5):356–370. [PubMed: 21716805]

33. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. Journal of Chemical Information and Modeling. 2012; 52(7):1757–1768.

34. Catlett, C.; Allcock, WE.; Andrews, P.; Aydt, R.; Bair, R.; Balac, N.; Banister, B.; Barker, T.; Bartelt, M.; Beckman, P.; Berman, F.; Bertoline, G.; Blatecky, A.; Boisseau, J.; Bottum, J.; Brunett, J.; Bunn, J.; Butler, M.; Carver, D.; Cobb, J.; Cockerill, T.; Couvares, PF.; Dahan, M.; Diehl, D.; Dunning, T.; Foster, I.; Gaither, K.; Gannon, D.; Goasguen, S.; Grobe, M.; Hart, D.; Heinzel, D.; Hempel, C.; Huntoon, W.; Insley, J.; Jordan, C.; Judson, I.; Kamrath, A.; Karonis, N.; Kesselman, C.; Kovatch, P.; Lane, L.; Lathrop, S.; Levine, M.; Lifka, D.; Liming, L.; Livny, M.; Loft, R.; Marcusiu, D.; Marsteller, J.; Martin, S.; McCaulay, S.; McGee, J.; McGinnis, L.; McRobbie, M.; Messina, P.; Moore, R.; Moore, R.; Navarro, JP.; Nichols, J.; Papka, ME.; Pennington, R.; Pike, G.; Pool, J.; Reddy, R.; Reed, D.; Rimovsky, T.; Roberts, E.; Roskies, R.; Sanielevici, S.; Scott, JR.; Shankar, A.; Sheddon, M.; Showerman, M.; Simmel, D.; Singer, A.; Skow, D.; Smallen, S.; Smith, W.; Song, C.; Stevens, R.; Stewart, C.; Stock, RB.; Stone, N.; Towns, J.; Urban, T.; Vildibill, M.; Walker, E.; Welch, V.; Wilkins-Diehr, N.; Williams, R.; Winkler, L.; Zhao, L.; Zimmerman, A. TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications. In: Grandinetti, L., editor. High Performance Computing (HPC) and Grids in Action. Vol. Vol. 16. Amsterdam: IOS Press; 2007.

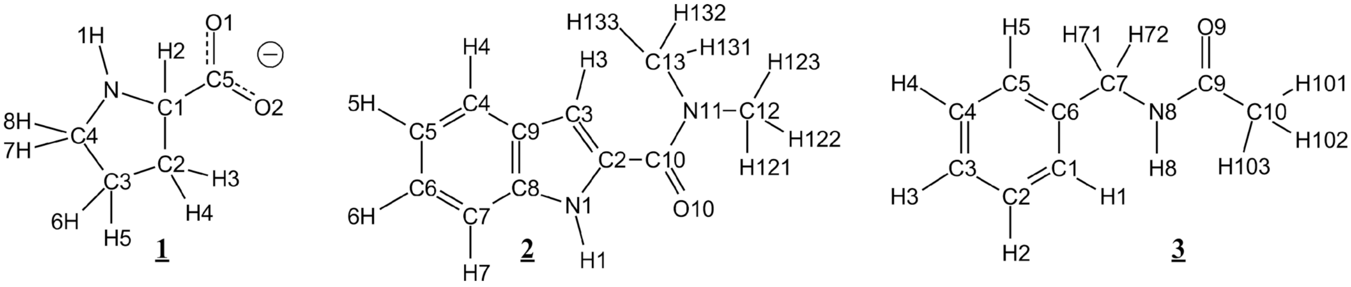

Figure 1.

Chemical structures of compounds discussed in this article, including atom naming. **1**: negatively charged proline, used as an example for illustrating the different ways of calculating per-charge penalties (see Table 4); **2**: 2-(dimethylcarbamoyl)indole, case study 1; **3**: N-benzylacetamide, case study 2.

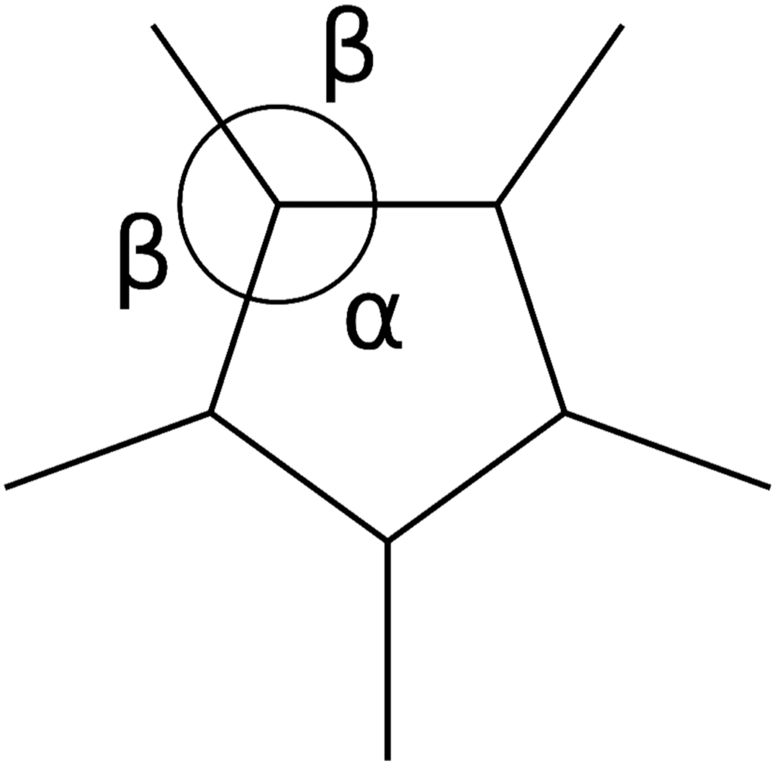

Figure 2.
Generic planar 5-membered ring.

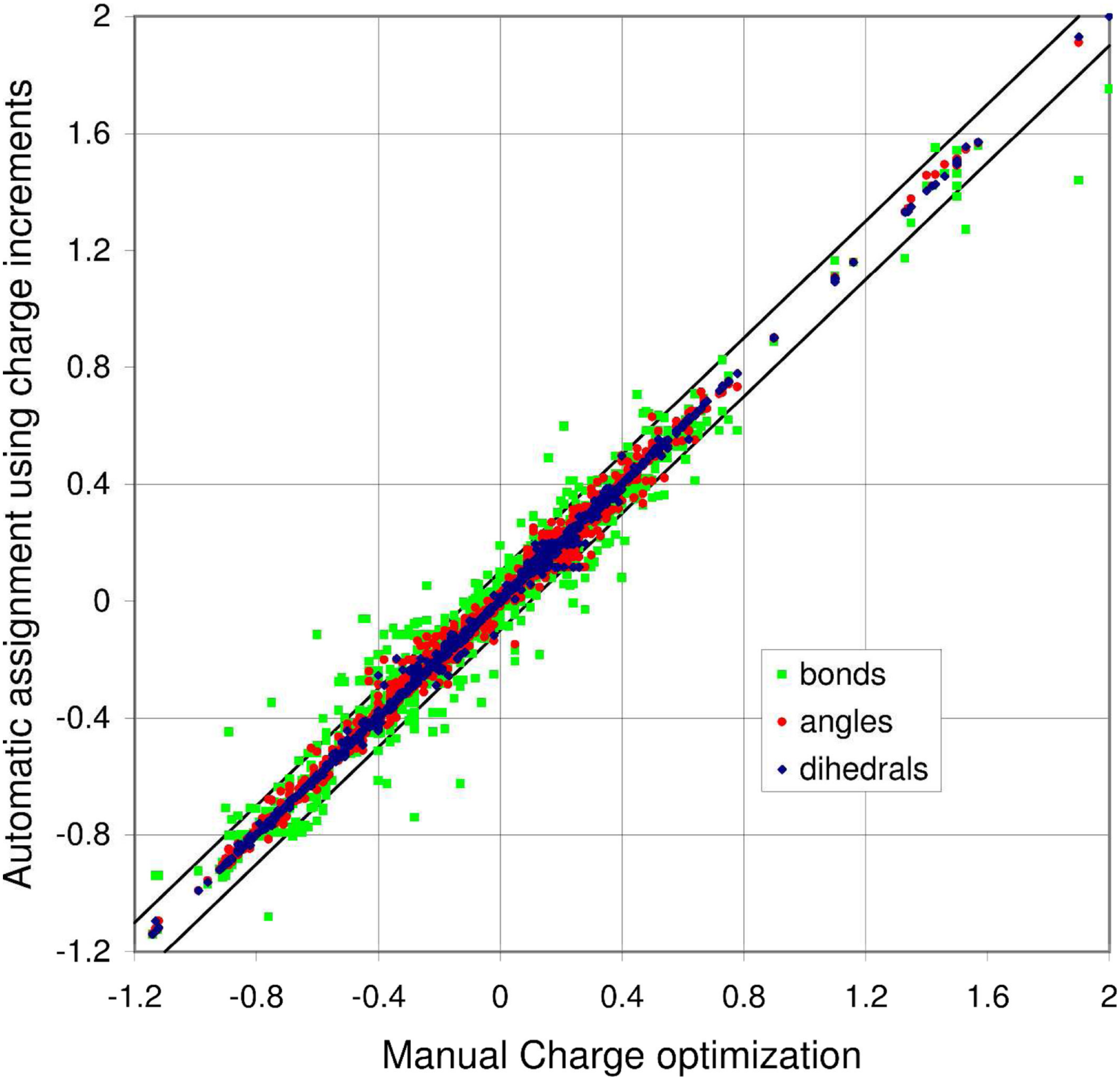

Figure 3.

Results of charge increment fitting by highest term. The green squares represent the bond charge increments only, the red dots bonds and angles, and the blue diamonds bonds, angles and dihedrals. The solid diagonal lines are deviations of ±0.1 *e* from identity.

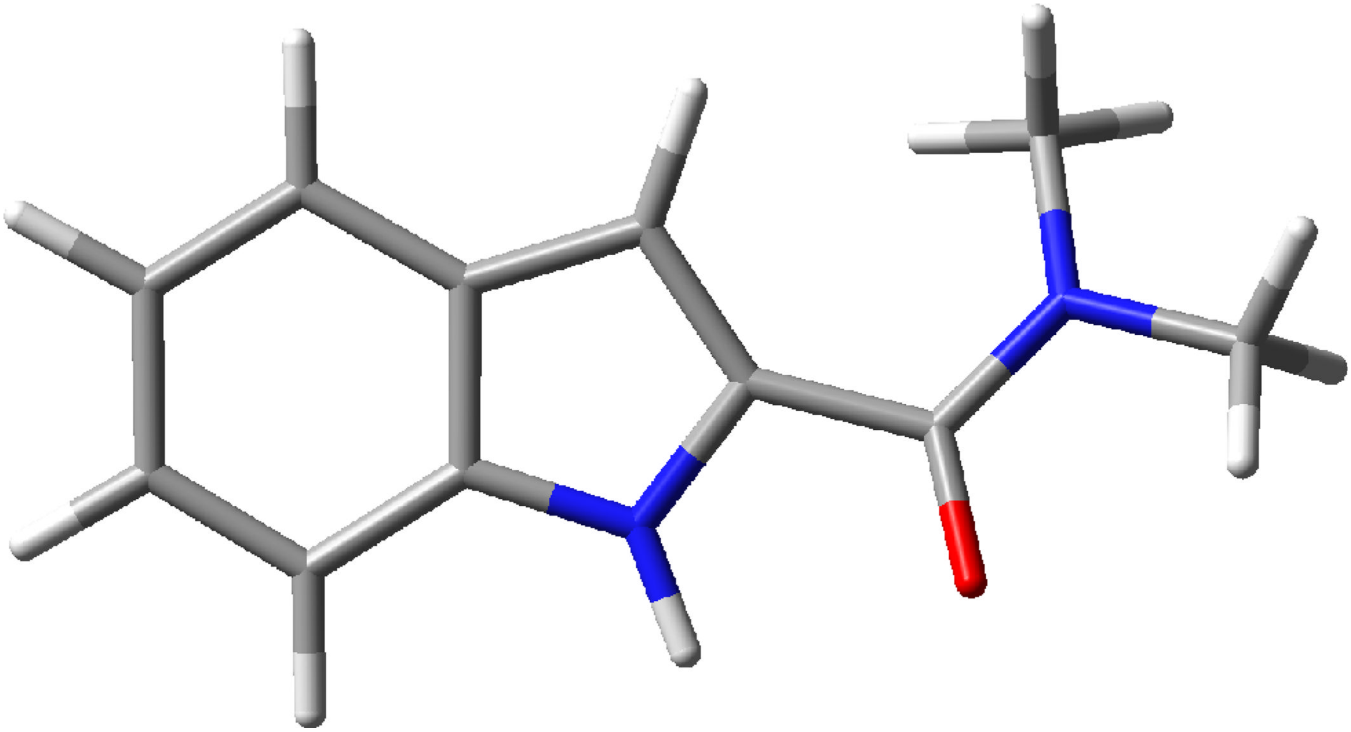

Figure 4.
2-(dimethylcarbamoyl)indole (**2**) MP2 minimized conformation.

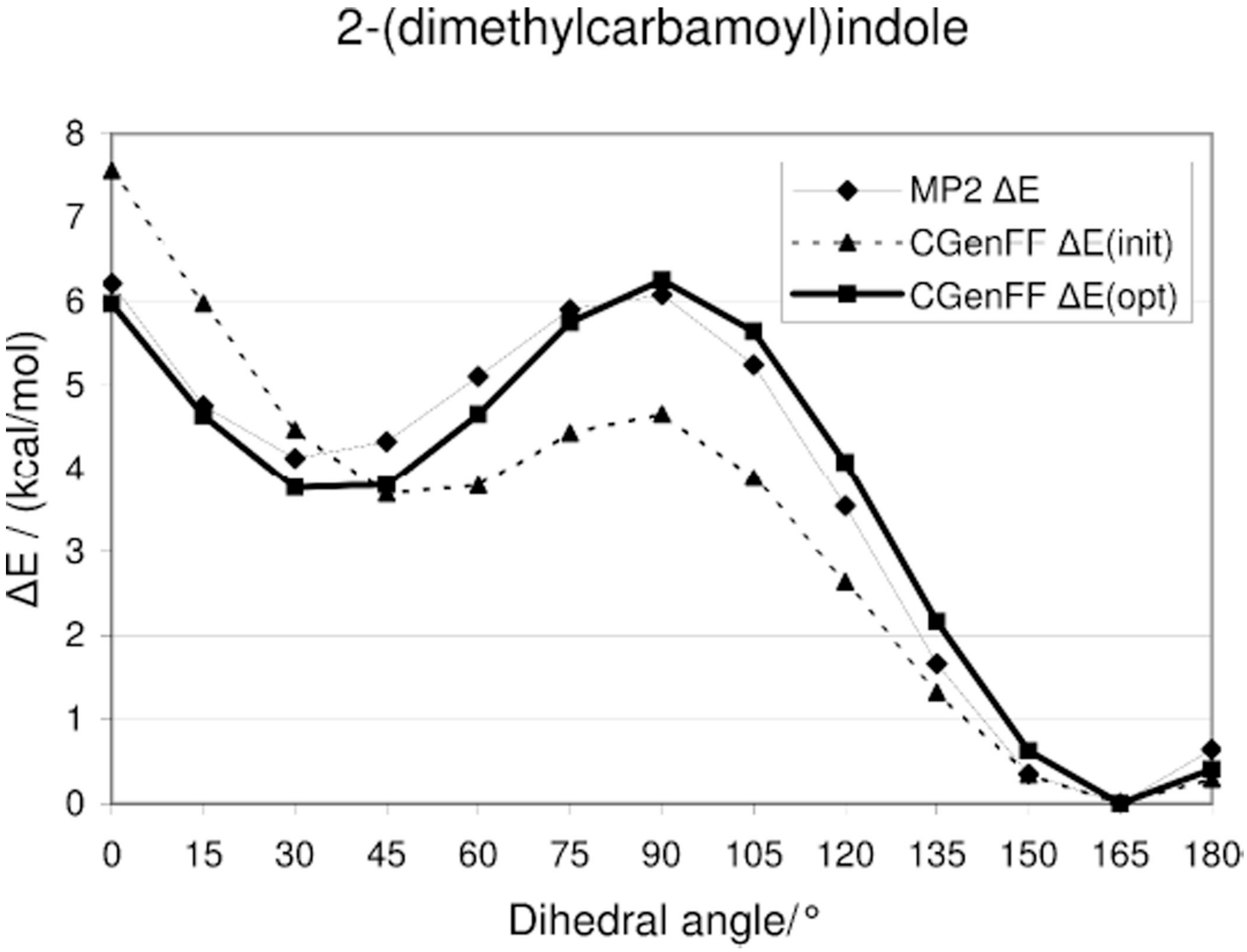

Figure 5.
Potential energy scan around the C2-C10 bond in 2-(dimethylcarbamoyl)indole (**2**).

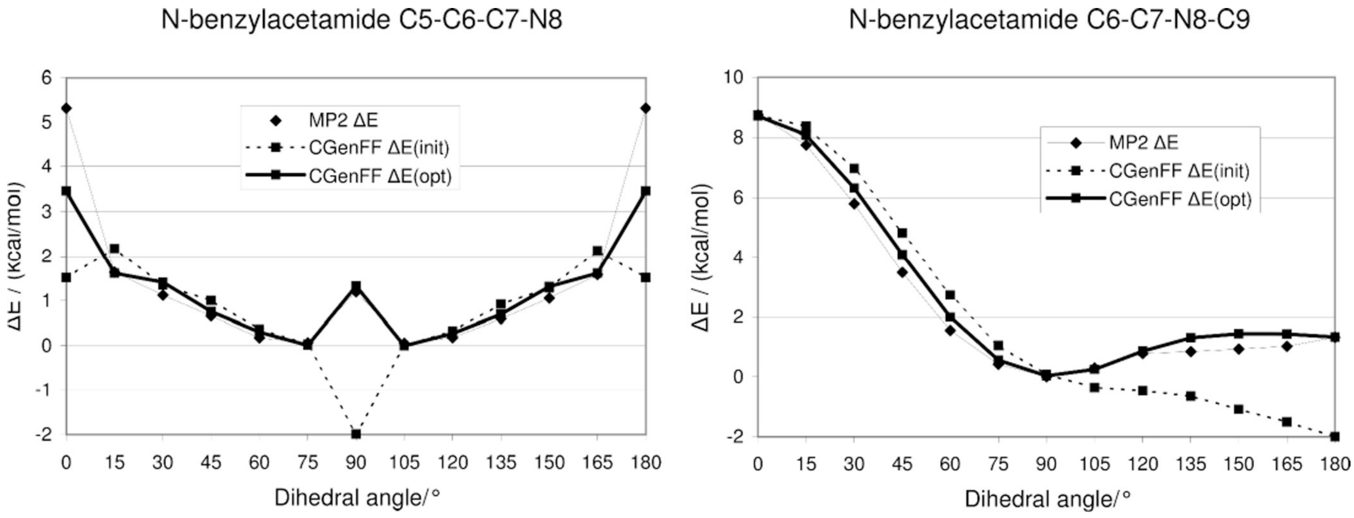

Figure 6.

Potential energy scans around the C6-C7 and C7-N8 bonds in N-benzylacetamide (**3**).

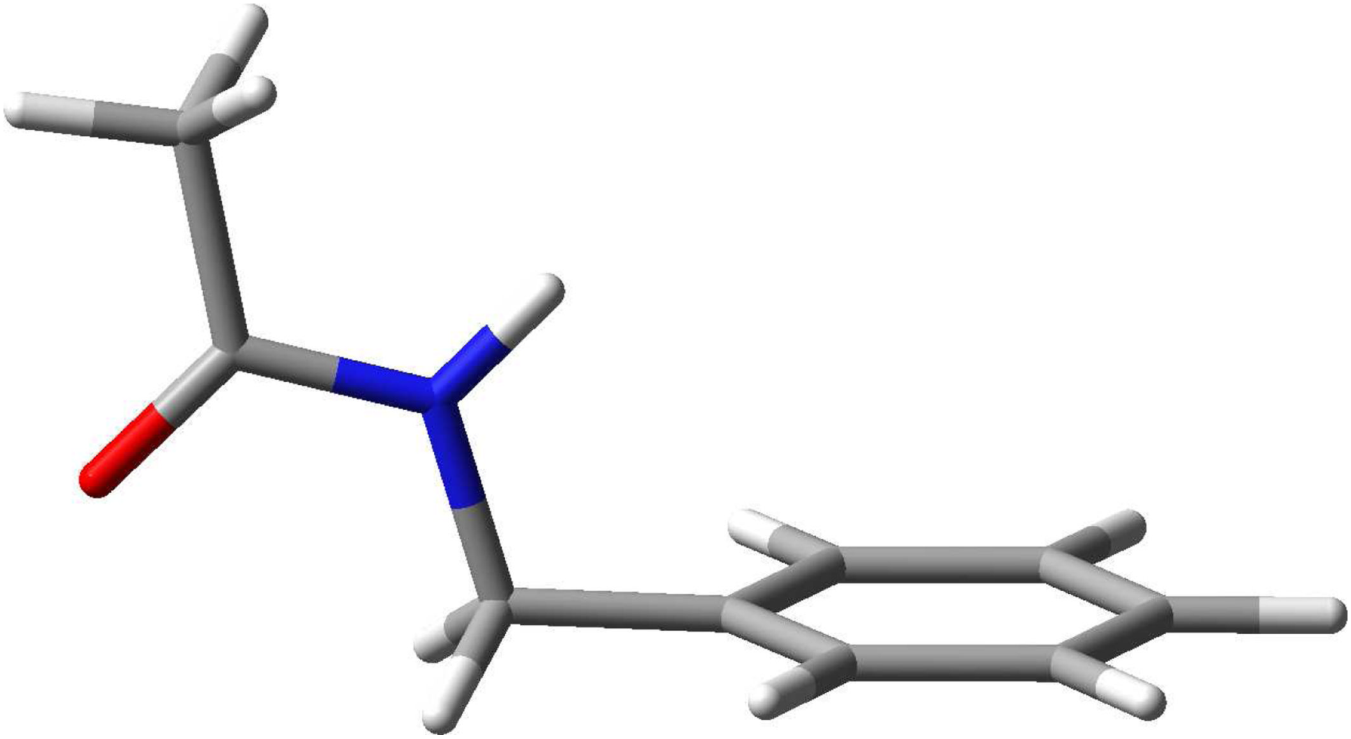

Figure 7.
N-benzylacetamide (**3**) conformation for ϕ(C5-C6-C7-N8)=90° and ϕ(C6-C7-N8-C9)=180°.

Table 1

Extract of the parameter assignment rule file that handles *sp*^{3} nitrogen atom types.

```
cat NG3
sub NG3P   : pri  0 alt NG3N 2 up 12
sub NG3N   : pri  5 alt NG3P 2 up 12
end

cat NG3P
typ NG3P2  : pri 0 alt NG3P1 1 alt NG3P3 3 alt NG3P0 4 up 8
typ NG3P3  : pri 1 alt NG3P2 1 alt NG3P1 2 alt NG3P0 4 up 8
typ NG3P1  : pri 3 alt NG3P2 1 alt NG3P0 3 alt NG3P3 4 up 8
typ NG3P0  : pri 4 alt NG3P1 1 alt NG3P2 2 alt NG3P3 4 up 8
end

cat NG3N
typ NG321  : pri 0   alt NG311 1   alt NG301 1.5 alt NG3N1 2.5 alt NG3C51 3   alt NG331 4   up 8
typ NG311  : pri 0.5 alt NG301 0.5 alt NG321 1   alt NG3N1 1.5 alt NG3C51 2   alt NG311 5   up 8
typ NG301  : pri 1   alt NG311 0.5 alt NG321 1.5 alt NG3N1 2   alt NG3C51 2.5 alt NG331 5.5 up 8
typ NG3N1  : pri 1.5 alt NG311 1.5 alt NG301 2   alt NG321 2.5 alt NG3C51 3.5 alt NG331 6.5 up 8
typ NG3C51 : pri 2.5 alt NG311 2   alt NG301 2.5 alt NG321 3   alt NG3C51 4   alt NG331 7   up 8
typ NG331  : pri 4   alt NG321 4   alt NG311 5   alt NG301 5.5 alt NG3N1  6.5 alt NG3C51 7   up 8
end
```

Table 2

CGenFF minimized geometry of 2-(dimethylcarbamoyl)indole (**2**) arising from parameters by analogy without and with bond groups, compared to MP2. r: bond lenght; θ: angle; φ: improper dihedral.

		Analogy without bond groups			**Analogy with bond groups**			**Manual optimization**	
	MP2	**CGenFF**	**Difference**	**Penalty**	**CGenFF**	**Difference**	**Penalty**	**CGenFF**	**Difference**
r(C2–C10)	1.483	1.516	0.034	40	1.514	0.032	40	1.517	0.035
r(C10-N11)	1.365	1.379	0.014	0	1.379	0.014	0	1.378	0.013
θ(N1-C2-C10)	114.5	112.1	−2.3	26.5	128.2	13.7	64	114.1	−0.4
θ(C3-C2-C10)	136.5	116.8	−19.7	26.5	125.9	−10.6	71	137.9	1.4
θ(N1-C2-C3)	108.6	108.0	−0.6	0	105.8	−2.8	0	107.1	−1.6
θ(C2-C10-N11)	119.2	119.3	0.1	9.5	121.3	2.1	9.5	121.4	2.1
θ(C2-C10-O10)	118.2	119.4	1.3	8.5	117.0	−1.2	8.5	117.1	−1.1
θ(O10-C10-N11)	122.6	121.3	−1.3	0	121.7	−0.9	0	121.6	−1.0
φ(C2-C3-N1-C10)	3.6	29.8	26.2	0	2.4	−1.2	0	5.5	1.9

Table 3

Bond groups in version 0.9.6 of the parameter assignment rule file

```
# bond groups for improving conjugated double bonds - considered as equivalent!
bgrp 40 CG2DC1 CG2D1O CG25C1 CG251O CG2DC3 CG2D1 CG2D2 NG2D1 NG2P1
bgrp 40 CG2DC2 CG2D2O CG25C2 CG252O CG2DC3 CG2D1 CG2D2 NG2D1 NG2P1

# bond group for planar 7-membered rings
bgrp 40 CG2R71 CG2RC7
# bond group for planar 6-membered rings
bgrp 20 CG2R61 CG2R62 CG2R63 CG2R64 CG2R66 NG2R60 NG2R61 NG2R62 CG2RC0 NG2RC0 CG2R67
# bond group for planar 5-membered rings
bgrp 20 CG2R51 CG2R52 CG2R53 NG2R50 NG2R51 NG2R52 NG2R53 OG2R50 SG2R50
# bond group for all 5-membered rings
bgrp 20 CG3C50 CG3C51 CG3C52 CG3C53 CG3C54 NG3C51 OG3C51 CG3RC1 CG2R51 CG2R52 CG2R53 NG2R50
NG2R51 NG2R52 NG2R53 OG2R50 SG2R50 CG25C1 CG25C2 CG251O CG252O CG2RC0 NG2RC0 CG2RC7
# bond group for 4-membered rings
bgrp 60 CG3C41 CG3RC1
# bond group for 3-membered rings
bgrp 80 CG3C31 CG3RC1

# bond group for biphenyls
bgrp 47 CG2R67

# bond group for distinguishing triple bonds
bgrp 27 CG1T1 CG1N1 NG1T1
```

Table 4

Different ways of calculation per-charge penalties for negatively charged proline (**1**). The atom names are explained in Figure 1. The power means $M_{+\infty}$, M_1 and M_2 (as defined in equation 2) are respectively equivalent to the maximum (max), arithmetic mean (A), and Root Mean Square (RMS).

		Charge	**power mean (equation 2)**						**equation 3**			**equation 4**
Atom name	**Atom type**		**$M_{+\infty}$ (max)**	**M_1 (A)**	**M_2 (RMS)**	**M_3**	**M_4**	**M_8**	**q=1**	**q=½**	**q=⅓**	
N	NG3C51	−0.847	50	1.46	4.39	8.14	11.60	20.00	4.26	15.31	24.12	24.38
H1	HGP1	0.366	34.5	0.57	4.42	8.76	12.35	20.64	2.67	9.60	14.71	14.71
C1	CG3C51	−0.011	50	14.52	23.31	28.72	32.22	39.22	17.27	37.12	50.53	50.78
H2	HGA1	0.090	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.57
C2	CG3C52	−0.187	50	0.25	1.12	1.89	2.52	4.19	0.52	2.72	4.73	5.93
H3	HGA2	0.090	7.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43
H4	HGA2	0.090	7.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43
C3	CG3C52	0.099	34.5	0.09	1.57	4.36	7.31	15.88	1.11	6.22	11.06	11.06
H5	HGA2	0.090	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
H6	HGA2	0.090	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
C4	CG3C52	−0.209	7.5	0.03	0.48	1.20	1.89	3.77	0.24	1.34	2.38	2.38
H7	HGA2	0.090	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H8	HGA2	0.090	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C5	CG2O3	0.679	50	4.14	13.17	19.74	24.35	34.13	17.21	37.90	51.71	51.95
O1	OG2D2	−0.760	50	0.11	1.87	5.49	9.52	21.82	1.63	9.00	15.95	16.34
O2	OG2D2	−0.760	50	0.11	1.87	5.49	9.52	21.82	1.63	9.00	15.95	16.34

Table 5

Interaction energies (kcal/mol) and distances (Å) of complexes of 2-(dimethylcarbamoyl)indole (**2**) with water in different orientations.

interaction geometry	**ΔE(HF)***	**ΔE (CGenFF)**	**ΔΔE**	**r(HF)**	**r(CGenFF)**	**Δr**	**Penalty**	**Penalty (parent)**
C4H…OHH	−1.75	−1.71	0.04	2.60	2.64	0.04	0	0
C5H…OHH	−1.39	−1.43	−0.04	2.65	2.66	0.01	0	0
C6H…OHH	−1.61	−1.49	0.12	2.60	2.65	0.05	0	0
C7H…OHH	−2.25	−1.63	0.62	2.51	2.64	0.13	0	0
N1H…OHH	−2.07	−2.50	−0.43	2.17	1.98	−0.19	18.23	41.87
C3H….OHH	−2.73	−2.70	0.03	2.83	2.71	−0.12	24.57	41.32
O10…HOH	−6.35	−6.07	0.28	2.01	1.79	−0.22	6.07	67.99
AD			0.09			−0.04		
RMSD			0.31			0.13		
AAD			0.22			0.11		

CGenFF data were calculated using the MP2 geometry for 2. HF/6-31G(d) interaction energies are scaled by a factor 1.16. HF interaction distances are not scaled; however, empirical hydrogen bond minimum distances (ie. involving model compound hydrogen bond donors and acceptors) should be approximately 0.2 Å shorter than HF/6-31G(d) values as required to yield adequate bulk phase properties. The meaning of the symbols describing the interaction geometries as well as the scaling factor and offset are explained in reference[2]. Results include Average Deviation (AD), Root Mean Square Deviation (RMSD) and Absolute Average Deviation (AAD). The values in the column "Penalty" are the penalties of the atoms that directly interact with the water probe, while the "Penalty (parent)" values are associated with the atom to which the latter is bound.

Table 6

Components of the dipole moment of 2-(dimethylcarbamoyl)indole (**2**) calculated with CGenFF and compared to HF and MP2. Units are Debye except for the percentual and angular differences. All dipole moments are calculated on the MP2 optimized geometry.

	HF	MP2	CGenFF	% Difference (HF)	% Difference (MP2)
X	1.51	1.80	2.20	45%	22%
Y	−2.02	−1.27	−2.46	22%	93%
Z	1.13	1.09	1.09	−3%	0%
total	2.76	2.46	3.48	26%	41%
		HF vs. MP2		CGenFF vs. HF	CGenFF vs. MP2
angle difference		16°		7°	14°

Table 7

Comparison between CGenFF charges assigned by the present algorithm and standard CHARMM charges for compound **3**.

	CGenFF	CHARMM	Difference	Penalty
C1	−0.118	−0.115	−0.003	24.49
H1	0.115	0.115	0	0
C5	−0.118	−0.115	−0.003	24.49
H5	0.115	0.115	0	0
C2	−0.11	−0.115	0.005	0
H2	0.115	0.115	0	0
C4	−0.11	−0.115	0.005	0
H4	0.115	0.115	0	0
C3	−0.115	−0.115	0	0
H3	0.115	0.115	0	0
C6	−0.022	0	−0.022	39.41
C7	−0.053	−0.02	−0.033	37.51
H71	0.09	0.09	0	0
H72	0.09	0.09	0	0
N8	−0.432	−0.47	0.038	37.02
H8	0.315	0.31	0.005	18.09
C9	0.517	0.51	0.007	12.78
O9	−0.51	−0.51	0	0
C10	−0.269	−0.27	0.001	0
H101	0.09	0.09	0	0
H102	0.09	0.09	0	0
H103	0.09	0.09	0	0