J-CAMD 289

MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry

Paul R. Gerber* and Klaus Müller

Pharmaceutical Research and Development, F. Hoffmann-La Roche AG, CH-4002 Basel, Switzerland

Received 19 October 1994
Accepted 28 November 1994

Keywords: Force field; Potential energy function; Molecular structure; Molecular modelling

Summary

The mathematical formulation, parametrization scheme, and structural results of a new, generally applicable molecular force field are presented. The central features are a scheme for automatic parameter assignments, the consistent united-atom approximation, the absence of atom types other than elements, the replacement of electrostatic terms by geometrical hydrogen-bonding terms, the concomitant lack of a need for partial atomic charge assignment and the strict adherence to a finite-range design. As a consequence of omitting all hydrogen atoms, optimal hydrogen-bond patterns are computed dynamically by appropriate network analyses. For a test set of 1589 structures, selected from the Cambridge Structural Database solely on the grounds of a given element list and criteria for high structure refinement, the agreements are on average 2 pm for bonds, 2° for valence angles and 10 to 20 pm for the root-mean-square deviation of atom positions, depending somewhat on size and flexibility of the structures. More qualitative testing of large-scale structural properties of the force field on proteins and DNA oligomers revealed satisfactory performance.

Introduction

All-atom force fields have developed in a variety of versions to standard tools in molecular modelling [1–8]. They provide reasonably accurate descriptions of molecular geometries and energies. Depending on the initial focus of the designer, each force field has its strengths and qualities for particular classes of molecular structures and systems. With the commercialization of modelling packages, substantial efforts have been made to extend the application range of force fields. Unfortunately, this resulted in a concomitant increase of the number of atom types per element, such that a user has to be increasingly aware of the rich possibilities and intricacies of atom type assignments in cases where automatic assignments fail. Apart from this practical aspect, the more fundamental question of additivity of chemical functionalities arises in cases where conjugative effects come into play. While such problems could be addressed by ad hoc modifications or extensions of the parametrization schemes for individual structural subsets, we have felt the need for a more generally applicable force field, able to cope with the bewildering richness of molecular structures encountered in the pharmaceutical research environment.

This need prompted us to develop a parametrization scheme that avoids the notion of chemical functionality altogether and tries to tackle the problem at a more fundamental level. Chemical functionalities and the corresponding parameter values that enter the energy expressions are properties derived from more generic underlying principles and, consequently, from a much more limited set of basic parameters.

Regarding nonbonded interactions, published force fields are similarly demanding with respect to assigning atomic charges, as well as choosing from the wide and ill-established possibilities of modified Coulomb interaction forms. For instance, some force fields attribute hydrogen bonding entirely to electrostatic interactions, not without assigning special charge values to the atoms involved. Others make use of additional energy terms. While for commonly encountered functional groups the assignment of partial atomic charges is often well established and

*To whom correspondence should be addressed.

tested for consistency within a given force field, less common cases are subject to substantial arbitrariness. Furthermore, variable cutoff distances for nonbonded interactions enlarge the spectrum of adjustable parameters and lead to an irritating richness, not to say confusion, of variability. The fast-decaying dispersion interaction has been less subject to argumentation and has become fairly standardized by the use of Lennard-Jones-type interaction forms, although there is little reason to be too confident in this matter.

With the intention to model the predominant aspects of hydrogen-bonding interactions in a consistent way, in accordance with experimental findings, we introduce a purely geometrical energy term for this type of interaction. Other charge effects are discarded in order to eliminate the need for partial charge distributions and to avoid the uneconomical effort to recalculate the shielding of charges, so effectively achieved in nature by the long-range Coulomb law.

In the following, first we present the energy expression in some detail, because some of the important terms deviate from conventional ones. Then, we present the parametrization scheme with its two layers of parameters, a set of basic parameters adhering to physical concepts, and a set of derived parameters which respond to actual structural contexts and eventually enter the potential energy expressions. Finally, the performance of the force field is illustrated and documented with the aid of a series of test compounds.

Energy expression

Since efficiency is a central issue in the performance of the force field, we work in principle with united atoms. However, there is no problem in carrying along explicit hydrogen atoms, and there are systems for which it is even advisable to do so. Nevertheless, this is more the exception than the rule, and we stick to the united-atom simplification for this presentation.

The total energy E of a molecular system is modelled in our MAB force field by the expression

$$E = E_b + E_v + E_\tau + E_p + E_{14} + E_{1n} + E_{hb} \qquad (1)$$

The first four terms constitute the valence energy terms, i.e., the bond stretching energy E_b, the valence angle bending energy E_v, the dihedral angle distortion energy E_τ and the pyramidality deformation energy E_p. The nonbonded interaction consists of the dispersion interaction energies E_{1n} and E_{14} and the hydrogen-bonding energy E_{hb}. The (1,4) interaction could be equally well attributed to the bonded interaction set, but we prefer the given assignment because of its functional similarity to the (1,n)-type nonbonded interaction term. The historical name MAB was derived from an early concept of the force field which envisaged a strict nearest-neighbor principle among atoms and bonds.

Valence energy terms

The bond-stretching expression is purely harmonic; it sums over all bonds determined by the topology of the chemical system:

$$E_b = \sum_i^{\text{bonds}} A_{bi}(b_i - b_{0i})^2 \qquad (2)$$

where b_i is the actual bond length of bond i, and b_{0i} and A_{bi} are derived parameters, representing actual reference values of the bond length and force constant.

Similarly, valence angle distortions in the small distortion regime are described by a harmonic potential:

$$E_v = \sum_i A_{vi}(v_i - v_{0i})^2 \qquad (3)$$

Here, the summation index i runs over all valence angles. Deviations of the actual value of the valence angle v_i from the reference value v_{0i} are penalized with a bending force constant A_{vi}. For larger distortions, the quadratic potential leads to too strong a rise in energy. To avoid this undesired feature, we introduce two measures. If $v < v_0$, each term in Eq. 3 is multiplied by the following function in $\Delta v = (v - v_0)$:

$$f_v(\Delta v) = \frac{1 + \log\left(1 + \frac{|\Delta v|}{\Delta v_{th}}\right)}{1 + \frac{|\Delta v|}{\Delta v_{th}}} \qquad (4)$$

in which Δv_{th} is a characteristic difference angle value at which a crossover from quadratic to linear behaviour takes place. For $v > v_0$, the following polynomial form in the variable $x = \pi - v$ applies:

$$E_v = A_{v0} + A_{v2}x^2 + A_{v8}x^8 \qquad (5)$$

in which the coefficients A_{vi} are chosen such that the value of E_v as well as its first and second derivatives match Eq. 3 at $v = v_0$.

The dihedral angle distortion energy is also of a conventional nature:

$$E_d = \frac{1}{2}\sum_i \sum_k^{M_i} D_{ki}[\cos(m_{ki}\phi_i - \phi_{0ki}) - 1] \qquad (6)$$

in which the summation index i enumerates all dihedral angles ϕ_i. In order not to make the results dependent on a particular choice of a single angle per bond, all possible dihedral angles occur in this sum, i.e., for a bond between two atoms with n and m additional ligands there are n×m contributions. For each angle, M_i expressions contribute to the energy. The integer multiplicity m_{ki} determines the number of minima encountered on a full turn and ϕ_{0ki} determines their position. The amplitudes or rotational barriers are given by D_{ki}. In most cases a single term is required, and the shifts ϕ_{0ki} are always such that $\phi = 0$ corresponds to a minimum or a maximum of the expression.

Planar geometries at sp^2-hybridized atoms are not sufficiently stabilized by the valence angle bending energy. Conventionally, deviations from planarity are restored by energy contributions that are quadratic in improper dihedral angles defined by the planar center and its three ligands. In order to avoid the inherent ambiguities in labelling the ligands, which can lead to slightly varying results in strained conformations, we use the following energy expression to restore planarity:

$$E_p = \sum_{i}^{sp^2} A_{pi} p_i^2 \qquad (7)$$

which is quadratic in the pyramidality p, defined by Eq. 8:

$$p = \det(\hat{b}_1 \hat{b}_2 \hat{b}_3) \qquad (8)$$

where the hatted quantities b_i are unit vectors to or from the three ligands. For atoms on the edge between sp^2 and sp^3 hybridization, the term in Eq. 7 is somewhat generalized to:

$$e_p = A_{p2} p^2 + A_{p4} p^4 \qquad (9)$$

This expression offers the possibility to describe double minima, a useful feature for nitrogen atoms in conjugated environments.

As a technical aside, we mention that bond unit vectors are calculated at the outset of each energy calculation and are then available in all valence energy terms, as well as in the hydrogen bond expressions (see below).

Dispersive interactions

The (1,4)-type repulsion accounts for the energy difference of trans and gauche conformers, and is singled out from the general (1,n)-type dispersion in order not to have the parametrization of the latter be influenced by the peculiarities of the former. Its formal expression employs the following distance function, in which the distance of the pair of atoms is r_p:

$$R(r_p) = \begin{cases} (r_c - r_p)/(r_c - r_0) & \text{for } r_p < r_c \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

in which r_c is a threshold distance beyond which the function vanishes, and r_0 determines its rise with decreasing values of r_p. The value of r_0 must be smaller than that of r_c. The actual values of r_0 and r_c are determined by the types of the two atoms in question.

In terms of the function R, the energy expression reads:

$$E_{14} = A_{14} \sum_{p} [R(r_p)]^4 \qquad (11)$$

where the sum runs over all pairs p of atoms which are connected via three bonds and have no other shorter connection. The parameter A_{14} is a pair-independent constant.

The general (1,n)-type nonbonded interaction contributions also employ the distance function of Eq. 10 and are thus strictly of finite range. The functional form of all our interaction potentials for nonbonded interactions are designed to be of finite range. Consequently, no cutoffs, with all their concomitant potential artifacts, need to be introduced. Our actual functional form of the interaction term reads

$$E_{1n} = \frac{1}{7} \sum_{i,j}^{\text{pairs}} e_{m,ij} \{[R(r_{ij})]^{16} - 8\,[R(r_{ij})]^2\} \qquad (12)$$

The sum runs over all pairs of atoms (i,j) which have no topological path of three or less than three bonds between them. The functional form of the single terms is chosen such that the parameter $e_{m,ij}$ represents the depth of the energy minimum. This minimum is positioned at $R = 1$ or, equivalently, at $r_{ij} = r_0$. The two distances r_0 and r_c are sums of corresponding atom-specific radii, while $e_{m,ij}$ is the geometrical mean of atom-specific energy parameters.

In contrast to the conventional Lennard-Jones-type potential which is fixed by two parameters, i.e., the energy minimum and its position, the form of Eq. 12 has three parameters (r_0, r_c and e_m). This feature offers the flexibility to adjust the rise in the small-distance repulsive domain independently of the position and depth of the energy minimum (deeper minima need not go together with harder cores!). However, for most cases we have not bothered to refine literature values of the parametrization and have chosen the ratio r_c/r_0 as well as the values of the two exponents in Eq. 12 so as to optimally reproduce the functional form of the conventional 6–12 potentials. Figure 1 gives an example of the mutual agreement.

For distances r_{ij} smaller than $(8r_0 - 3r_c)/5$ (i.e., $R > 8/5$) the functional form is changed to a harmonic potential in such a way that the function and its derivative remain continuous. This has virtually no effect on relevant energy values, but is advantageous in that it makes the model considerably more stable in molecular dynamics runs, for which larger step sizes can be used to save computational time.

Hydrogen bonding energies

Since this part of our force field deviates most distinctly from conventional ones, it is appropriate to describe it in some more detail. The prominent features are the strict adherence to the united-atom approximation, the directionality of the hydrogen bonding geometry, and the evaluation of an optimal hydrogen bonding pattern.

Experimental findings [9] give evidence that hydrogen

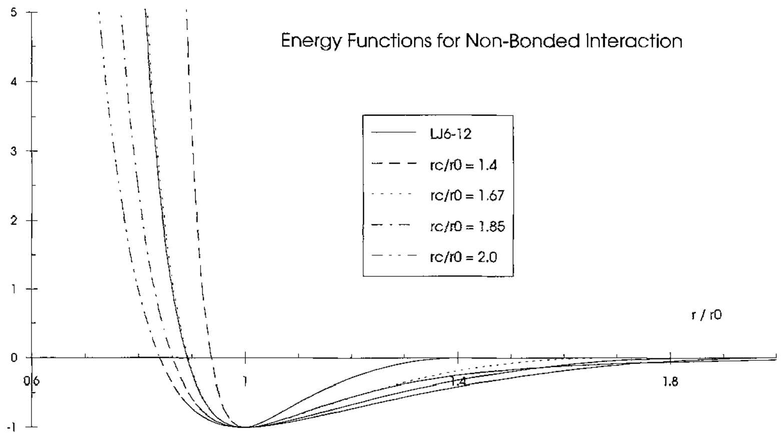

Fig. 1. Graph of a conventional 6–12 Lennard-Jones potential together with a series of potentials corresponding to the functional form of Eq. 20 for varying values of r_c / r_0 (Eq. 10). Positions and values of the minima are identical for all curves.

bonds preferentially show a linear geometry, suggesting that it may be possible to construct a suitable energy expression for a hydrogen bond without carrying along the coordinates of the hydrogen atom. However, omitting the hydrogen atoms requires the donor site of the H-bond to account for the directionality of the now omitted covalent donor-hydrogen bond. Similarly, the acceptor site should show an analogous directionality [9], owing to the H-bond-promoting lone pair orientations. This somewhat symmetrical treatment of donor and acceptor sites can be rationalized by recalling the resonance state description of the H-bond, in which the proton resides in part on the lone pair of the acceptor. Lone pair directionality has been accounted for previously [10]. The following expression is used for the energy of a single H-bond:

$$e_h = A_h B_r(r) \Theta_d(\hat{r}) \Theta_a(\hat{r}) \tag{13}$$

where r is the distance between the donor (d) and the acceptor (a) atom, and the hat indicates a unit vector in the direction from the donor to the acceptor.

The energy is determined by the parameter A_h, while the distance dependence follows a bathtub-shaped function:

$$B_r(r) = -\left[1 - \left(\frac{r - r_{h0}}{w_r}\right)^n\right]^m \tag{14}$$

This expression applies whenever the expression in the outer brackets is positive; otherwise B vanishes. The exponents n and m govern the curvatures of the potential function near the minimum and at the edges of the function (see Fig. 2) and have been chosen as n = 2 and m = 4. The parameter r_{h0} is the equilibrium distance between the donor and acceptor sites, while w_r (which generally has different values for $r > r_{h0}$ and $r < r_{h0}$) determines the range or width of the tub. The (1,n)-type dispersion energy (Eq. 12) remains in effect, although with a slightly reduced range, and still provides the short-distance repulsion.

The directionality is governed by the function Θ, the form of which depends on the valence states of the atoms in question. For isolated atoms (most importantly the

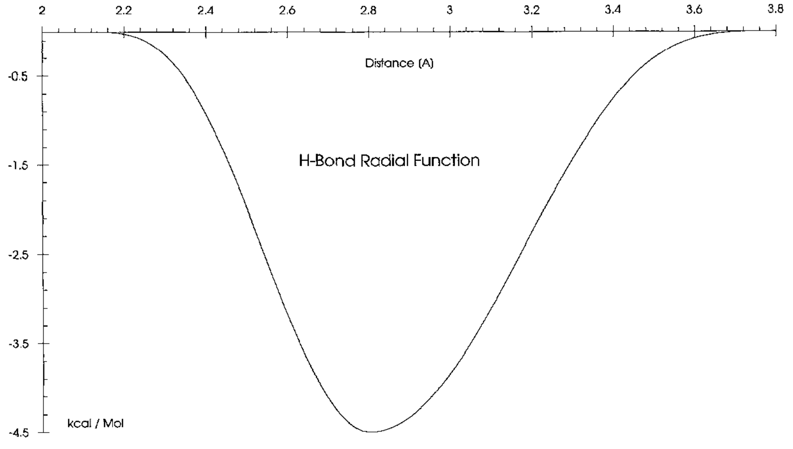

Fig. 2. Graph of the distance dependence of the radial energy function for the hydrogen bond potential.

oxygen atoms of water) it is equal to one, while for covalently bound atoms the following product expression applies:

$$\Theta(\hat{r}) = \prod_{i}^{\text{bonds}}\left[1-\left(\frac{\alpha_{ri}-\alpha_0}{\beta}\right)^2\right]^2 \tag{15}$$

where α_{ri} is the angle between the argument unit vector and the bond vector of bond i. The function is taken to vanish whenever the expression within either one of the outer brackets in the product becomes negative. The parameter α_0 fixes the optimal direction and depends on the valence state of the atom. The softness of the potential is controlled by the parameter β, which depends not only on the atom state, but also has varying values depending on whether α is larger or smaller than α_0.

In case of sp^2 hybridization of the donor or acceptor atom, Eq. 15 is multiplied by an additional factor Ψ, accounting for the influence of the π-electron system, i.e., penalizing deviations from coplanar arrangements which reduce π-conjugation at the donor site. Its functional form is given by Eq. 16:

$$\Psi(\hat{r}) = \left[1-\left(\frac{\det(\hat{b}_1,\hat{b}_2,\hat{r})}{w_p}\right)^2\right]^2 \tag{16}$$

If the expression within the brackets is negative, Ψ is taken to be zero and the whole term vanishes. The unit vectors b_i point along the bonds of the atom. In the (frequently occurring) case where the atom is only singly coordinated, the second unit vector points towards the next-nearest neighbor. If there are two of them, their average contribution to Ψ is taken to avoid ambiguities. The parameter w_p controls the softness of this plane bending correction and depends on the coordination number.

To obtain the total H-bond energy, all possible single H-bond contributions are evaluated and from this set, that H-bond pattern is determined which minimizes the total H-bond energy and accounts for the maximum H-bond capacities of donor and acceptor sites. In other words, no donor (acceptor) can occur more often as a donor (acceptor) site for H-bonds than its number of hydrogens (lone pairs) indicates. For highly complex H-bond networks, the computational efforts to explore all possible patterns may grow exponentially with increasing size of the network. Thus, by default we employ the approximation described in the next paragraph, though the exact algorithm is also available.

The approximation in our pattern evaluation starts from a sorted list of single H-bonds. Working up from the weakest ones, we eliminate each hydrogen bond which violates a donor or acceptor limit. Having no violations left, we work back on the eliminated H-bonds and reinstate those for which the donor and acceptor sites both have gained a vacancy.

This way of calculating the H-bond energy avoids the fixation to an a priori pattern as given by an initial spatial distribution of polar hydrogen atoms. Thus, switches of hydrogen bond patterns are possible and indeed frequently occur in the course of an energy minimization.

This scheme cannot account for bifurcated hydrogen bonding. However, in certain situations increasing the hydrogen count of heavy atoms may provide an acceptable work-around.

Optional constraints

Various types of optional constraints can be applied to the system. They are needed at several stages in molecular modelling activities.

The simplest form of a constraint is to have atoms exempt from participating as dynamical variables in the minimization or optimization process. Their coordinates are kept fixed, but occur in the energy terms. A related tool is the application of positional constraints which, when acting on an atom, subject it to a harmonic force proportional to its distance from the center of the constraint. Still more general, a constrained distance can be set between two atoms both participating in the process of optimization. The constraining potential has the following general form:

$$U_d(d) = \frac{1}{2}u_d(d-d_0)^2 \tag{17}$$

where d_0 is the target distance, and the force constant u_d can be non-zero for $d > d_0$ only (upper constraint), for $d < d_0$ only (lower constraint), or for any value of d (two-sided). This type of constraint is useful in structure determination from experimental data (NOE) as well as in the process of positioning and docking of molecules.

Torsional angles ϕ are constrained to a target value ϕ_0 by an expression of the form given in Eq. 18:

$$c(\phi,\phi_0) = (\cos\phi - \cos\phi_0)^2 + (\sin\phi - \sin\phi_0)^2 \tag{18}$$

Depending on whether a penalty-free region of $\pm\Phi$ is required or not, one of the following expressions for the penalty is applied. Without a free range ($\Phi = 0$), we use:

$$U_t = u_t\, c\,(\phi,\phi_0) \tag{19}$$

otherwise the expression reads

$$U_t = \begin{cases} u_t\,(c(\phi,\phi_0)-c_0)^2 & \text{for } |\phi-\phi_0|>\Phi \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where

$$c_0 = \left(2\sin\left(\frac{\Phi}{2}\right)\right)^2 \tag{21}$$

Torsional constraints are useful in the initial stages of relaxation of highly deformed conformations, where the danger exists that cis–trans isomerization might occur.

Finally, pyramidal constraints have an energy penalty of the form:

$$U_p = \begin{cases} u_p(|p - p_0| - p_f)^2 & \text{for } |p - p_0| > p_f \\ 0 & \text{otherwise} \end{cases} \qquad (22)$$

where p_0 is the target value and p_f again a range in which the constraint has no effect.

Pyramidal constraints are important to ensure the preservation of the proper stereochemistry during relaxation of highly deformed initial structures. This is particularly important in united-atom force fields, in which stereo inversion is much facilitated when one of the ligands is a hydrogen atom. The introduction of free ranges offers the possibility to apply constraints that prevent unwanted flips of conformation, but do not operate in the small-deformation domain.

Generic parametrization

In a force-field model, it is understood that the parameters which occur in the energy expressions are independent of the actual molecular geometry. Thus, the parameter values are determined entirely by the molecular topology.

In most currently available force fields, the parameter assignment is done on a chemical basis, with parameter values optimized for a particular type of molecule or functionality. This procedure has the advantage of high accuracy, at the cost of generality. As an additional consequence, usually a large variety of atom types is needed, each corresponding to a particular structural context. The number of parameters for bond stretching and angle bending usually grows respectively with the second and third power of the number of atom types, unless, as not uncommon, values are simply missing. For the dihedral potentials, the explosion of the number of parameter values is usually counteracted by introducing generic values in the sense that the values are determined by the types of the two central atoms alone. This is certainly a welcome simplification, although it affects just those degrees of freedom that are most susceptible to energy input on the scale of thermal energies (kT).

Evidently this situation is of no concern for work on molecules within a common family such as e.g. proteins and peptides. However, for the broad spectrum of compounds encountered in pharmaceutical research, use of such force fields requires much skill in choosing the appropriate atom types and in defining, often in an ad hoc fashion, appropriate values for missing parameters.

In order to overcome this difficulty, we proceed along a more generic line, which emphasizes the physical aspect of this parametrization problem. The procedure can be divided into two steps. The first entails determination of the hybridization states of the atoms and of the bond orders of all bonds; the second consists of the derivation of actual parameter values from these quantities. Furthermore, we stick as far as possible to a nearest-neighbor-type concept among atoms and bonds. This means that bond stretching parameters should only depend on the characteristics of the bond and the adjacent atoms (true in all force fields), valence angle bending parameters only on the characteristics of the central atom and the adjacent bonds, etc.

In the following sections formulae appear that contain parameters. Values for these parameters are found in the tables of the Appendixes; the number of the Appendix is indicated in each case. Appendix 1 contains basic atom-related parameters.

Atom hybridizations and topological bond order (Appendixes 1 and 2)

In a first round, all atoms for which the number of ligands l (including virtual hydrogen atoms) is smaller than their formal valence v are grouped into a set of unsaturated atoms. Their state of hybridization is set to

$$h = 3 + l - v \qquad (23)$$

In the next round, all atoms that are connected to an unsaturated one and which have a formal valence of less than four are added to the set. They are assigned a hybridization state h = 2. This step is repeated until no new atoms are added. Finally, the hybridization of doubly coordinated nitrogen atoms, bonded to an sp-hybridized atom, is set to one.

In the next step, this set of atoms is grouped into subsets of directly connected atoms. Between the subsets no direct bond exists. For each of these subsets bond orders are determined from a recursive Hückel-type calculation.

This calculation starts off in the well-known way [11, 12] by assigning diagonal elements α_i to the π-electron Hamiltonian according to:

$$\alpha_i = \alpha_{0i} + \alpha_e \frac{(n_{ei} - 1)}{n_{np,i}} \qquad (24)$$

where α_{0i} is a value specific to the core number of atom i, n_{ei} is the number of π-electrons contributed by the atom, and $n_{np,i}$ the number of its neighbours within the π-system.

Off-diagonal elements are initialized to a random value between a minimal value β_0 and one, when a bond exists between two atoms, and to zero otherwise. The randomization is applied in order to avoid any symmetry in the initial conditions which would then not change in subsequent iterations.

The Hückel matrix is diagonalized and atom occupation numbers n_i as well as π-bond orders p_{ij} are calculated. No bond order is introduced for pairs of atoms that are not covalently linked.

More importantly, the off-diagonal elements are modified such that higher bond orders lead to increased values of the matrix elements:

$$\beta_{ij} = \{\beta_0 + (1 - \beta_0)p_{ij}\}R_{sr} \qquad (25)$$

where R_{sr} is a reduction factor for the case that one of the atoms i or j is from the second row in the Periodic System.

After each diagonalization of the Hückel matrix, a new one is set up by inserting the resulting n_i's and p_{ij}'s in Eqs. 24 and 25. This cycle is terminated when self-consistency is reached. The final bond orders, p_{ij}, and the partial atomic π-charges, q_i, as obtained from the population analysis, are then used to obtain the derived force-field parameters (see next section). The carbonyl bond orders appeared systematically low. Thus, they are corrected according to Eq. 26:

$$b(CO) = p_H + (1 - p_H)B_{CO} \qquad (26)$$

where p_H is the bond order as it emerges from the Hückel calculation.

A remark on our operational policy may be in order here. The introduction of expressions like Eqs. 24 or 25 is usually initiated by some deficiency of the force field. The actual form and the variables entering the correction expressions are often justified by intuitive chemical argumentation. However, we felt free to modify these expressions in unconventional ways whenever the overall quality of the force field could be improved by doing so.

Force-field parameters for valence energy terms

In order to outline the rationales for introducing the various functional dependencies of the force-field parameters on topological and atomic properties, we proceed in a deductive way. Starting from simple functional forms, we add improvements as required to fit experimental structural data taken mostly from the tables of Landolt-Börnstein [13] and the Cambridge Structural Database (CSD) [14].

Bond stretching parameters (Appendix 3)

The reference bond length b_0 in Eq. 2 is in principle a sum of bond radii of the two adjacent atoms. However, the bare radii r_{b0} as listed in Appendix 1 are modified by the structural context of the atoms:

$$r_b = r_{b0}\,(1 + R_v\,i_v) \qquad (27)$$

In this expression i_v is the excess of the actual number of ligands (including virtual H-atoms) over the formal ligand number (as observed for uncharged single-, double- and triple-bonded atoms). Positive values lead to an increasc in bond radius. The most prominent case is nitrogen in a fourfold coordination or in a pyrrol-like structural context.

The sum of the two bond radii (Eq. 27) is augmented by the following expression if both atoms have lone pairs (n_{LP}):

$$\Delta b_{LP} = B_{LP}\,n_{LP,1}\,n_{LP,2} \qquad (28)$$

For single bonds, the final expression for the reference bond length reads:

$$b_0 = (r_{b1} + r_{b2} + \Delta b_{LP})\,(1 - B_h|h_1 - h_2|) \qquad (29)$$

where a reduction is made if one of the atoms is not sp^3 hybridized.

For higher order bonds we apply a reduction factor f_h if the two atoms are both from the first row but have different core numbers:

$$f_h = 1 + B_d\,(i_{v1} + i_{v2} - 1) \qquad (30)$$

where the excess ligand number i_v has been introduced in Eq. 27.

Finally, the whole expression is multiplied by a quadratic function of the π-bond order p_b, to yield, together with Eqs. 29 and 30, the following reference bond length:

$$b_0 = b_{0s}\,f_h\,(1 - B_1\,p_b + B_2\,p_b^2) \qquad (31)$$

where b_{0s} is the value obtained from Eq. 29.

Multiple bonds to four-coordinated hypervalent second-row atoms are treated separately, although less systematically. The sum of the bare radii is reduced by a single decrement:

$$b_0 = (r_{b0,1} + r_{b0,2})\,(1 - R_a) \qquad (32)$$

The force constants A_b (Eq. 2) are of little importance in the cases which are relevant for our applications, because of their large values. They are taken to be linearly dependent on the bond order p_b:

$$A_b = A_{b0}\,(1 + A_{b1}\,p_b) \qquad (33)$$

Note that, up to now, we have defined only about a dozen parameters from which all bond stretching parameters for the force field are derived.

Valence angle bending parameters (Appendix 4)

The reference values for the valence angles v_0 appearing in Eq. 3 are obtained from a basic value V_h, specific for the state of hybridization h of the central atom. For sp-hybridized atoms this value is always 180°.

The case of an sp^2-hybridized atom is more involved. To account for the influence of electronegativity [15] and partial charges (as emerging from the Hückel calculation), excesses in electronegativities χ and π-charges q of the two ligands to the central atom are used to modify a standard value for the angle at the central atom:

$$\begin{aligned} \Delta\chi_l &= \chi_l - \frac{1}{L}\sum_{j=1}^{L}\chi_j \\ \Delta q_l &= q_l - \frac{1}{L}\sum_{j=1}^{L} q_j \end{aligned} \qquad (34)$$

where the sum runs over all ligands. For the electronegativity, the sum also includes a contribution from a lone pair, if present. Furthermore, we introduce the average π-bond order p_a about the atom as well as a relevant π-bond order p_r as follows:

$$\begin{aligned} p_a &= \frac{1}{L}\sum_{l=1}^{L} p_l \\ p_r &= \frac{(p_a L + 2p_m)}{3} \end{aligned} \qquad (35)$$

where p_m is the maximum π-bond order to the atom. With the average order p_a, our expression for the reference valence angle reads:

$$\begin{aligned} v_0 = {} & V_{0,sp^2} + V_\chi(\Delta\chi_1 + \Delta\chi_2) \\ & + V_q(\Delta q_1 + \Delta q_2) + V_p(p_1 + p_2 - 2p_a) \end{aligned} \qquad (36)$$

This value is reduced by V_{LP} if a lone pair is present. A further correction is applied to cases in which the coordination number is higher than the formal one, corresponding to sp^2 hybridization (e.g. pyrrol nitrogen). These cases are assumed when the maximum π-order p_m of all bonds to the central atom is below a threshold value p_{th}, resulting in a reduced tendency to planarity (see the following section and Eq. 9). Consequently, the corresponding reference valence angles are reduced by:

$$\Delta v_0 = (v_0 - V_{0,sp^3})\left(1 - \frac{p_m}{p_{th}}\right)^2 \quad \text{for } p_m < p_{th} \qquad (37)$$

In the case of a saturated sp^3-hybridized central atom, the reference value v_0 starts off from the near-tetrahedral value V_{0,sp^3}. A first correction is applied to this value whenever the central atom is involved in a three- or four-membered ring. The correction V_r applied in this case depends on the ring size and on whether or not the angle is endocyclic.

A second type of correction concerns the second and higher row atoms. For these, a reduction by a factor V_{r2} is made whenever the coordination is below four; otherwise the reference angle is set to the tetrahedral value V_{0,sp^3}.

For the force constants A_v (Eq. 2) a hybridization-dependent basic value A_h is used. For an sp^2-hybridized atom, this value is modified according to:

$$A_v = A_{sp^2}(1 + p_1 + p_2) \qquad (38)$$

Furthermore, this value is multiplied by the factor A_{12} if the atom contains an implicit hydrogen. All angle bending force constants are finally augmented by a row-dependent factor, $1 + A_{rw}(r - 1)$, in which r is the row of the central element.

Finally, for four-coordinated second and higher row elements the force constant is substantially reduced by a factor A_{rw4}.

Pyramidality parameters (Appendix 5)

The reference value for pyramidality at sp^2-hybridized atoms is always zero. In normal cases, only the harmonic part of the potential (Eq. 9) is present. Its force constant is given the following value:

$$A_p = A_{sp^2} p_m \qquad (39)$$

However, if the actual valence is higher than that corresponding to the atom in its hybridization state, a double minimum is possible when the relevant π-bond order p_r (Eq. 35) is below the threshold value p_{th}. In this case we take:

$$\begin{aligned} A_{p2} &= A_{sp^2}(p_r - p_{th}) \\ A_{p4} &= \begin{cases} A_4(p_{th} - p_r) & \text{for } p_r < p_{th} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \qquad (40)$$

Parameters for dihedral angle potentials (Appendix 6)

These terms are very important, because substantial conformational distortions can be induced by small input energies.

As mentioned in connection with Eq. 6, all possible atom quadruples specifying the torsion about a given bond are considered as separate dihedral terms with equal potential. This allows us to describe the parameters bondwise. Each term is characterized by a multiplicity m and a phase parameter ϕ_0 locating the positions of the energy maxima. However, ϕ_0 is always selected so that a dihedral of zero corresponds to either a maximum or a minimum of the potential. Consequently, this information can be packed into the sign of the multiplicity m. A positive value of m signifies the case of a maximum at $\phi = 0$. Apart from a few exceptions, only a single potential type is attributed to a bond.

If both central atoms are sp^3 hybridized, we apply a multiplicity of m = 3 and an amplitude of the form

$$D = T_3(3 - \Delta g) \qquad (41)$$

in which Δg is the absolute difference between the element groups in the Periodic Table to which the two atoms belong. Alternative values are only used when the two atoms are both in group 5 or both in group 6. Then we take m = 2. Disulfides represent the most prominent case of this type, with $D = T_{ss}$.

If just one of the two central atoms, say number three in Fig. 3, is sp^2 hybridized, the multiplicity becomes m = –3 while the amplitude of the potential depends on the π-bond order p_{adj} of the unsaturated bond to atom four. It takes the following value:

$$D = T_{23}\, p_{adj} \qquad (42)$$

If two adjacent bonds are present with equal bond order, they usually counteract each other to the extent of yielding no net torsion potential.

Again, if the number of ligands of the sp^2-hybridized atom is increased over the formal value as defined according to Eq. 27, we employ a switching between sp^2- and sp^3-type behaviour at the threshold condition $p_{adj} = p_{th}$. In this case the torsional barrier touches zero with the functional form

$$D = T_{23}\sqrt{|1 - (p_{adj}/p_{th})|} \qquad (43)$$

The multiplicity takes the value m = 3 for $p_{adj} < p_{th}$, and m = –3 otherwise. The value of T_{23} applies for both cases.

If both central atoms are sp^2 hybridized, the multiplicity is always m = –2. The amplitude of the torsional potential has essentially a linear dependence on the π-bond order p, with a quadratic correction:

$$D_2 = T_2 p\,(1 + T_{2q}(p - 1)) \qquad (44)$$

If there are two adjacent bonds with non-zero π-bond order on any of the two central atoms, Eq. 44 is modified by the following correction factors:

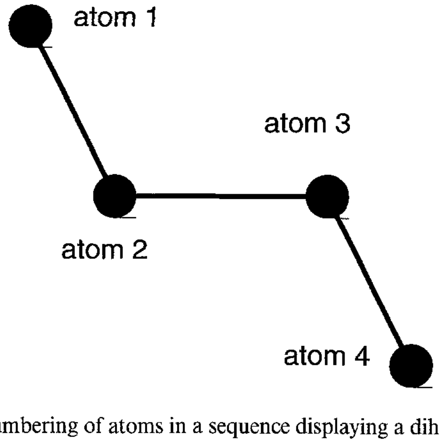

Fig. 3. Numbering of atoms in a sequence displaying a dihedral angle.

$$D = D_2\left[1 - T_{2m}\left(1 - \frac{\Delta p_{adj}}{\sum p_{adj}}\right)_1\right]\left[1 - T_{2m}\left(1 - \frac{\Delta p_{adj}}{\sum p_{adj}}\right)_2\right] \qquad (45)$$

where p_{adj} are the π-bond orders of the adjacent bonds at the atom indicated by the subscript of the bracket. The difference Δ between the two bond orders is taken to be positive.

To account for anomeric effects, a torsional potential of multiplicity m = 2 is added if the following conditions are met. (i) Both central atoms are sp^3 hybridized; (ii) atom 2 (see Fig. 3) belongs to group 4 in the Periodic Table, atom 3 to group 6; and (iii) if the group of atom 1 is higher than 4, then the torsional potential has the amplitude:

$$D = T_{am}(\chi_1 - \chi_2) \qquad (46)$$

where χ_j is the electronegativity of atom j.

Allenes and cumulenes have dihedral angle terms that involve the atoms at either end of the cumulated bond, together with an outer ligand of each of these atoms. In this case the multiplicity alternates with the number n of cumulated bonds between the values m = –2 and m = 2, the latter being the value for allene (n = 2). Our torsion potential decreases exponentially with n, according to:

$$D = T_2(2)^{-n+1} \qquad (47)$$

Parameters for dispersive interactions (Appendix 7)

The range-determining distance parameters r_0 and r_c of Eq. 10 are sums of corresponding atom-specific values r_{0j} and r_{cj}. The two are related by an atom-specific factor:

$$r_{cj} = C_j\, r_{0j} \qquad (48)$$

The values of the minimum radii r_{0j} depend on the number n_h of implicit hydrogen atoms lumped together in the heavy atom within the united-atom approximation, as well as on the number of lone pairs n_{LP}:

$$r_{0j} = R_{0j} + W_h\, n_{hj} + W_{LP}\, n_{LPj} \qquad (49)$$

In the (1,4)-type interaction, these values are multiplied by a constant factor R_{14}. Furthermore, if these atoms are involved in a non-vanishing hydrogen bond interaction, the corresponding interaction distances r_0 and r_c are reduced by a factor R_{hb}.

In the case of the (1–4)-type interaction the energy parameter A_{14} is independent of the two atoms involved, whereas for the dispersive (1,n)-type nonbonded interaction potential (Eq. 12), the minima values e_m of the energy are taken as geometric means of element-specific quantities of the two atoms (i,j) involved:

$$e_{m.ij} = \sqrt{E_{mi}E_{mj}} \qquad (50)$$

TABLE 1
COMPARISON OF MINIMIZED STRUCTURES WITH EXPERIMENTAL TEST STRUCTURES

Name	Molecular formula	Heavy atoms	rmsd (pm)	rmsb (pm)	maxb (pm)	rmsa (°)	maxa (°)	Reference code
Butane	C_4H_{10}	4	2.22	0.087	0.093	2.290	2.301	LB 551
Ethane	C_2H_6	2	0.20	0.400	0.400	0.000	0.000	LB 274
Propane	C_3H_8	3	1.35	0.609	0.616	1.871	1.871	LB 428
Dimethylether	C_2H_6O	3	1.08	1.612	1.624	1.180	1.180	LB 297
Ethanol	C_2H_6O	3	2.53	1.464	1.999	2.721	2.721	LB 296
Methanol	CH_4O	2	0.06	0.100	0.100	0.000	0.000	LB 124
Methylamine	CH_5N	2	0.25	0.500	0.500	0.000	0.000	LB 129
Piperazine	$C_4H_{10}N_2$	6	2.14	1.232	1.538	1.009	1.749	LB 555
Methylmercaptan	CH_4S	2	0.10	0.200	0.200	0.000	0.000	LB 125
Methylphosphine	CH_5P	2	0.11	0.200	0.200	0.000	0.000	LB 130
Dimethyldisulfide	$C_2H_6S_2$	4	6.32	3.633	6.227	4.585	4.589	LB 301
Dimethylsulfide	C_2H_6S	3	0.90	1.413	1.426	0.785	0.785	LB 300
Dimethylsulfone	$C_2H_6O_2S$	5	0.85	1.070	1.488	0.314	0.674	LB 299
Dimethylsulfoxide	C_2H_6OS	4	1.94	1.104	1.826	1.064	1.765	LB 298
Trimethylphosphineoxide	C_3H_9OP	5	0.82	0.778	1.549	0.187	0.198	LB 452
Benzene	C_6H_6	6	0.49	0.491	0.508	0.014	0.021	LB 667
Butadiene	C_4H_6	4	0.69	0.402	0.492	0.545	0.553	LB 500
2-Butene	C_4H_8	4	1.91	0.979	1.414	1.459	1.463	LB 526
Ethene	C_2H_4	2	0.33	0.662	0.662	0.000	0.000	LB 225
Hexatriene	C_6H_8	6	1.80	0.907	1.913	1.414	1.700	LB 674
Isobutene	C_4H_8	4	0.90	0.457	0.535	0.952	1.346	LB 527
Propene	C_3H_6	3	1.44	0.255	0.359	1.974	1.974	LB 395
Methylformate	$C_2H_4O_2$	4	3.86	2.012	3.105	3.503	4.947	LB 245
Acetaldehyde	C_2H_4O	3	1.29	0.653	0.923	2.064	2.064	LB 243
Acetamide	C_2H_5NO	4	1.88	2.076	2.797	0.924	1.251	LB 270
Acetone	C_3H_6O	4	0.37	0.460	0.527	0.347	0.491	LB 403
2-Propenal	C_3H_4O	4	1.45	0.836	1.129	1.267	1.625	LB 375
Aniline	C_6H_7N	7	1.57	1.334	3.303	0.480	0.707	LB 673
Anisole	C_7H_8O	8	8.23	1.097	2.934	2.939	5.565	FEGFUS
Azomethane	$C_2H_6N_2$	4	2.70	3.181	3.555	3.761	3.764	LB 292
Benzoquinone	$C_6H_4O_2$	8	1.39	1.453	2.026	0.199	0.327	LB 653
Glycolicaldehyde	$C_2H_4O_2$	4	1.56	1.089	1.501	0.948	1.007	LB 247
N-Methylmethyleneimine	C_2H_5N	3	1.36	2.320	3.219	0.069	0.069	LB 268
N-Methylacetamide	C_3H_7NO	5	1.62	2.022	3.422	0.635	0.803	LB 424
N-Methylformamide	C_2H_5NO	4	3.15	1.481	1.525	2.143	2.916	LB 271
Nitroethene	$C_2H_3NO_2$	5	3.87	1.942	3.197	2.686	3.812	LB 222
Phenol	C_6H_6O	7	1.02	0.392	0.669	0.768	1.286	LB 671
Allene	C_3H_4	3	1.99	2.438	2.438	0.000	0.000	LB 368
Ethine	C_2H_2	2	0.02	0.038	0.038	0.000	0.000	LB 179
Propine	C_3H_4	3	0.49	0.863	1.173	0.000	0.000	LB 367
Phenylisocyanide	C_7H_5N	8	0.75	0.636	0.922	0.535	1.074	LB 722
Acrylonitrile	C_3H_3N	4	1.38	1.159	1.954	1.256	1.743	LB 363
Benzonitrile	C_7H_5N	8	0.96	0.848	1.682	1.087	2.209	LB 721
Cyclopentadiene	C_5H_6	5	0.47	0.556	1.022	0.089	0.116	LB 590
Dimethylenecyclobutene	C_6H_6	6	1.37	1.529	2.137	0.484	0.735	LB 665
Fluorene	$C_{13}H_{10}$	13	1.46	0.749	1.847	0.884	1.399	FLUREN02
Fulvene	C_6H_6	6	1.55	1.136	2.418	1.090	1.816	LB 666
Ethyleneepoxide	C_2H_4O	3	1.87	2.388	2.604	1.541	2.180	LB 244
Vinylenecarbonate	$C_3H_2O_3$	6	1.87	2.571	5.251	0.505	0.695	LB 353
Adenine	$C_5H_5N_5$	10	3.95	1.814	4.147	2.735	4.493	FUSVAQ01
Furan	C_4H_4O	5	1.50	1.178	2.156	1.054	1.588	LB 488
Guanine	$C_5H_5N_5O$	11	4.20	1.966	4.315	2.192	4.527	GUANMH10
Imidazole	$C_3H_4N_2$	5	2.75	2.725	4.022	2.123	3.562	IMAZOL06
Indole	C_8H_7N	9	1.76	1.588	3.063	1.043	2.016	EADIND
1,2,5-Oxadiazole	$C_2H_2N_2O$	5	2.24	2.881	4.219	0.755	1.364	LB 190
1,3,4-Oxadiazole	$C_2H_2N_2O$	5	1.58	1.808	3.860	0.647	1.269	LB 191
Pyrazole	$C_3H_4N_2$	5	1.32	1.277	1.909	1.242	1.835	LB 371
Pyridine	C_5H_5N	6	1.06	0.660	0.988	1.254	1.830	LB 583
Pyrrole	C_4H_5N	5	0.90	1.142	1.665	0.561	0.692	LB 498
Phosphabenzene	C_5H_5P	6	2.98	3.042	4.637	1.231	2.517	LB 586

TABLE 1
(continued)

Name	Molecular formula	Heavy atoms	rmsd (pm)	rmsb (pm)	maxb (pm)	rmsa (°)	maxa (°)	Reference code
1,2,5-Thiadiazole	$C_2H_2N_2S$	5	1.87	2.513	3.209	1.175	2.138	LB 192
1,3,4-Thiadiazole	$C_2H_2N_2S$	5	1.38	1.006	1.338	1.005	1.241	LB 193
Thiazole	C_3H_3NS	5	1.78	1.387	2.339	1.262	1.853	LB 366
Thiophene	C_4H_4S	5	1.34	0.762	1.415	1.008	1.602	LB 491
Bicyclo[1.1.1]pentane	C_5H_8	5	1.97	0.397	0.409	1.769	2.160	LB 609
Bicyclo[2.1.0]pentane	C_5H_8	5	1.72	1.607	2.397	0.496	1.025	LB 610
Bicyclobutane	C_4H_6	4	1.30	0.515	0.588	0.704	1.337	LB 504
Cyclobutane	C_4H_8	4	0.78	0.603	0.604	0.209	0.217	LB 528
Cycloheptadiene	C_7H_{10}	7	3.06	0.993	1.983	1.269	1.654	LB 738
Cycloheptatriene	C_7H_8	7	10.66	1.264	1.634	3.633	6.138	LB 727
1,3-Cyclohexadiene	C_6H_8	6	1.92	1.189	1.741	0.526	0.769	LB 675
Cyclohexene	C_6H_{10}	6	0.63	0.733	1.147	0.361	0.498	LB 689
Cyclooctatetraene	C_8H_8	8	3.23	0.500	0.775	1.145	1.255	LB 750
Cyclopropane	C_3H_6	3	0.56	0.975	1.000	0.012	0.016	LB 396
Cyclopentene	C_5H_8	5	2.12	1.432	1.836	0.802	1.092	LB 607
Spiropentane	C_5H_8	5	4.94	3.571	3.825	2.102	3.065	LB 608
6-Enlacton	$C_5H_6O_2$	7	4.56	1.802	2.864	1.505	2.799	–
Butyrolactone_d1	$C_9H_{14}O_2$	11	7.95	2.143	4.861	1.697	3.692	–
Butyrolactone_d2	$C_9H_{14}O_2$	11	3.56	1.479	2.946	1.352	2.425	–
Cyclobutylbromide	C_4H_7Br	5	1.87	2.063	4.259	0.461	0.556	LB 521
Cyclopropylchloride	C_3H_5Cl	4	1.69	2.447	4.591	0.543	0.855	LB 387
Glutarimide	$C_6H_9NO_2$	9	5.91	1.316	1.829	1.712	3.956	–
Oxabicyclo[2.2.1]heptan	$C_6H_{11}O$	7	2.55	1.936	3.621	2.080	2.723	LB 695
Thietane	C_3H_6S	4	1.70	1.212	1.505	1.507	2.426	LB 414
Mean		5.2	2.04	1.306	2.032	1.140	1.651	
Standard deviation		2.3	1.84	0.822	1.385	0.968	1.392	

Structures are characterized with a short name and the molecular formula. The third column gives the number of heavy atoms. Results include the rms deviation of atom positions after superimposing the minimized structure with the experimental one (column 4). For bonds and valence angles, rms deviations (columns 5 and 7, respectively) and maximum absolute deviations (columns 6 and 8, respectively) are given. Reference codes correspond to entries in Ref. 13 (LB...) or else in Ref. 14.

Parameters for hydrogen bonding (Appendix 8)

The position of the minimum r_0 of the radial function, Eq. 14, is taken to be a sum of element-specific radii:

$$r_{0da} = R_{hb,d} + R_{hb,a} \qquad (51)$$

Similarly, for the bond strength we use an average of element-specific values:

$$A_{Hda} = \frac{H_d + H_a}{2} \qquad (52)$$

The only distinction of atom types is for nitrogen and oxygen. To any other atom of the fifth or sixth group we assign default values.

The softness parameter w_r in Eq. 14 has a value W_{rl} when the distance r is lower than r_0, otherwise W_{rh} is taken.

For the angular dependence, Eq. 15, the reference value for the optimal direction α_0 as well as the widths β_l and β_h depend on the state of hybridization h, which is indicated by corresponding subscripts in the table. The width w_p of the planarity function, Eq. 16, has a single value.

Comparison with experimental structures

In this paper, we concentrate on structural aspects. We examine two sets of molecules for which the calculated geometries are compared with experimental data. The first contains some 80 structures of small molecules for which data are mostly derived from microwave-spectroscopic or electron diffraction data. They serve to adjust the parameters that govern the valence energy terms, and exemplify the extent to which the force field can cope with various topologies. The second set of molecules is a selection of small molecules for which reliable X-ray structures have been reported in the literature. It comprises 1589 structures of compounds containing only atoms from a limited list of elements (see below). These

TABLE 2
STATISTICS OF FORCE FIELD RESULTS

Quantity	Set 0	Set 1	Set 2
Number of atoms	16 ±7	16 ±6	18 ±6
Average bond length rmsd (pm)	2.1±1.4	2.1±1.2	2.0±0.9
Average valence angle rmsd (°)	2.1±1.1	2.1±1.0	2.3±1.2
Average match rmsd (pm)	13 ±15	17 ±14	24 ±18

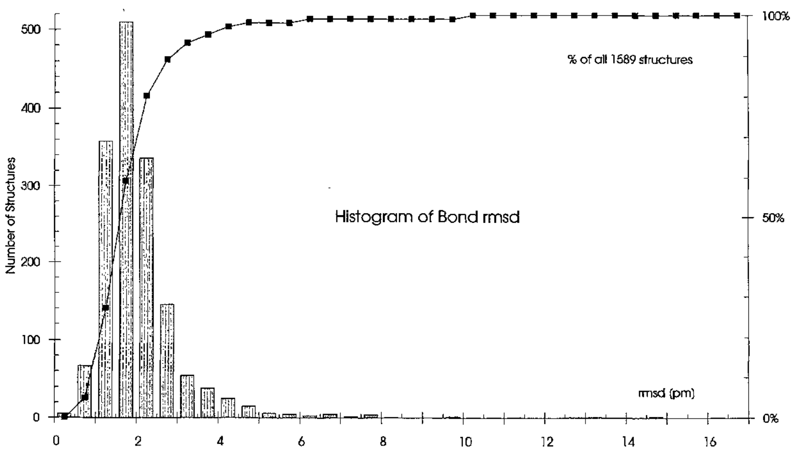

Fig. 4. Histogram of number of structures against rms deviation of bond lengths after minimization from the corresponding experimental values.

structures have been selected from the Cambridge Structural Database solely on the basis of the degree of structural refinement (see below). Consequently, the set encompasses the full spectrum of topological and structural diversity, including for example acyclic, cyclic, polycyclic, aromatic and non-aromatic conjugated, strained as well as non-strained compounds.

To avoid the compilation of an enormous reference list, we restrict ourselves to citing reference codes either of the structural tables of Landolt and Börnstein (as LB n, where n is the running number) [13] or of the CSD File [14].

The following test protocol was applied to each structure. In a first step all hydrogens were stripped, but their count was kept with each heavy atom to provide the required input for the united-atom approximation. Then every atomic coordinate was augmented by a random number in the range [–10,10] pm. Subsequently, the structure was minimized by the force field, using the conjugate-gradient procedure with restart (see Ref. 16). The result was compared with the original structure with respect to bond lengths, valence angles and rms deviation of atomic positions after rigid superimposition of all heavy atoms.

Structures for parameter determination

The parameters of the force field have been optimized with a set containing structures that more or less span the chemical range of interest in our applications. Typically, if a class of structures of particular interest appears to be badly reproduced by the force field, a representative of this class is entered into this set. The force field is then modified to again reproduce the structural features of the whole set satisfactorily. The results are listed in Table 1.

Validating structures

To provide an extended comparison with the vast set of experimentally determined structures as compiled in the CSD File [14], we have extracted all structures satisfying the following criteria.

(i) No error flag is set and the structure has no disorder.

(ii) The R-factor as given by the authors is below 4%.

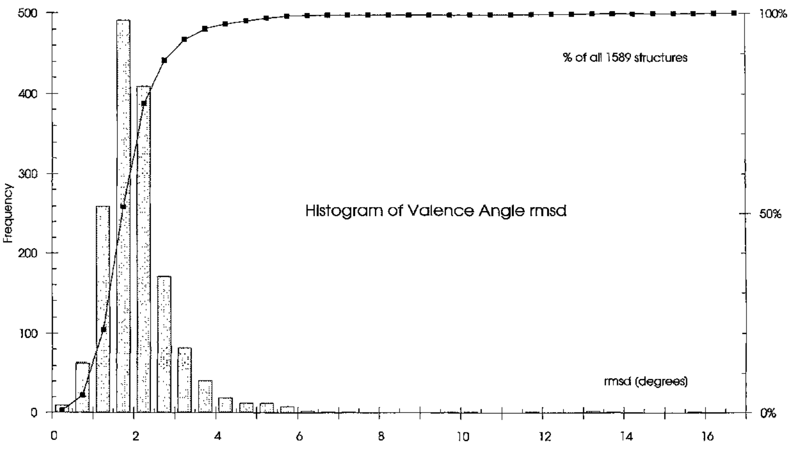

Fig. 5. Histogram of number of structures against rms deviation of valence angles after minimization from the corresponding experimental values.

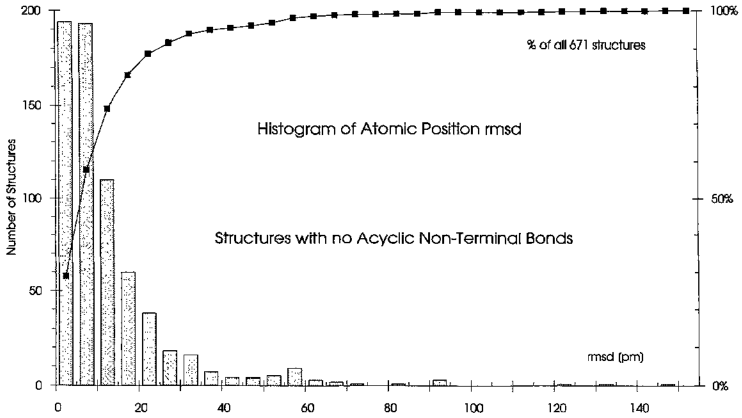

Fig. 6. Histogram of number of structures against the rms deviation of all atomic positions upon superimposition of the minimized structure onto the experimental one. Only structures that contain no rotatable bond have been included.

(iii) The authors give an rms deviation of less than 0.5 pm for C-C single bonds from standard values.

(iv) All atoms declared in the molecular formula are also given with coordinates. This also means that coordinates for all hydrogen atoms are given. This is important, because our force field needs the number of hydrogens attached to each atom in order to arrive at the correct bonding topology.

(v) The structure contains only the elements H, C, N, O, F, Si, P, S, Cl, Br or I.

(vi) Whenever duplicates occur in the database (with identical first six digits in the reference code), only the alphabetically first structure was taken.

In each structure from this set only the representative structural unit was considered. Counterions or solvent molecules were omitted. Whenever the representative molecule was found more than once in the asymmetric unit, only a single one was selected at random. No other criteria were applied to eliminate additional structures. These structures were sorted according to the number of acyclic, nonterminal bonds. Only the classes with zero, one and two such bonds were considered further. After removal of a few obviously erroneous structures, these classes contained 671, 437 and 481 structures, respectively.

The results are summarized in Table 2. For bond lengths and valence angles the agreement between experimental and calculated values is essentially independent of the class, as was to be expected. The histograms in Figs. 4 and 5 show that agreement in the ranges of 2 pm and 2° is achieved essentially over the whole range of structures.

The overall agreement, as expressed in the rms deviation of the atomic positions after rigid superimposition of the minimized model onto the experimental structure, shows on average an increase by about a factor of 1.4 with each additional rotatable bond. In addition, these numbers increase markedly with the number of atoms in a structure. The average increase is about 1 pm per atom in the structure. Figures 6, 7 and 8 show histograms of the atomic deviations for the three categories.

In this set of structures all major functional groups are represented. It also contains conjugated π-systems with or without heteroatoms, heterocycles of any kind, systems with extended heteroatom clusters, highly twisted π-systems, macrocycles, sterically hindered and polyfunctional

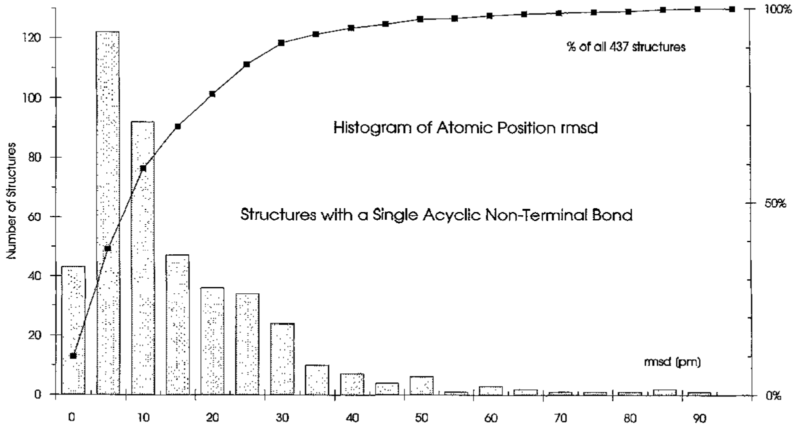

Fig. 7. Histogram of number of structures against the rms deviation of all atomic positions upon superimposition of the minimized structure onto the experimental one. Only structures that contain a single rotatable bond have been included.

TABLE 3
OCCURRENCE OF CHEMICAL CLASSES

Number	Chemical class	n	N	rmsd	bd	ang
1	Aliphatic acids/derivatives	34	2280	19	2	4
2	Aliphatic acid salts	5	621	12	2	4
3	Aliphatic amines	40	3182	16	3	5
4	Aliphatic (N and S)	6	255	12	3	6
5	Aliphatic miscellaneous	4	682	18	2	3
6	Enolates	1	201	29	2	5
7	Nitriles	11	1153	8	2	3
8	Urea compounds	21	913	10	3	6
9	N-N compounds	6	786	7	3	4
10	N-O compounds	14	494	10	3	6
11	S and Se compounds	25	2104	18	2	5
12	Carbonium ions etc.	3	624	9	4	10
13	Benzoic acid derivatives	13	1340	12	2	4
15	Benzene nitro compounds	16	1223	15	2	3
16	Anilines	19	1625	12	1	3
17	Phenols and ethers	30	2292	14	2	4
18	Benzoquinones	9	224	13	2	4
19	Benzene miscellaneous	7	1193	11	2	4
20	Monocycles (3,4,5)	21	1037	21	2	4
21	Monocycles (6)	29	1067	12	2	4
22	Monocycles (7,8)	4	194	11	1	2
23	Monocycles (9,9+)	6	139	11	1	4
24	Naphthalene compounds	14	711	27	2	4
25	Naphthoquinones	4	124	13	2	3
26	Anthracene compounds	9	397	5	1	3
27	Polycycles (2 fused)	34	992	18	2	3
28	Polycycles (3 fused)	31	892	12	2	4
29	Polycycles (4 fused)	25	483	22	2	4
30	Polycycles (5+ fused)	11	450	23	2	3
31	Bridged ring compounds	126	2698	14	2	5
32	Hetero-N (3,4,5 ring)	61	2282	20	2	5
33	Hetero-N (6 ring)	60	3037	16	2	4
34	Hetero-N (7,7+ ring)	13	273	17	2	3
35	Hetero-N (2 fused)	103	2990	18	2	4
36	Hetero-N (2+ fused)	111	2560	19	2	4
37	Hetero-N (bridged)	54	1323	16	2	4
38	Hetero-O	269	6917	21	2	4
39	Hetero-S,Se	114	2875	19	3	6
40	Hetero-(N and O)	77	2224	22	2	6
41	Hetero-(N and S)	102	2028	20	3	7
42	Hetero-mixed miscellaneous	28	513	24	3	7
43	Barbiturates	6	156	15	2	4
44	Pyrimidines, purines	89	2128	22	2	4
45	Carbohydrates	146	2915	21	2	4
46	Phosphates	7	603	18	3	5
47	Nucleosides/-tides	64	1151	28	2	4
48	Amino acids, peptides	30	2260	20	2	4
49	Porphyrins, corrins	1	1086	29	2	4
50	Antibiotics	14	814	25	2	4
51	Steroids	38	1504	16	1	3
52	Monoterpenes	1	127	10	2	5
53	Sesquiterpenes	30	935	15	2	4
54	Diterpenes	29	861	19	2	4
56	Triterpenes	8	345	20	1	4
58	Alkaloids	30	1452	14	2	4
59	Miscellaneous natural products	41	1275	21	2	4
60	Molecular complexes	15	1661	7	2	4
61	Clathrates	11	1503	25	2	5
64	Phosphorus compounds	60	4754	24	3	8

Listed are the number n of occurrences of chemical classes in the test data set, the total number of occurrences N in the CSD File (115 930 entries), average rms deviation of atomic positions, rmsd (pm), average bond deviation, bd (pm) and average angle deviation, ang (°), within each class.

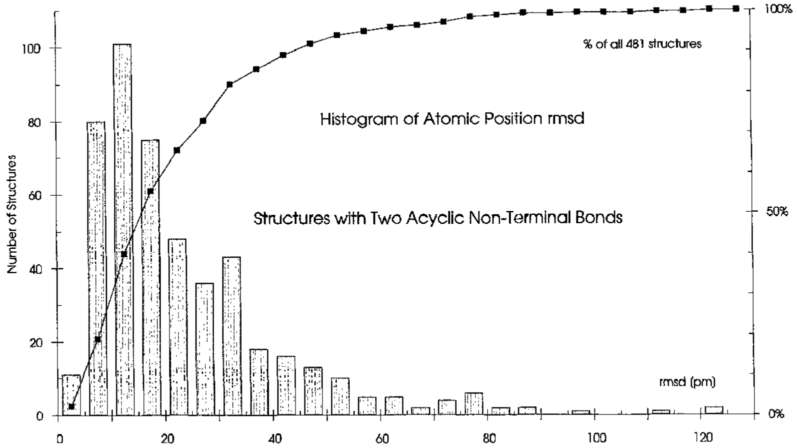

Fig. 8. Histogram of number of structures against the rms deviation of all atomic positions upon superimposition of the minimized structure onto the experimental one. Only structures that contain two rotatable bonds have been included.

compounds, including complex natural compounds. Of the 86 chemical classes defined in the CSD File, 62 are compatible with our specification of test compounds. Table 3 shows that most classes are represented, and well handled. However, it is also instructive to consider some of the less well handled cases, which constitute the tails of the histograms of Figs. 4–8. Here we find some rather unusual structures with extended heteroatom clusters, cases of extreme strain, a few cases of hypervalent second-row atoms, some highly distorted conjugated heterocycles, but also cases where the deposited structures are known to be inaccurate for reasons of intricate data analysis. Specific examples of poorly handled cases (for structures see Fig. 9) are thiapentalene derivatives (e.g. reference codes gipvoq, gipvuw02) with hypervalent sulfur atoms, conjugated bicyclic hydrazines in which nitrogen pyramidalities are underestimated (e.g. the nonplanar urazourazole, reference code gohvuu, but see also reference code biygul in which the hydrazine unit is indeed planar), or anthracenophane derivatives (e.g. reference code doxwui), where crystal packing effects may influence the conformation of the highly flexible polyether loops.

A complete list of all calculated structures with reference codes and deviations in geometrical parameters is available from the authors on request.

Further tests

It is clear that the testing described above concerns mostly the bonded interaction potentials. However, in the area of ligand design a reliable description of the nonbonded interactions is of crucial importance. An equally thorough testing of these terms is infinitely more complex. It would mean calculation of crystal packing of small molecules, a task involving thermodynamic calculations, which is far beyond the scope of the present documentation. Nevertheless, some more qualitative testing can be achieved by considering folded proteins. Many secondary structure features are stabilized by hydrogen bond patterns. Thus, if experimental structures stay essentially unchanged under energy minimization this is an indication, though by far not a proof, that essential features of the nonbonded interactions are properly modelled.

Energy minimization of protein structures within the present force field lead to structures that deviate from the experimental ones by a few tens of a pm, even with some intermediate randomization of the atomic coordinates with amplitudes of several tens of a pm. As an example, after minimization of the structure of lysozyme (entry 2LZT of the Brookhaven Protein Database (PDB) [17]), a 23 pm rms deviation for all atomic positions was obtained. We are aware that such calculations have only a limited value as preliminary tests, but they indicate, nevertheless, that the systems of interest are not grossly distorted, but stay within the experimental resolution upon minimization.

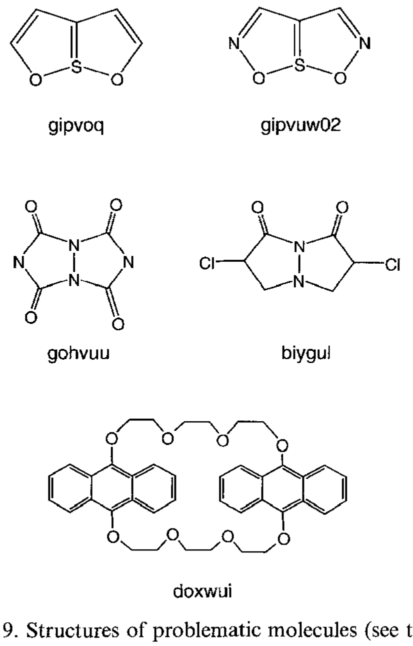

Fig. 9. Structures of problematic molecules (see text).

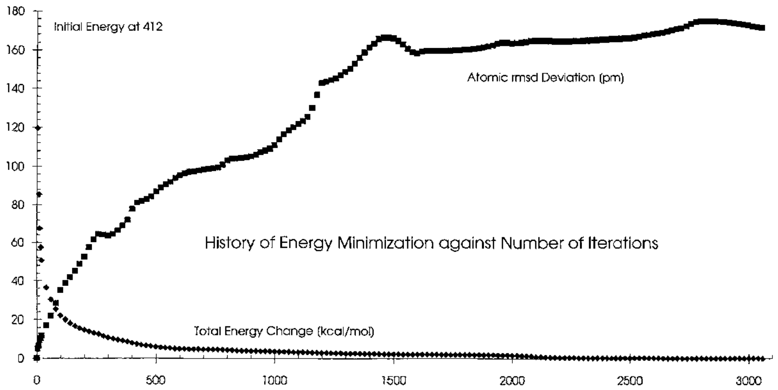

Fig. 10. Energy change and overall rms displacement of atoms from experimental positions during energy minimization of a dodecamer double-stranded DNA [18]. On the abscissa the number of iteration steps is given. Evident is a fast conservative structure relaxation to a state with energy near the final value, followed by a slow phase with little energy gain and large-scale structural modifications.

A further example is the unconstrained relaxation of Dickerson's DNA double helix (entry 9BNA of the Brookhaven PDB) [18]. In this case it is interesting to follow the minimization trajectory. Figure 10 shows the development of the energy change and of the rms deviation of all heavy atom positions against the number of iterations. It is evident that after only 100 iterations, the energy has approached its final value to within about 20 kcal/mol, which is less than a single kcal/mol per monomer, while the positions of the atoms have changed by about 35 pm per atom. During the remaining 3000 iteration steps the energy hardly changes, while a slow conformational adjustment takes place, which mainly affects large-scale features. This low-energy rearrangement leads to an average atomic displacement in the range of 200 pm. It is characterized by fully preserved base pairing, as evident from the H-bond pattern. The most pronounced changes occur in global features, such as a reduction in helix pitch by 100 to 200 pm over the whole dodecamer. Furthermore, a widening of the minor groove by about 100 pm and a corresponding narrowing of the major groove develop. These features are reflected in the rmsd values of partial rigid superimpositions. Single bases superimpose within about 30 pm, base pairs within 80 pm, while for the whole structure we obtain about 170 pm. The energy record illustrates that these conformational changes involve very little strain.

Of importance for our applications is that conformations of inhibitors embedded in proteins, as measured for example by X-ray diffraction, are properly reproduced by the force field. A prominent example is the complex of methotrexate with the dihydrofolate reductase of *Escherichia coli*. Starting from the experimental structure [19] (which is of somewhat poor geometry with respect to small-scale features), the complex develops under energy minimization into a conformation that shows the described H-bonding pattern excellently. Although such statements are of little use for quality assessment of the force field, they are valuable for interactive modelling, where the development of a particular H-bond pattern is often a useful, if not fully quantitative, result.

Finally, to illustrate the computational efficiency a few data for lysozyme (2LZT) are given. The setup phase takes 15 s on a Silicon Graphics INDY R4000SC, while each iteration takes 3 s.

Conclusions

We have presented a detailed description of our MAB force field with respect to the mathematical form of the energy terms, as well as the structure of the parametrization scheme. Extensive structural test calculations are given, demonstrating that a wide spectrum of topological and structural cases can be treated in an accuracy range of 2 pm on bond lengths and 2° on valence angles. Root mean square deviations of atomic positions depend on the number of atoms in the structure as well as on the number of acyclic nonterminal bonds. They increase roughly by 1 pm per atom and by a factor of 1.4 per acyclic nonterminal bond. A particular feature is the automatic search for optimal hydrogen bond patterns. Energetic aspects have only been touched upon in the present work. However, the quality of structural aspects in the rich spectrum of situations covered is an indication that the energies of the valence terms are well balanced. Otherwise, strained structures which react quite sensitively to bad tuning of these terms would not be treated satisfactorily. Nevertheless, it is planned to add more detailed investigations on conformational energy aspects as well as on nonbonded interaction assessment in future publications.

Testing of larger scale structural features revealed, on a more qualitative level, satisfactory behaviour on folded proteins (23 pm agreement) or DNA oligomers (20 to 200 pm agreement, depending on length scale), as well as on inhibitor–receptor complexes.

Acknowledgements

We have benefitted from excellent retrieval software for the structures of the Cambridge Structural Database by H.J. Ammann. The many comments of colleagues, who work on a daily basis with the force field, have helped to identify weak points and awkward performances. In this context we thank in particular K. Gubernator, C. Broger, D. Bur, T. Schaumann, N. Borkakoti and A. Kroehn.

References

1 Li, J.H. and Allinger, N.L., J. Comput. Chem., 12 (1991) 186.

2 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.

3 Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A., J. Phys. Chem., 79 (1975) 2361.

4 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.

5 Van Gunsteren, W.F. and Berendsen, H.J.C., GROMOS, Biomos, Groningen, 1987.

6 Maple, J.R., Dinur, U. and Hagler, A.T., Proc. Natl. Acad. Sci. USA, 85 (1988) 5350.

7 Clark, M., Cramer III, R.D. and Van Opdenbosch, N., J. Comput. Chem., 10 (1989) 982.

8 Momany, F.A. and Rone, R., J. Comput. Chem., 13 (1992) 888.

9 Taylor, R. and Kennard, O., J. Am. Chem. Soc., 105 (1983) 5761.

10 Vedani, A. and Dunitz, J.D., J. Am. Chem. Soc., 107 (1985) 7653.

11 Heilbronner, E. and Bock, H., The HMO Model and its Application, Wiley, London, 1976.

12 Trinajstic, N., Chemical Graph Theory, CRC Press, Boca Raton, FL, 1983.

13 Callomon, J.H., Hirota, E., Kuchitsu, K., Lafferty, W.J., Maki, A.G. and Pote, C.S., In Hellwege, K.H. (Ed.) Landolt-Börnstein, Vol. 7: Structure Data of Free Polyatomic Molecules, Springer, Berlin, 1976.

14 Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rogers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.

15 Allred, A.L. and Rochow, E.G., J. Inorg. Nucl. Chem., 5 (1958) 264.

16 Powell, M.J.D., Math. Programming, 12 (1977) 241.

17 Ramanadham, M., Sieker, L.C. and Jensen, L.H., Acta Crystallogr., A37 (1981) 33.

18 Fratini, A.V., Kopka, M.L., Drew, H.R. and Dickerson, R.E., J. Biol. Chem., 257 (1982) 14686.

19 Matthews, D.A., Bolin, J.T., Burridge, J.M., Filman, D.J., Volz, K.W., Kaufman, B.T., Bedell, C.R., Champness, J.N., Stammers, D.K. and Kraut, J., J. Biol. Chem., 260 (1985) 381.

Appendixes

1. Atomic parameters

Atom	r_{b0}: Bond radius (Eq. 26)	R_0: vdW radius (Eq. 49)	C: vdW range factor (Eq. 48)	E_m: vdW minimum energy (Eq. 50)	Electro-negativity χ	Hückel α_0
Def.	$0.426 \times R_0$		1.67	0.12		0
LP					1.6	
H	0.345	1.20	–	–	2.70	–
He	1.4	1.40	–	–	1.00	–
Li	–	0.68	–	–	0.97	–
Be	–	0.40	–	–	1.47	–
B	–	0.35	–	–	2.01	–
C	0.766	1.80	1.67	0.16	2.50	–
N	0.71	1.64	1.4	0.03	3.07	0.3
O	0.66	1.50	1.4	0.02	3.50	0.6
F	0.628	1.47	–	–	4.10	–
Ne	1.54	1.54	–	–	1.00	–
Na	–	0.97	–	–	1.01	–
Mg	–	0.66	–	–	1.23	–
Al	–	0.51	–	–	1.47	–
Si	1.1	2.10	–	–	1.74	–
P	1.09	2.00	1.85	0.2	2.06	0.2
S	1.05	1.95	1.85	0.3	2.44	0.3
Cl	1.02	1.75	1.85	0.2	2.83	–
Ar	–	1.88	–	–	1.00	–
Br	1.17	1.85	2.0	0.3	2.74	–
I	1.37	2.20	2.0	0.3	2.21	–

Default values, which are indicated by a (–), are given in the first row. For the bond radius they are taken to be 0.426 times the van der Waals radius R_0. A default value is mostly an indication that the corresponding situation has rarely ever occurred. Electronegativities are taken from Ref. 15. Distances are in Å, energies in kcal/mol.

2. Parameters for bond-order determination

Parameter	Defined in Eq.	Value
α_e	24	0.5
β_0	25	0.45
R_{sr}	25	0.9
B_{CO}	26	0.7

3. Parameters for bond stretching

Parameter	Defined in Eq.	Value	Unit
R_v	27	0.04	
B_{LP}	28	0.04	Å
B_h	29	0.02	
B_d	30	0.018	
B_1	31	0.1531	
B_2	31	0.0228	
R_a	32	0.054	
A_{b0}	33	300	kcal/(mol Å^2)
A_{b1}	33	1.17	

4. Parameters for angle bending

Parameter	Defined in Eq.	Value	Unit
Δv_{th}	4	4	degree
V_{0,sp^1}	~37	180	degree
A_{sp^1}	~38	0.02	kcal/deg^2
V_{0,sp^2}	36	120	degree
A_{sp^2}	38	0.01	kcal/deg^2
A_{l2}	~38	1.5	kcal/deg^2
V_{0,sp^3}	37	110.5	degree
A_{sp^3}	~38	0.02	kcal/deg^2
V_χ	36	–5	degree
V_q	36	–12	degree
V_p	36	5	degree
V_{LP}	~36	5	degree
V_{r2}	~37	0.88	degree
A_{rw}	~38	0.2	
A_{rw4}	~38	0.4	
p_{th}	37, 40	0.53	

5. Parameters for pyramidality terms

Parameter	Defined in Eq.	Value	Unit
A_{sp^2}	39, 40	30	kcal/mol
A_4	40	25	kcal/deg^2
p_{sp^3}		0.77	
A_{sp^3}		50	kcal/deg^2

6. Parameters for dihedral angle terms

Parameter	Defined in Eq.	Value	Unit
T_3	41	3.0	kcal/mol
T_{ss}	~41	2.5	kcal/mol
T_{23}	42, 43	1.0	kcal/mol
T_2	44	30.0	kcal/mol
T_{2q}	44	0.5	
T_{2m}	45	0.4	
T_{am}	46	1.5	kcal/mol

7. Parameters for dispersive interactions

Parameter	Defined in Eq.	Value	Unit
W_h	49	0.05	Å
W_{LP}	49	0.05	Å
R_{hb}	~49	0.9	
R_{14}	~49	1.05	
A_{14}	~50	200	kcal/mol

8. Parameters for hydrogen bonding

Parameter	Defined in Eq.	Value	Unit
$R_{hb,N}$	51	1.45	Å
$R_{hb,O}$	51	1.4	Å
$R_{hb,X}$	51	1.5	Å
$R_{hb,M}$	51	0.2	Å
H_N	52	4.0	kcal/mol
H_O	52	4.5	kcal/mol
H_X	52	2.6	kcal/mol
H_M	52	14.0	kcal/mol
W_{rl}	14, ~52	0.7	Å
W_{rh}	14, ~52	1.0	Å
α_{0,sp^1}	15, ~52	180	degree
β_{sp^1}	15, ~52	90	degree
α_{0,sp^2}	15, ~52	120	degree
β_{l,sp^2}	15, ~52	60	degree
β_{h,sp^2}	15, ~52	90	degree
α_{0,sp^3}	15, ~52	109.47	degree
β_{l,sp^3}	15, ~52	60	degree
β_{h,sp^3}	15, ~52	90	degree
w_p	16	0.75	

A ~ sign in the second column in Appendixes 4, 6, 7 and 8 indicates that the corresponding value is defined near the corresponding equation, without being explicitly written in any equation.