

# Molecular Force Field Parametrization using Multi-Objective Evolutionary Algorithms

S. Mostaghim\*

Electrical Engineering Department,  
University of Paderborn,  
Germany

mostaghim@date.upb.de

M. Hoffmann\*, P. H. König\*, Th. Frauenheim

Physics Department,  
University of Paderborn,  
Germany

{hoffmann,koenig,frauenheim}@phys.upb.de

J. Teich

Computer Science Department,  
Friedrich-Alexander-University,  
Erlangen,  
Germany

teich@cs12.de

**Abstract**—We suggest a novel tool for the parametrization of molecular force fields by using multi-objective optimization algorithms with a new set of physically motivated objective functions. The new approach is validated in the parametrization of the bonded terms for the homologous series of primary alcohols. Multi-Objective Evolutionary Algorithms (MOEAs) and particularly Multi-Objective Particle Swarm Optimization (MOPSO) are applied. The results show that in this case MOPSO finds solutions with higher convergence than the MOEA method. Physical analysis of the results confirms the performance of the MOPSO method and the choice of objective functions.

## I. INTRODUCTION

Molecular force fields are used to explore, explain and predict a large variety of molecular properties like vibrational spectra, energy changes and binding affinities e.g., for pharmaceuticals and to simulate chemical reactivity as in transition states, minimum energy pathways and free energy differences.

The parametrization of molecular force fields is a tedious task involving a large number of parameters as well as a number of objectives. The conventional methods in solving such problems are iterative techniques and the weighting methods. Iterative methods optimize each of the objectives one by one until a self-consistent state has been reached [1].

Weighting methods reduce the dimensionality of the problem by suitable weighting parameters [2]. However, the classical weighting method finds one solution for the chosen weighting parameters after each simulation run. The choice of weights for different objectives is challenging if different properties of different physical nature are combined. Also different weights have to be used if the quality of one property should be increased at the cost of another one [2]. Therefore *a priori* knowledge is necessary to find suitable weights.

An alternative approach is the use of multi-objective optimization algorithms in which all of the objectives are optimized at once. The result is not a single parameter set (or a small number thereof) which satisfies the dynamical process outlined above, but rather a variety of parameter sets with trade-offs in terms of the objective functions. Within this variety one objective can not be improved without loss of performance for other objectives. The consideration of accuracy of the description of the individual objectives can

hence be postponed until after the optimization. A suitable solution can then be chosen according to demands of accuracy and physical arguments.

One important aim for the force field parametrization is to get the best description of the reference data. For finding solutions close to the global best solution for this optimization problem genetic algorithms are well suited.

Different approaches for force field parametrization have been made using genetic algorithms. Busold and Strassner [3], [4] describe an approach for parametrizing metalloorganic compounds in the framework of the MM3 (Molecular Mechanics version 3) force field [5] using genetic algorithms. They only use Cartesian, geometric root mean square deviations (*rmsd*) for fitting all force field parameters (including force constants). Hence they neglect other essential properties such as the curvature of the potential energy surface and structure and energetics of different conformations, which are actually needed for physically determining all of the parameters involved. In similar (earlier) publications of Huttner et al. natural parameters for molybdenum compounds in the framework of the MM2 (Molecular Mechanics version 2) force field [6] were determined using the *rmsd* values to a large number of reference structures as the objective function [7], [8]. The authors optimize the parameters which can directly be deduced from the objective function.

Wang and Kollmann [2] describe the parametrization using genetic algorithms and a combination of different objectives with weighting functions. This approach however has the intrinsic problem outlined above.

So far there is no research on the application of Multi-Objective Evolutionary Algorithm (MOEA) [9], [10] for force field parametrization. Therefore in this paper we study the application of MOEA on this problem. MOEA methods were tested on different other problems [9] and are recorded to be able to find solutions with high diversity and convergence particularly for problems with a high number of parameters and objectives. On the other hand, Multi-Objective Particle Swarm Optimization (MOPSO) [11] methods are also recorded to solve such problems successfully. Therefore, the MOPSO method is also applied as another alternative optimization method on the parametrization of the force fields and compared with the result of the MOEA.

\* These authors contributed equally to this work

This paper is organized as follows: A brief introduction on molecular force fields is given in Section II. In Section III, methods, the objective functions and parameters are introduced. The validation of the methods both from a technical as well as a physical point of view is shown in Section IV.

## II. MOLECULAR FORCE FIELDS

To describe deformations of a molecular structure a suitable functional form has to be chosen which maps coordinate fluctuations onto an energy function. Various functional forms and parametrizations have been introduced as reviewed in [12]. The work presented here is focused on the parametrization of a class II force field, namely in the framework of the force field implemented in CHARMM (Chemistry at HARvard Molecular Mechanics) [13]. The molecular force field is of the functional form denoted in Equations (1)-(3). In these equations interactions in molecules are classified into interactions mediated by bonds, termed intramolecular (2) and interactions between atoms separated by three or more bonds, termed intermolecular (3). The total energy term describing a molecule is the sum of inter- and intramolecular energy contributions:

$$E = E_{intra} + E_{inter} \quad (1)$$

A natural choice of coordinates for the intramolecular interactions is one which is directly derived from the topology of the molecule using bond lengths, angles and dihedral angles (see Figure 1). The intramolecular potential describes the variation of energy connected to deviations from given geometrical parameters (natural parameters)<sup>1</sup>.

$$E_{intra} = \sum_{bonds} \frac{1}{2} k_b (b - b_0)^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{torsions} \sum_n k_{n\phi} (1 + \cos(n\phi - \delta_n)) \quad (2)$$

$$E_{inter} = \sum_{\substack{nonbonded, \\ i < j}} \frac{q_i q_j}{r_{ij}} + \sum_{\substack{nonbonded, \\ i < j}} \epsilon_{ij} \left[ \left( \frac{R_{min,ij}}{r_{ij}} \right)^6 + \left( \frac{R_{min,ij}}{r_{ij}} \right)^{12} \right] \quad (3)$$

The first term in (2) describes the necessary energy for lengthening and shortening bond lengths  $b$  relative to a natural bond length  $b_0$ . The potential is harmonic with a force constant  $k_b$ . In the same fashion deviations of bond angles  $\theta$  from a natural bond angle  $\theta_0$  are described. The description of the dihedral coordinates  $\phi$  differs as the function is required to satisfy periodicity.

Force field parametrization includes the determination of suitable values for the force constants of bonds  $k_b$ , angles  $k_\theta$  and dihedral angles  $k_{n\phi}$  as well as their associated natural values  $b_0$ ,  $\theta_0$  and the phase for the dihedral angles  $\delta_n$ . For the

<sup>1</sup>For a discussion of the term natural parameters vs. equilibrium parameters see the review by Jensen [14]

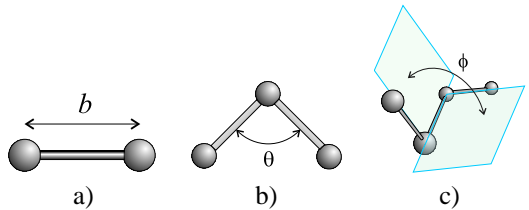


Fig. 1. Relevant intramolecular geometry measures used in molecular force fields: a) bond lengths b) bond angles c) dihedral angles.

description of nonbonded interactions the partial charges  $q$  for the electrostatic energy and the Lennard-Jones Parameters  $\epsilon_{ij}$  and  $R_{min,ij}$  for the description of van der Waals interactions have to be determined.

To obtain a reasonable description all of these parameters have to be adapted for different *atom types* i.e., different chemical elements and bonding situations. Hence the force field parameters are specific for each term summed over. For instance a carbon-carbon bond in a benzene ring and a carbon-carbon bond in an alcohol represent different bonding situations and different values for both  $k_b$  and  $b_0$  are required. Simultaneously force field parameters should be as transferable as possible meaning that the parameters should be applicable to a different molecule showing a similar bonding situation. In many cases parameters can be adapted directly to other molecules (parametrization by analogy) while in some cases the description can be improved significantly by reparametrization [1].

## III. MULTI-OBJECTIVE OPTIMIZATION

In this section, definitions in multi-objective optimization are briefly reviewed. Then the objectives of this study are outlined.

### A. Definitions

A multi-objective optimization problem has several objective functions which are to be minimized or maximized at the same time. In the following, we state the multi-objective optimization problem in its general form:

$$\text{minimize } \vec{y} = \vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \dots, f_m(\vec{x}))$$

The *decision vectors* (parameters)  $\vec{x} = (x_1, x_2, \dots, x_n)^T$  belong to the feasible region  $S \subset \mathbb{R}^n$ . We denote the image of the feasible region by  $Z \subset \mathbb{R}^m$  and call it a feasible objective region. The elements of  $Z$  are called *objective vectors* and consist of objective values  $\vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \dots, f_m(\vec{x}))$ .

- A decision vector  $\vec{x}_1 \in S$  is said to *dominate* a decision vector  $\vec{x}_2 \in S$  (denoted  $\vec{x}_1 < \vec{x}_2$ ), iff the decision vector  $\vec{x}_1$  is not worse than  $\vec{x}_2$  in all objectives and strictly better than  $\vec{x}_2$  in at least one objective.
- A decision vector  $\vec{x}_1 \in S$  is said to *weakly dominate* a decision vector  $\vec{x}_2 \in S$  (denoted  $\vec{x}_1 \preceq \vec{x}_2$ ), iff the decision vector  $\vec{x}_1$  is not worse than  $\vec{x}_2$  in all objectives.
- A decision vector  $\vec{x}_1 \in S$  is called *Pareto-optimal* if there does not exist another  $\vec{x}_2 \in S$  that dominates it. An objective vector is called Pareto-optimal if the corresponding decision vector is Pareto-optimal.

The non-dominated set of the entire feasible search space  $S$  is the Pareto-optimal set. The Pareto-optimal set in the objective space is called *Pareto-optimal front*.

### B. Quantitative convergence metric

For quantitatively comparing two non-dominated sets in terms of the convergence there are several different quantitative metrics, like the diversity metric and the number of non-dominated solutions [15]. Another measure, the  $C$  metric is introduced in [16] and is aimed to compare the convergence of two non-dominated sets of  $A$  and  $B$  as follows:

$$C(A, B) = \frac{|\{b \in B | \exists a \in A : a \preceq b\}|}{|B|} \quad (4)$$

The value of  $C(A, B) = 1$  means that all the members of  $B$  are weakly dominated by the members of  $A$ . We can also conclude that  $C(A, B) = 0$  means non of the members of  $B$  are weakly dominated by the members of  $A$ . However,  $C(A, B)$  is not equal to  $1 - C(B, A)$  and we have to consider both of the  $C(A, B)$  and  $C(B, A)$  for comparison.

### C. Description of objective functions

For the parametrization of force field terms both experimental and *ab initio*<sup>2</sup> data can be employed. Here we will focus on the reproduction of *ab initio* molecular geometries, molecular vibrations and rotational barriers. In this study we consider three objective functions as follows.

1) *Reproduction of molecular geometries*: Some parametrization studies use the Cartesian *rmsd* of the atoms with respect to a reference structure. However, this can lead to an unphysical weighting of different deviations of optimum force field parameters which in turn can result in an unbalanced optimization. An alternative approach employed by Dasgupta et al. [17] is to take the forces exerted on the atoms of the reference structure using the test parameters. If the potential generated by the force field has a minimum at the same place as the reference structure the forces i.e. the gradients should be zero. We follow these arguments and suggest to take the maximum component of the energy gradient in Cartesian coordinates as the objective function for the reproduction of molecular geometries (5)<sup>3</sup>.

$$f_1 := \max_{1 \leq i \leq N} |\nabla_i E| \quad (5)$$

Here,  $N$  specifies the number of atoms in the molecule and  $\nabla_i$  is the gradient vector with respect to the position of the  $i$ -th atom.

2) *Reproduction of molecular vibrations*: The interpretation of vibrational spectra suffers from the identification problem. It describes the task of assigning each calculated vibration to the corresponding reference vibration, i.e. an experimentally observed one or one calculated using *ab initio* methods. Due to this problem a direct fitting which only compares the

<sup>2</sup>*ab initio* refers to quantum mechanical calculations where no a priori information is required.

<sup>3</sup> The units for objective functions  $f_1$  and  $f_3$  are  $\text{kcal} (\text{mol } \text{\AA})^{-1}$  and  $(\text{kcal mol}^{-1})^2$ .

vibrations directly sorted by frequency can lead to unphysical results and may prevent correct optimization.

To circumvent this problem a better choice is to compare the force matrices in internal coordinates. Molecular vibrations in harmonic approximation, i.e. vibrational frequencies and normal coordinates are obtained using the second derivatives of the energy. The elements  $F_{ij}^{(c)}$  of the Cartesian force matrix  $\mathbf{F}^{(c)}$  in direction of the Cartesian displacement coordinates  $q_i^{(c)}$  are calculated according to (6).

$$F_{ij}^{(c)} = \frac{\partial^2 E}{\partial q_i^{(c)} \partial q_j^{(c)}} \quad 1 \leq i, j \leq 3N \quad (6)$$

In internal coordinates the deformations of a molecule are not described using displacements in terms of an external Cartesian coordinate system but regarding changes of internal coordinates of the molecule, e.g. bond lengths, bond angles and dihedral angles. The definition of the internal coordinates used in this work is based on Pulay et al. [18]. These internal coordinates  $\mathbf{q}^{(i)}$  can be described as a function of the displacement in cartesian coordinates  $\mathbf{q}^{(c)}$ :

$$\mathbf{q}^{(i)} = \mathbf{q}^{(i)}(\mathbf{q}^{(c)}) \quad (7)$$

This function is not linear in general. But under the assumption of small displacements the transformation from cartesian to internal coordinates can be linearized.

$$\mathbf{q}^{(i)} = \mathbf{B} \mathbf{q}^{(c)} \quad (8)$$

The construction of the matrix  $\mathbf{B}$  was first described by Wilson et al. [19]. In this work the  $(3N, 3N - 6)$ -dimensional  $\mathbf{B}$  matrix respective its pseudoinverse  $\mathbf{B}^\dagger$  is used to transform the force matrix into internal coordinates [20].

$$\mathbf{F}^{(i)} = \mathbf{B}^{T\dagger} \mathbf{F}^{(c)} \mathbf{B}^\dagger \quad (9)$$

One of the advantages of this approach is that the problem is decoupled from the external coordinate system. So a more natural description of the system can be achieved at the costs of choosing a proper set of non-redundant internal coordinates and transforming the calculated matrices into internal coordinates. In order to minimize the identification problem the relative deviation (10) of the diagonal elements of the force constant matrix in internal coordinates with respect to a reference matrix is used as objective. Comparing only the diagonal elements is possible because after the transformation the off-diagonal elements are smaller by several orders of magnitude than the diagonal elements.

$$f_2 := \max_{0 \leq j \leq 3N-6} \left( \frac{F_{jj}^{(i)} - F_{jj}^{(i),ref}}{F_{jj}^{(i),ref}} \right)^2 \quad (10)$$

The choice of absolute deviations leads to a good description of vibrations associated to diagonal elements which are large (i.e., bond stretching) but a weak description of vibrations where the associated diagonal elements are small. Hence we choose relative deviations.

3) *Reproduction of energetics for different conformations:* Of course a force field should be able to reproduce the relative energies for different conformations as well as barriers between them. Therefore the third objective function is defined by the sum over quadratic energy deviations for all relevant conformations and barriers between them<sup>3</sup>:

$$f_3 := \sum_i (E_i - E_i^{ref})^2 \quad (11)$$

#### D. Choice of parameters

As components of the decision vector we use the force field parameters  $k_b$ ,  $b_0$ ,  $k_\theta$ ,  $\theta_0$  and  $k_{n\phi}$ . The number of components of the decision vector  $\vec{x}$  is defined by the particular parametrization problem. As explained above for each specific chemical bonding situation different parameters are required. In the first step the search space  $S$  is chosen such that all of the bonding situations described so far in the CHARMM27 force field [21] are covered. The final intervals for  $S$  are then increased in order not to exclude alternative solutions from the beginning.

### IV. VALIDATION

#### A. Computational Details

Reference molecular geometries, energetics for different conformations and vibrational force matrices were calculated using a 6-311G\*\* basis set and the B3LYP hybrid density functional as implemented in the *ab initio* program package Gaussian98 [22]. The cartesian force fields were transformed into internal force fields using the UNRAVEL program suite [20]. The force field calculations were performed using CHARMM [13].

The optimizations were carried out using the MOPSO and MOEA methods. The Sigma method [11] is selected as the MOPSO and SPEA2 [10] as the MOEA method. The MOEA and MOPSO methods are run with the parameters: population size: 300, archive size: 200  
MOEA specific: mutation probability: 0.01, 0.1, cross-over probability: 0.8  
MOPSO specific: inertia weight: 0.4, turbulence factor: 0.01

#### B. Exemplaric study

For validation the force field parameters were determined for primary alcohols.

While the force field parameters for these alcohols are available in common force fields we use them for *validation* of the methods described here. They are of simple structure and allow the testing of the new suggested procedure. In each run of parametrization multiple alcohols were parametrized simultaneously. This guarantees a certain degree of transferability ensuring that deviations (of physical origin) in small molecules do not determine the parameters for all of the molecules of the series. Also in a previous study it was pointed out that the parametrization problem can be under-determined in some cases [17]. By fitting to multiple alcohols simultaneously we can ensure that the problem is well defined.

Aliphatic alcohols whose parameters were fitted in this study have the general structure depicted in Figure 2. In the series  $k$  is a nonnegative integer describing the length of the alcohol.

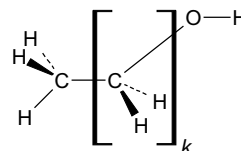


Fig. 2. General structure of non-branched, primary, aliphatic alcohols. The index  $k$  indicates how often the unit contained within the brackets is repeated.

#### C. Case Study: Set of two alcohols

Here we want to find force field parameters for a set of two alcohols: methanol ( $k=0$ ) and ethanol ( $k=1$ ). The number of parameters and objectives are 36 and 3 respectively. The set of parameters is composed of 6 parameters for each  $k_b$  and  $b_0$ , 9 parameters for each  $k_\theta$  and  $\theta_0$  and 6 parameters for  $k_{n\phi}$ . As objectives we use the functions introduced in (5), (10) and (11) where the values for the two molecules are summed up, e.g.  $f_1^\Sigma = f_1^{methanol} + f_1^{ethanol}$ . For the computation of the objective function  $f_2$  the number of internal coordinates i.e., the dimension of the force matrices  $F^{(i)}$  according to (9) are 12 and 21 (for methanol and ethanol respectively).

The evaluation of the sums in the contributions to  $f_3$  in (11) is carried out using a number of control points which describe the rotational profile sufficiently. In the case of ethanol 4 resp. 3 points are included for the rotations about the H-C[-]C-O and the C-C[-]O-H bonds. The rotation about the H-C[-]O-H bond in methanol can be described using 2 points due to symmetry.

1) *Comparison of different algorithms and algorithm related parameters:* First we study the performance of the MOEA method for this problem for different probabilities of mutation ( $p_m$ ): 0.01 and 0.1. The optimization process was terminated after 3000 generations. For each of the different  $p_m$  parameters we ran 10 individual optimizations with different random seeds. Out of the results we picked out the run with the best archive, where the criteria was the  $C$  metric. Figures 3 (a) and (b) show these results. The quantitative  $C$  metric values are  $C(MOEA_{0.1}, MOEA_{0.01}) = 8\%$ ,  $C(MOEA_{0.01}, MOEA_{0.1}) = 73\%$ . This means that with lower value of  $p_m$  better convergence can be achieved.

For additional comparisons we also compare the results of the 10 runs of MOPSO method after 3000 generations. Again we picked out the best run the same as the MOEA results. Figure 3 (c) shows the non-dominated set in the objective space. The values of the quantitative measures ( $C$  metric) are as follows:

$$\begin{aligned} C(MOPSO, MOEA_{0.1}) &= 97\% \\ C(MOEA_{0.1}, MOPSO) &= 0\% \end{aligned}$$

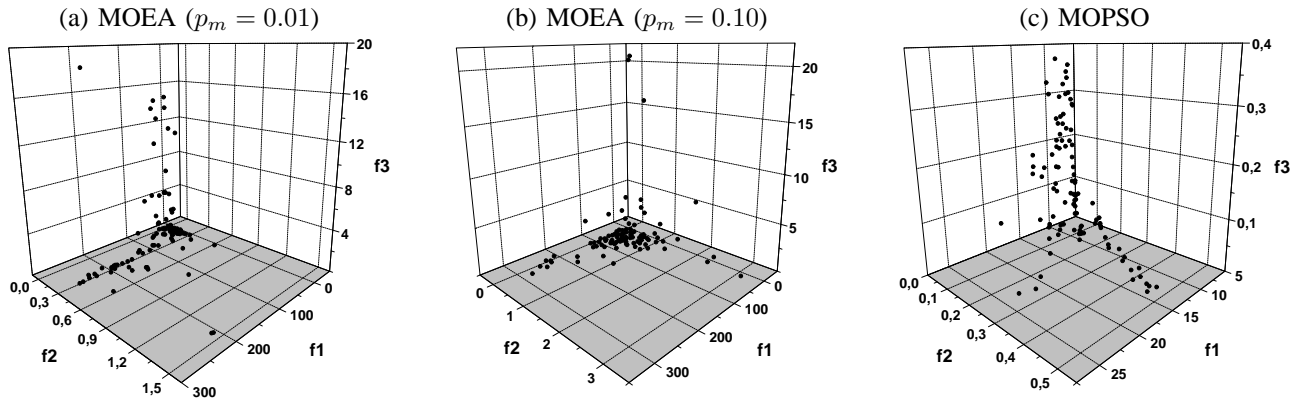


Fig. 3. The non-dominated front in objective space obtained after 3000 generations for the best three runs of a) MOEA,  $p_m=0.01$  b) MOEA,  $p_m=0.1$  c) MOPSO

$$C(MOPSO, MOEA_{0.01}) = 42\%$$

$$C(MOEA_{0.01}, MOPSO) = 2\%$$

These values show that the solutions of the MOPSO dominate most of the solutions of MOEA with  $p_m = 0.1$  and 42% of the solutions of the MOEA with  $p_m = 0.01$ . This can also be observed in Figure 3. Although, MOEA with a lower  $p_m$  is able to obtain solutions with better convergence than with higher probability, 42 % of its solutions are still dominated by the solutions of the MOPSO. In other words, for the same number of generations, MOPSO is able to achieve solutions which show a higher convergence than those obtained with MOEA. This has also been studied in detail and confirmed for standard test functions in [11].

2) *Stopping Criteria:* As the single evaluations of the objective functions are computationally expensive the applicability for real world parametrizations depends critically on the number of generations required. Further the results of the optimization are in general not known. Hence we require a reliable criteria for estimating whether further generations significantly improve the physical results obtained. However, the following issues should be considered in this application:

- A relatively small set of solutions is needed i.e., here it is not necessary to use unconstrained archives to find the whole set of Pareto-optimal solutions.
- Convergence of solutions is more important than a well distribution of solutions along the approximated Pareto-optimal front. Further convergence only makes sense to a certain threshold after which the improvements are no longer of physical significance because of other sources of errors (e.g. in the reference or the model).
- As calculating the objective functions is the most time consuming part in this application, the number of generations, particles in the population should be selected as low as possible.

It can be observed that the MOPSO obtained lower objective values than when it is run for 3000 generations. This is also

confirmed quantitatively:

$$C(MOPSO^{10000}, MOPSO^{3000}) = 85\%$$

$$C(MOPSO^{3000}, MOPSO^{10000}) = 0\%$$

For finding the least number of generations, the following experiment was performed: After running 10 independent optimizations using MOPSO for each of the runs we compared the non-dominated set ( $A_t$ ) of generation  $t$  with the non-dominated set of generation  $t - \Delta t$  ( $A_{t-\Delta t}$ ) using the  $C$  metric. Figure 4 shows the trend of this measure averaged over these 10 runs together with its standard deviation. We further picked one representative we used for further examination (run1, as shown).

We can observe that after a certain number of generations (approx. 6500) the measure nearly constant and below 20%. From this we assume that the differences between following non-dominated sets [ $A_{t-\Delta t}, A_t$ ] are small and the cost/gain of accuracy ratio i.e. the amount of generations required for further improvements is high. Hence, we suggest to take a  $C$  metric threshold as termination criteria instead of running it for a fixed number of generations. In the next section we

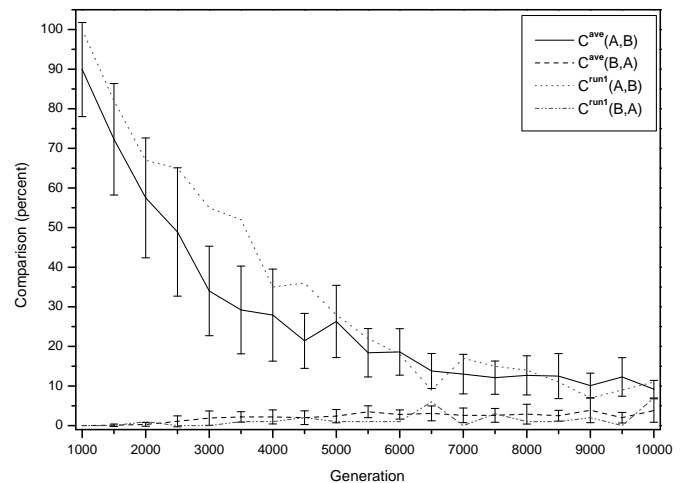


Fig. 4.  $C$  metric comparison of the non-dominated set of generations  $t$  and  $t - \Delta t$  ( $A = A_t$ ,  $B = A_{t-\Delta t}$ ,  $\Delta t = 500$ )

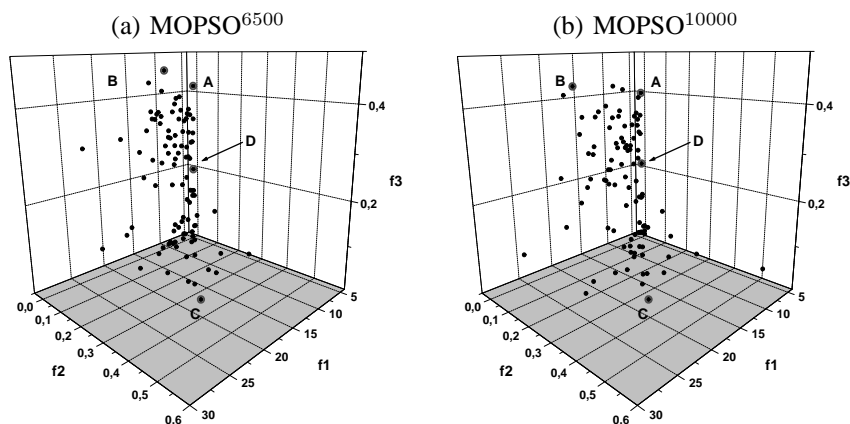


Fig. 5. The non-dominated set obtained using MOPSO after 6500 (a) and 10000 (b) generations for run1

will analyze the results obtained after the threshold set has been reached (6500 generations) and the final results (10000 generations) for run1. When comparing the non-dominated sets in objective space we can see that the differences between them are minor.

*Justification:* The selected comparison metric ( $C$  metric) compares the convergence of two non-dominated sets. In comparison to other metrics, e.g., the  $v$  metric [23], it is more applicable in our approach because we compare the solutions of one archive during several generations. The non-dominated solutions survive in the archive and the diversity of solutions can vary only when one non-dominated solution is inserted into the archive. Otherwise the diversity of solutions remains equal. Indeed, we are interested to measure the number of non-dominated solutions which remain in the archive after some generations, i.e., if there is improvement in terms of the convergence and not the diversity. This can not be achieved by a metric like the  $v$  metric, because it doesn't yield this information.

3) *Analysis of physical properties:* In the following we take a closer look at four sets of parameters taken from the non-dominated set from run1 after 6500 and 10000 generations.

In the following we regard the parameter sets having the best objective functions (A,B,C) and one point manually picked out (D) of the center of the surface defined by the non-dominated set. This point was chosen under the assumption that it represents a reasonable compromise for the different objective functions. The position of the parameter sets in objective space are shown in Figure 5. The corresponding objective values are listed in Table I.

The objectives values defined by  $f_1$  (5) and  $f_2$  (10) are not directly related to observable physical properties and hence it is difficult to judge the quality of the force field parameters obtained. Therefore we analyze more ostensive properties in the following.

*Structural Analysis:* The molecules were optimized with respect to the energy function defined by the force field (1-3) using the parameter sets A-D after 6500 and 10000 steps. Some characteristic geometrical parameters obtained using the parameter sets obtained after 6500 generations are shown

TABLE I  
OBJECTIVE VALUES FOR THE PARAMETER SETS PICKED FROM RUN 1 AFTER 6500 AND 10000 GENERATIONS. THE LINE MAX. DENOTES THE MAXIMUM VALUE OF THE RESPECTIVE OBJECTIVE WITHIN THE NON-DOMINATED SET.

	$f_1$	$f_2$	$f_3$
A <sup>6500</sup> /A <sup>10000</sup>	6.3 / 6.3	0.038 / 0.037	0.42 / 0.40
B <sup>6500</sup> /B <sup>10000</sup>	11.5 / 17.7	0.036 / 0.034	0.46 / 0.43
C <sup>6500</sup> /C <sup>10000</sup>	17.9 / 17.9	0.345 / 0.345	0.008 / 0.008
D <sup>6500</sup> /D <sup>10000</sup>	7.3 / 7.0	0.060 / 0.055	0.21 / 0.22
max. <sup>6500</sup> /max. <sup>10000</sup>	24.5 / 27.8	0.355 / 0.547	0.46 / 0.43

in Table II. For all of the sets the geometrical data shows good agreement with the reference. The deviations from the reference are less than the usual error bar and hence not significant from a physical point of view. The data obtained with the parameter sets after 10000 steps of MOPSO show the same picture (data not shown). We conclude that the accuracy obtained here after 6500 generations is sufficient for application purposes no matter which of the parameter sets A-D is chosen.

TABLE II  
CHARACTERISTICAL GEOMETRY PARAMETERS FOR MINIMUM GEOMETRIES FOR PARAMETER SETS A-D. DISTANCES  $d$  ARE IN Å, ANGLES  $\alpha$  IN °.

	A <sup>6500</sup>	B <sup>6500</sup>	C <sup>6500</sup>	D <sup>6500</sup>	Ref.
Methanol					
$d(\text{H-O})$	0.96	0.96	0.96	0.96	0.96
$d(\text{O-C})$	1.42	1.42	1.43	1.42	1.43
$\alpha(\text{H-O-C})$	108.3	108.3	108.5	108.3	108.9
Ethanol					
$d(\text{H-O})$	0.96	0.96	0.96	0.96	0.96
$d(\text{O-C})$	1.43	1.43	1.43	1.42	1.43
$d(\text{C-C})$	1.51	1.51	1.50	1.51	1.52
$\alpha(\text{H-O-C})$	109.1	109.0	109.1	109.0	109.0
$\alpha(\text{O-C-C})$	107.3	107.6	107.9	107.5	108.0

*Vibrational spectra* were calculated for each of the parameter sets at the energetic minimum. A simple ordering of the vibrational frequencies by magnitude can lead to wrong assign-

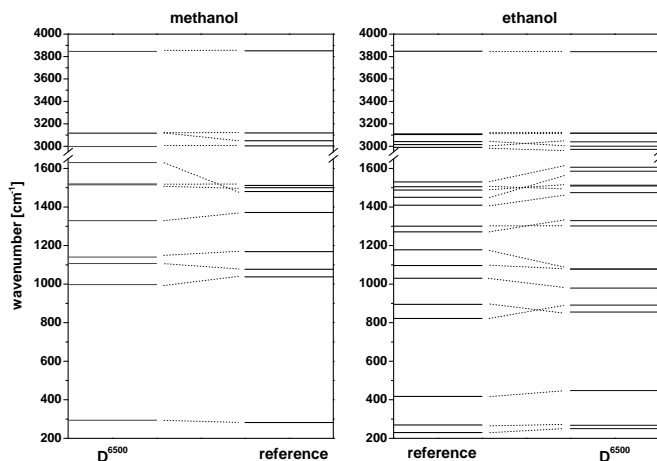


Fig. 6. Reference vibrational spectra and vibrational spectra obtained with parameter set  $D^{6500}$ . The dotted lines connect corresponding modes.

ments of vibrations (see the identification problem described in the section "Reproduction of Molecular Vibrations"). Hence the vibrational modes of the different test sets were manually assigned those of the reference. The vibrational spectrum for the two alcohols calculated with parameter set  $D^{6500}$  in comparison to the reference is shown in Figure 6. For most vibrations of both alcohols good correspondence is found though various swappings of vibrational modes can be observed. This is common when comparing different methodologies and is not a problem *per se* as relative shifts are more meaningful.

A few of the vibrational modes are significantly shifted leading to the large maximal absolute deviations as listed in Table III.

Closer observation of the vibrational spectrum shows that the modes being responsible for the maximum deviations are the same ones for all parameter sets and similar ones for both alcohols. In all of these description these modes are too high compared to the reference. We can also observe this problem with the CHARMM27 parametrization [21]. Not considering this problem the maximum absolute deviations are about  $100\text{ cm}^{-1}$  or less. The large maximum absolute deviation for ethanol using parameter set  $B^{10000}$  is hence not significant for the overall quality of the fit.

Apart from this deviation the comparison of the maximum absolute deviations shows clearly that vibrational spectra are generally better described with parameter set B or D than with parameter set C. Parameter set A describes the vibrational modes with comparable quality to the descriptions of sets B and D. This correlates well with the corresponding values of  $f_2$ . The same tendency can be observed for the *rmsd* and the mean absolute deviation. The parameter sets B and D again give the best results whereas set C gives poor results. The quality of description obtained with set A is different for ethanol and methanol. The trends of the quality of description are well reflected in the objective values  $f_2$ .

*Energetics of different conformations:* From both the deviations from the rotational profile by  $f_3$  and the individual rotational profiles it is obvious that for all of the sets in the

TABLE III  
VIBRATIONAL FREQUENCY DEVIATIONS (IN TERMS OF  
WAVENUMBERS  $\text{cm}^{-1}$ ) FOR PARAMETER SETS A-D.

	$A^{6500}$ ( $A^{10000}$ )	$B^{6500}$ ( $B^{10000}$ )	$C^{6500}$ ( $C^{10000}$ )	$D^{6500}$ ( $D^{10000}$ )
<b>Methanol</b>				
rmsd	51.4 (51.5)	52.1 (52.7)	64.8 (64.8)	52.2 (52.2)
mean abs. dev.	32.2 (32.3)	32.9 (33.8)	44.9 (44.9)	33.8 (33.5)
max abs. dev.	145.5 (145.7)	148.0 (149.9)	187.6 (187.6)	150.1 (148.9)
<b>Ethanol</b>				
rmsd	51.1 (51.1)	51.6 (55.2)	62.0 (62.0)	51.6 (51.6)
mean abs. dev.	38.6 (38.6)	39.3 (39.7)	49.4 (49.4)	38.2 (38.3)
max abs. dev.	130.6 (130.1)	126.0 (174.4)	157.8 (157.7)	135.3 (134.2)

Pareto points the deviations are well below the accuracy of the reference calculations ( $\Delta E \leq \sqrt{f_3}\text{ kcal/mol}$ ).

Also the qualitative features are well represented for all of the parameter sets considered here. This is depicted in an example for one rotational profile in Figure 7 for one bond in ethanol and the parameters obtained after 6500 steps. This particular torsional angle was chosen for illustration as it shows the most prominent deviations. Even those parameters which lead to the larger deviations in terms of  $f_3$  show the required features of maxima and minima at the respective angles. Nevertheless we observe that parameter set C and D best reproduce the rotational profile. When taking a closer look at the curves at  $60^\circ$  we get the impression that the curve generated by parameter set D is superior to set C in its description of the rotational profile. This is counterintuitive on first glance with respect to the objective values  $f_3$ . However we have to keep in mind that  $f_3$  measures the fit of *all* rotational profiles in some of which the deviations of D are larger.

Similar behavior as for this dihedral rotation is observed for the other rotational profiles (data not shown).

## V. CONCLUSION AND FUTURE WORK

In this paper, we studied the parametrization of molecular force fields using multi-objective evolutionary algorithms (MOEA) and particle swarm optimization methods (MOPSO).

The optimization methods were applied on a set of two alcohols and compared in terms of convergence. The results show that MOPSO achieves solutions with higher convergence than MOEA for the same number of generations. The MOPSO was run for a high number of generations and the termination criteria for an acceptable convergence of solutions is studied by the  $C$  metric. We showed that when using this measure as a termination criteria a reasonable description of the physical properties can be obtained while limiting the time required.

The non-dominated results achieved by MOPSO were also analyzed from an application point of view. It is obvious that both the accuracies reached for geometrical properties and the conformations are sufficient. In both cases the accuracy

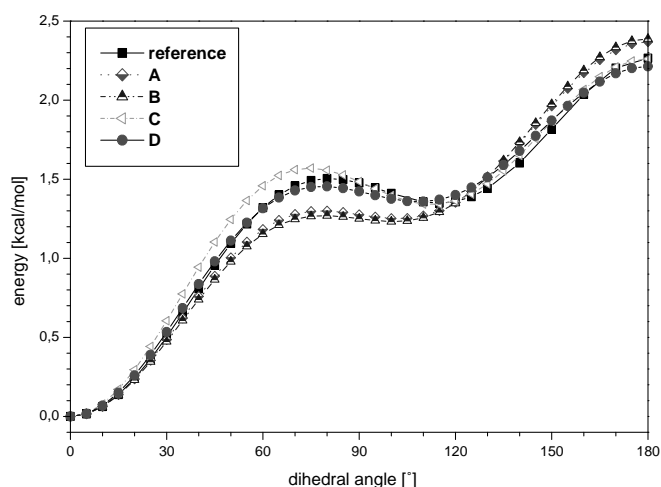


Fig. 7. Comparison of the rotational profile around the C-C[...]-O-H bond in ethanol. The rotational angles and energies are relative to the conformational minimum at 0° (staggered conformation).

is within the usual error bar of force fields. This is different for the vibrational modes where significant errors can be found in some cases.

From this result it is clear that the utilization of multi-objective optimization is superior to weighting methods. A wrong choice of weights can easily lead to a faulty description of single objectives involved. In contrast to weighting methods the approach presented here delivers a multitude of different solutions. A reasonable choice of a parameter set from the non-dominated set can lead to a good description of *all* physical properties concerned.

Further we introduced a set of objective functions for the purpose of parametrization. The results confirm that the choice of objective functions is suitable for this task.

Having introduced and validated a new method for parametrizing molecular force fields using an admittedly simple test case, we are looking forward to presenting more work on biologically relevant molecules not yet parametrized as an application of the methods introduced here. Further we will show the ability of the methods outlined here to solve larger parametrization problems ( $\approx 60$  parameters).

#### ACKNOWLEDGMENT

S. Mostaghim and P. H. König are grateful for stipends from the DFG sponsored Graduiertenkolleg “Application oriented modelling and development of algorithms”. The authors would like to thank R. Rebertisch for making the UNRAVEL package available to us.

#### REFERENCES

- [1] A. D. MacKerell. Empirical force fields: Overview and parameter optimization. In *43th Sanibel Symposium*, 2003.
- [2] J. Wang and P. A. Kollman. Automatic parameterization of force field by systematic search and genetic algorithms. *Journal of Computational Chemistry*, 22:1219–1228, 2001.
- [3] M. Busold. PhD thesis, Technische Universität München, Garching, Germany, 2001.

- [4] T. Strassner, M. Busold, and H. Radrich. FFGeneAtoR 2.0 - an automated tool for the generation of MM3 force field parameters. *Journal of Molecular Modeling*, 7:374–377, 2001.
- [5] Norman L. Allinger, Young H. Yuh, and Jenn Huei Lii. Molecular mechanics. the MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society*, 111:8551–66, 1989.
- [6] N. L. Allinger. Conformational analysis. MM2. a hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society*, 99:8127–34, 1977.
- [7] J. Hunger and G. Huttner. Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. *Journal of Computational Chemistry*, 20:455–471, 1999.
- [8] J. Hunger, S. Beyreuther, G. Huttner, K. Allinger, U. Radelof, and L. Zsolnai. How to derive force field parameters by genetic algorithms. modeling tripod-Mo(CO)<sub>3</sub> compounds as an example. *European Journal of Inorganic Chemistry*, pages 693–702, 1998.
- [9] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, 2001.
- [10] E. Zitzler, M. Laumanns, and L. Thiele. Spea2: Improving the strength pareto evolutionary algorithm. In *EUROGEN 2001, Evolutionary Methods for Design, Optimisation and Control with Applications to Industrial Problems*, 2001.
- [11] S. Mostaghim and J. Teich. Strategies for finding good local guides in multi-objective particle swarm optimization. In *IEEE Swarm Intelligence Symposium*, pages 26–33, 2003.
- [12] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Advances in protein chemistry*, 66:27–85, 2003.
- [13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.
- [14] F. Jensen. *Introduction to computational chemistry*. John Wiley & Sons, 1999.
- [15] J. Knowles and D. Corne. On metrics for comparing nondominated sets. In *IEEE Proceedings, World Congress on Computational Intelligence (CEC2002)*, pages 711–716, 2002.
- [16] E. Zitzler. *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. TIK-Schriftenreihe Nr. 30, Diss ETH No. 13398, Shaker Verlag, Germany, Swiss Federal Institute of Technology (ETH) Zurich, 1999.
- [17] S. Dasgupta and W. A. Goddard III. Hessian-biased force fields from combining theory and experiment. *Journal of Chemical Physics*, 90:7207–15, 1989.
- [18] P. Pulay, G. Fogarasi, F. Pang, and J. E. Boggs. Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives. *Journal of the American Chemical Society*, 101:2550–60, 1979.
- [19] Jr Wilson E. B., J. C. Decius, and P. C Cross. *Molecular Vibrations*. McGraw-Hill, 1955.
- [20] R. Rebertisch. PhD thesis, Universität zu Köln, Köln, Germany, 1998.
- [21] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, III Reiher W. E., B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wierkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102:3586–3616, 1998.
- [22] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople. *Gaussian 98*. Gaussian, Inc., Pittsburgh PA, 1998.
- [23] J.E. Fieldsend, R.M. Everson, and S. Singh. Using unconstrained elite archives for multi-objective optimisation. In *IEEE Transactions on Evolutionary Computation*, 2002.