# Crystal Structure Predictions Using Five Space Groups with Two Independent Molecules. The Case of Small Organic Acids

**BOUKE P. VAN EIJCK**
*Department of Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands*

*Received 30 July 2001; Accepted 12 October 2001*

**Abstract:** Crystal structure generations with two independent molecules have been performed for a series of carboxylic acids, using a slightly modified version of the OPLS force field. It was found that in this way the experimental structures with one independent molecule were produced as special cases, except for the molecules with four or more internal degrees of freedom. This work shows that a search with two independent molecules in only five space groups, although costly in computer power, can automatically also find structures with one independent molecule in many supergroups. Considering the observed abundances of structural classes, such a search should cover more than 95% of the possible homomolecular crystal structures.

 J Comput Chem 23: 456–462, 2002; DOI 10.1002/jcc.10042

**Key words:** crystal structure prediction; carboxylic acids

## Introduction

In *ab initio* crystal structure prediction the first task is to construct a list of hypothetical structures.[1,2] In this article we shall limit ourselves to crystal structures that are homomolecular, i.e., they contain only one type of molecule. Thus, solvates and other cocrystals are excluded; we have discussed the problems in crystal structure prediction for hydrates previously.[3] An overview of the occurrence of solvates in the Cambridge Crystallographic Database[4] has been published recently.[5]

A crystallographic unit cell contains a certain number of molecules, conventionally denoted by $Z$. Usually, most of these are related by symmetry elements, and a more important quantity is the number of independent molecules, which we have called $Z''$.[3] One problem in crystal structure prediction is that there are many possible space groups and that $Z''$ is unknown. In principle, one could circumvent this problem by constructing all structures in space group $P1$ with many different values for $Z$. However, this approach fails in all but very simple cases due to the overwhelming number of possible structures that can be created in that way.[3,6] Usually crystal structure prediction relies on the fact that most structures have one independent molecule and crystallize in a very limited number of space groups. This approach provides a natural way to limit the computational burden; it will be successful in the majority of cases but is bound to fail occasionally. To start with, about 8% of the structures reported up to 1987 have $Z'' > 1$.[7] Furthermore, the 10 most populated space groups cover about 96% of the reported homomolecular structures.[8] Thus, routine crystal structure predictions will fail in about 12% of the cases, even if we disregard other reasons why the search can miss a structure with $Z'' = 1$ in one of the investigated space groups.

In this article an intermediate approach is proposed, where the structure generation is carried out with two independent molecules in the five most abundant space groups, $P2_1/c$, $P\bar{1}$, $P2_1$, $P2_12_12_1$ and $P1$. This covers over 85% of the cases with $Z'' = 2$.[7–9] Furthermore, the search should also find structures that can be described with $Z'' = 1$ when the unit cell is halved or when another extra symmetry element is created. Thus, for $Z'' = 1$ one should find all minimal supergroups of the five space groups investigated. Inspection of the International Tables[10] leads to the list given in Table 1. This list comprises all monoclinic space groups, as well as the most populated orthorhombic ones and a few space groups of higher symmetry.

Even more space groups are accessible if structures with molecules on special positions are taken into account, because these structures can be constructed in a space group with lower symmetry. Belsky, Zorkaya, and Zorky[9] have used the concept of a structural class, which specifies the space group as well as the numbers of molecules in special and general positions. In a database of 19,642 homomolecular substances they encountered 305 structural classes, and they determined the numbers of structures in

***Correspondence to:*** B. P. Van Eijck; e-mail: vaneyck@chem.uu.nl

**Table 1.** Minimal Supergroups of Five Selected Space Groups.

| | $P1$ | $P\bar{1}$ | $P2_1$ | $P2_1/c$ | $P2_12_12_1$ | $A$ |
|---|---|---|---|---|---|---|
| $P1$ | x | | | | | 0.9 |
| $P\bar{1}$ | x | x | | | | 20.8 |
| $P2$ | x | | x | | | 0.03 |
| $P2_1$ | x | | x | | | 5.7 |
| $C2$ | x | | x | | | 0.9 |
| $Pm$ | x | | | | | 0.002 |
| $Pc$ | x | | | | | 0.4 |
| $Cm$ | x | | | | | 0.06 |
| $Cc$ | x | | | | | 1.0 |
| $P2/m$ | | x | | | | 0.02 |
| $P2_1/m$ | | x | x | x | | 0.7 |
| $C2/m$ | | x | | x | | 0.5 |
| $P2/c$ | | x | | x | | 0.5 |
| $P2_1/c$ | | x | x | x | | 35.8 |
| $C2/c$ | | x | | x | | 7.5 |
| $P222_1$ | | | x | | | 0.02 |
| $P2_12_12$ | | | x | | x | 0.5 |
| $P2_12_12_1$ | | | x | | x | 8.8 |
| $C222_1$ | | | x | | x | 0.2 |
| $I2_12_12_1$ | | | | | x | 0.01 |
| $Pmc2_1$ | | | x | | | 0.02 |
| $Pca2_1$ | | | x | | | 0.7 |
| $Pmn2_1$ | | | x | | | 0.09 |
| $Pna2_1$ | | | x | | | 1.5 |
| $Cmc2_1$ | | | x | | | 0.17 |
| $Pnna$ | | | | x | | 0.08 |
| $Pmna$ | | | | x | | 0.02 |
| $Pcca$ | | | | x | | 0.04 |
| $Pbam$ | | | | x | | 0.04 |
| $Pccn$ | | | | x | | 0.4 |
| $Pbcm$ | | | | x | | 0.13 |
| $Pnnm$ | | | | x | | 0.08 |
| $Pbcn$ | | | | x | | 0.9 |
| $Pbca$ | | | | x | x | 3.8 |
| $Pnma$ | | | | x | x | 1.5 |
| $Cmca$ | | | | x | | 0.17 |
| $P4_1$, $P4_3$ | | | x | | | 0.17 |
| $P4_12_12$, $P4_32_12$ | | | | | x | 0.4 |
| $P3$ | x | | | | | 0.03 |
| $P3_1$, $P3_2$ | x | | | | | 0.15 |
| $R3$ | x | | | | | 0.15 |
| $P\bar{3}$ | | x | | | | 0.10 |
| $R\bar{3}$ | | x | | | | 0.5 |
| $P6_1$, $P6_5$ | | | x | | | 0.11 |
| $P6_3$ | | | x | | | 0.07 |
| $P2_13$ | | | | | x | 0.07 |

$A$ is the abundance percentage in the CSD (issue October 2000).

each class. From that survey we found that in our proposed search method 127 structural classes would be accessible, comprising 95.9% of the structures. Interestingly, a quick and dirty counting of overall space group frequencies in the October 2000 issue of the CSD leads to a similar result: the space groups enumerated in Table 1 account for 95.7% of 220,000 entries. The most important missing structural classes contain general positions only; they are $Pca2_1$, $Pbca$ and $Pna2_1$ with $Z'' = 2$, followed by $Fdd2$ and $I4_1/a$ with $Z'' = 1$.

Thus, the proposed scheme should be potentially successful in at least 95% of the homomolecular crystal structures. Furthermore, there is an organizational advantage in investigating only five space groups with the same protocol. However, one has to pay a price. Allowing two independent molecules in the search leads to a large increase in the number of possible structures, and very many trial structures are necessary before one can be reasonably sure that all promising ones have been encountered at least once. With the present generation of computers, however, this disadvantage is only prohibitive for large or very flexible molecules. Extending the method to more than two independent molecules is not recommended, as structure generations tend to miss essential structures when the number of degrees of freedom exceeds about 20.[3]

The situation for chiral molecules needs careful treatment. For a pure enantiomorph the consideration of space groups $P2_1$, $P2_12_12_1$, and $P1$ will suffice. To construct all possible crystals that can be formed from a racemic mixture, however, it will be necessary to add not only the space groups $P2_1/c$ and $P\bar{1}$, but also $P2_1$, $P2_12_12_1$, and $P1$ with two enantiomers. Note that such a mixture might produce either racemic crystals or equal quantities of two enantiomorphs.[11]

In this article we apply the proposed scheme to crystal structure predictions for carboxylic acids. These molecules show an interesting variety of conformational possibilities and crystal structure symmetries, including special positions and two cases with $Z'' = 2$. Although some force field development was found to be necessary, the emphasis of the present work is on the performance of the proposed method of structure generation. After a list of possible structures has been obtained, more sophisticated methods can be used to select the most promising polymorph(s). The group of Price has already developed highly successful potentials for this class of compounds.[12,13] Although that work has been limited to rigid molecules, *ab initio* calculations for crystal structures of flexible molecules are becoming possible.[14,15] We are at present working on the extension of the necessary intermolecular potential to the case of carboxylic acids.

## Choice of Compounds

The compounds to be studied contained up to four carbon atoms, of which one or two are involved in a carboxyl group and each of the other ones may carry a hydroxyl group. They are listed in Table 2, with references to experimental data from the Cambridge Structural Database.[4] In generating possible structures for these compounds the carboxyl group was assumed to have its usual H . . . O *cis* planar form, but internal rotation about other single bonds ($N_f$ in number, not counting methyl groups) was allowed.

For several substances more than one well-characterized polymorph is known; some forms are only found in certain regions of pressure and temperature. In fact, the existence of another polymorph has also been reported for succinic acid,[16,17] but no useful atomic coordinates are available. There exists a form of DL-malic acid,[18] which exhibits disorder and cannot be modeled by static methods.

**Table 2.** Selected Carboxylic Acids.

| Substance | $N_f$ | Form | Reference | Space group | $S$ | $Z''$ |
|---|---|---|---|---|---|---|
| Formic acid | 0 | | FORMAC01 | $Pna2_1$ | | 1 |
| Acetic acid | 0 | *o* | ACETAC03 | $Pna2_1$ | | 1 |
| | | *m* | Ref. 26 | $P2_1/n$ | | 1 |
| Propionic acid | 1 | | PRONAC | $P2_1/c$ | | 1 |
| *n*-Butyric acid | 2 | | BUTRAC | $C2/m$ | *m* | 1 |
| Glycollic acid | 2 | | GLICAC10 | $P2_1/c$ | | 2 |
| L-Lactic acid | 2 | | YILLAG | $P2_12_12_1$ | | 1 |
| α-Hydroxyisobutyric acid | 2 | | HXIBAC | $P2_1/n$ | | 1 |
| Oxalic acid | 0 | *o* | OXALAC03 | $Pbca$ | $\bar{1}$ | 1 |
| | | *m* | OXALAC04 | $P2_1/c$ | $\bar{1}$ | 1 |
| Malonic acid | 2 | *t* | MALNAC02 | $P\bar{1}$ | | 1 |
| | | *o* | MALNAC03 | $Pbcn$ | 2 | 1 |
| Methylmalonic acid | 2 | | MEMALA | $P\bar{1}$ | | 1 |
| Succinic acid | 3 | | SUCACB03 | $P2_1/c$ | $\bar{1}$ | 1 |
| Tartronic acid | 3 | | HMALAC01 | $P2_12_22_1$ | | 1 |
| L-Malic acid | 4 | | COFRUK10 | $P2_1$ | | 2 |
| DL-Malic acid | 4 | | DLMALC11 | $P2_1/c$ | | 1 |
| D-Tartaric acid | 5 | | TARTAC01 | $P2_1$ | | 1 |
| Mesotartaric acid | 5 | | TARTAM | $P\bar{1}$ | | 1 |

$N_f$ is the number of single bonds about which internal rotation was allowed.
$S$ (if present) denotes the symmetry of a special position occupied by the molecule.
Polymorphs are indicated simply by the first letter of their crystal systems.

Several substances have been studied by neutron diffraction, so their hydrogen coordinates should be accurately determined. For the other structures these coordinates are not always reliable, especially in older work. The structure of mesotartaric acid is rather doubtful, as it contains a significantly nonplanar carboxylic group. For butyric acid and *t*-malonic acid no hydrogen positions were reported, and in L-malic acid all hydroxyl hydrogen atoms are missing. As the latter structure contains two independent molecules in the asymmetric unit, it was not immediately obvious where these atoms should be placed.

## Force Field Development

The first requirement of a force field is that it should reproduce the observed crystal structures without excessive changes in geometry. A second test is possible after generation of sufficient hypothetical structures, which produces a list of structures ordered to energy. In that list each structure has a ranking $R$ and a relative energy $\Delta E$ with respect to the global energy minimum. If the experimentally observed structure(s) have an excessively large relative energy (say, over 10 kJ/mol), this is a strong indication that there is something wrong with the force field.

In this study we used the OPLS force field[19–21] as basis. In previous work this was found to give reasonable results in similar work on carbohydrates,[22] although it was later seen to produce unrealistically low energies for the *trans* O—C—C—O conformation in polyalcohols.[23]

All necessary charges and van der Waals parameters were available from the OPLS force field. They are collected in Table 3. Some parameters for bond stretching and angle bending involving $sp^2$ carbon atoms were missing; they were taken over from comparable ones. More essential is the choice of missing torsional parameters. It was found necessary to add strong twofold terms to enhance the planarity of various groups of atoms: $V_2 = 40$ kcal/mol for HA—OA—C—X carboxyl groups, $V_2 = 10$ kcal/mol for X—C—C—X in oxalic acid, and $V_2 = 4$ kcal/mol for OH—CT—C—X in hydroxycarboxylic acids. The exact values for $V_2$ are not critical, but such terms are necessary to reproduce the observed trends in geometries.

The necessity of other modifications to the force field became only apparent after tentative structure generations; thus, the force field development must be an iterative procedure. For instance, the dihedral angle OH—CT—C—O is less than 30° in all hydroxy

**Table 3.** Charges and Lennard–Jones Parameters in the OPLS Force Field.

| | Atom types | $C_{12}$(Mcal mol$^{-1}$ Å$^{12}$) | $C_6$(kcal mol$^{-1}$ Å$^6$) | $q$ |
|---|---|---|---|---|
| C | C in COOH | 3248.06 | 1167.98 | 0.52 |
| O | ═O in COOH | 380.00 | 564.98 | −0.44 |
| OA | —O in COOH | 361.38 | 495.72 | −0.53 |
| HA | H in COOH | 0 | 0 | 0.45 |
| CT | aliphatic C | 892.11 | 485.30 | [a] |
| HC | H on C or CT | 7.15 | 29.30 | 0.06 |
| OH | O in $CH_nOH$ | 476.62 | 569.30 | −0.683 |
| HO | hydroxyl H on OH | 0 | 0 | 0.418 |

[a]The charges on CT are adjusted to obtain zero charge for the groups $CH_n$ and $CH_nOH$. The charge on HC in formic acid is zero.

**Table 4.** Torsional Parameters in the OPLS-AC Force Field.

| Torsion | $V_1$ | $V_2$ | $V_3$ |
|---|---|---|---|
| C(T)—CT—CT—C | −0.550 | 0 | 1.000 |
| CT—CT—CT—HC | 0 | 0 | 0.366 |
| C—CT—CT—HC | −6 | 0 | 0 |
| C—CT—CT—OH | −14 | 0 | 0 |
| OH—CT—CT—OH | −6 | 0 | 0 |
| HC—CT—CT—HC | 0 | 0 | 0.318 |
| OH—CT—CT—HC | 0 | 0 | 0.468 |
| HO—OH—CT—C(T) | 2.674 | −2.883 | 1.026 |
| HO—OA—CT—HC | 0 | 0 | 0.450 |
| O—C—CT—CT | 0 | 0.546 | 0 |
| OA—C—CT—CT | 1.000 | 0.546 | 0.450 |
| O—C—CT—C | 0 | 2 | 0 |
| OA—C—CT—C | 0 | 0 | 0 |
| O(A)—C—CT—HC | 0 | 0 | 0 |
| OH—CT—C—O | −2 | 4 | 0 |
| OH—CT—C—OA | 0 | 4 | 0 |
| O(A)—C—C—O(A) | 0 | 10 | 0 |
| HA—OA—C—C(T) | 3 | 40 | 0 |
| HA—OA—C—HC | 0 | 40 | 0 |
| HA—OA—C—O | 0 | 40 | 0 |

The standard OPLS notation has been followed: units are kcal/mol and the contribution of one torsional term to the energy is defined as $V_{\text{tors}} = V_1(1 + \cos\phi)/2 + V_2(1 - \cos 2\phi)/2 + V_3(1 + \cos 3\phi)/2$. Entries in three decimal places were taken from refs. 20 and 21, the other ones were roughly adjusted to reproduce observed conformational preferences.

acids except one (monoclinic mesotartaric acid hydrate[24] where it is 173°). To reproduce this preference for dihedral angles around zero, an additional term $V_1 = -2$ kcal/mol was added. Without this term the modeling of the experimental structures hardly changes, but in structure predictions many structures with dihedral angles around 180° are found to have improbably low energies. Likewise, the dihedral angles C—CT—CT—OH in malic acid and OH—CT—CT—OH in the tartaric acids would be *trans* for all low-energy structures, whereas *gauche* conformations are observed consistently. Such considerations led to the introduction of several new torsional parameters (Table 4). We shall denote the resulting force field with OPLS-AC, to emphasize that it has been extended for acids beyond the original parameterization.

## Reproduction of Experimental Structures

The experimental structures were subjected to energy minimization in the OPLS-AC force field. The results are given in Table 5. It is seen that the oxalic acids are reproduced badly. The difficulty for transferable force fields to model oxalic acid adequately has already been noted by Nobeli and Price.[12] For the other substances the agreement varies from adequate to poor. The good results for succinic acid are not surprising, because essential OPLS parameters were obtained from that molecule.[21] It is obvious that the force field needs improvement, but we did not pursue that matter as it was not the main objective of this study.

For tartronic acid the geometry shifts are not exceptionally large, but they are sufficient to change the space group symmetry from $P2_12_12_1$ into *Pnma*. In five structures the molecules occupy a special position (Table 2). UPACK cannot handle that situation, but these structures can be studied in appropriate subgroups. Then the original symmetry may be lost during energy minimization, but this was never observed.

## Generation of Crystal Structures

Hypothetical structures for all compounds were generated using the random search facility in the UPACK program package.[3] The number of independent molecules will be denoted by $G$; as explained above, a calculation for $G = 2$ creates structures with $Z'' = 2$ as well as $Z'' = 1$. Each independent molecule was placed with random position and random orientation in a crystal cell with random crystallographic parameters, constrained by the density found in a preliminary calculation. Random values were also assigned to the dihedral angle of each single bond around which internal rotation was allowed, giving another $G \times N_f$ degrees of freedom. A preliminary energy minimization, using the OPLS-AC force field with fully flexible molecules, was performed for possibly acceptable structures. For each substance 10,000 structures were generated with $G = 2$ in the five space groups discussed in the Introduction. For comparison, 5000 structures were generated with $G = 1$ in the ten space groups $P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $P2_1$, $Pbca$, $C2/c$, $Pna2_1$, $Cc$, $Pca2_1$, and $C2$. Equivalent structures were removed by clustering, after which the energy minimization was continued for structures within an energy window of 30 kJ mol$^{-1}$ until the residual forces were less than 0.0001 kJ mol$^{-1}$ Å$^{-1}$. A second clustering delivered the final list of possible structures.

Most calculations were carried out on a Gateway personal computer equipped with a Pentium II 500 MHz processor and running under Linux. For $G = 2$, the total computing times varied from a day for formic acid to 3 weeks for DL-malic acid. Large variations occurred between apparently similar compounds; especially for butyric acid the structure-generating algorithm turned out to be extremely slow for unknown reasons. The calculations for $G = 1$ were between 5 and 10 times faster.

Table 6 shows the numbers of hypothetical structures in an energy window of 7 kJ/mol ($N_7$). These numbers also exhibit a large variation between the various substances. Allowing structures with more than one independent molecule increases the number of possible structures considerably. Especially for acetic acid and propionic acid astoundingly many packings are possible. Nevertheless, only two instances were found where a $Z'' = 2$ structure had a lower energy than the best $Z'' = 1$ structure: propionic acid (0.3 kJ/mol) and D-tartaric acid (1.0 kJ/mol).

Ideally, the structure generation with $G = 2$ should rediscover all structures that were found with $G = 1$. In practice, this was not the case. Table 6 gives the numbers of missed structures [$M(2)$], which are by no means as small as would be desirable. Moreover, many of them have a low energy ($\Delta E_M(2)$) with respect to the global minimum. For the malic acids and tartaric acids the structure generation for $G = 2$ is entirely incomplete. This is not surprising, because the number of degrees of freedom here is

**Table 5.** Geometry Shifts Upon Energy Minimization in the OPLS-AC Force Field.

| Substance | $\Delta L$(%) | $\Delta\phi$(°) | $\Delta X$(Å) | $\Delta\tau$(°) | $\Delta R$(Å) | $\Delta\omega$(°) |
|---|---|---|---|---|---|---|
| Formic acid | 5.4 | | 0.16 | 8 | 0.03 | |
| *o*-Acetic acid | 8.7 | | 0.36 | 4 | 0.00 | |
| *m*-Acetic acid | 6.6 | 3.3 | 0.17 | 5 | 0.01 | |
| Propionic acid | 1.7 | 0.4 | 0.28 | 6 | 0.04 | 7 |
| *n*-Butyric acid | 7.4 | 2.5 | 0.08 | 3 | 0.06 | 0 |
| Glycollic acid | 1.9 | 0.4 | 0.24 | 13 | 0.04 | 9 |
| L-Lactic acid | 6.0 | | 0.72 | 8 | 0.05 | 4 |
| $\alpha$-Hydroxyisobutyric acid | 3.6 | 2.8 | 0.32 | 4 | 0.04 | 11 |
| *o*-Oxalic acid | 20.3 | | 0.00 | 11 | 0.09 | 0 |
| *m*-Oxalic acid | 14.1 | 0.8 | 0.03 | 9 | 0.09 | 0 |
| *t*-Malonic acid | 2.1 | 4.7 | 0.18 | 6 | 0.07 | 30 |
| *o*-Malonic acid | 4.1 | | 0.08 | 1 | 0.02 | 5 |
| Methylmalonic acid | 3.0 | 2.3 | 0.10 | 3 | 0.08 | 8 |
| Succinic acid | 2.5 | 3.1 | 0.08 | 8 | 0.01 | 6 |
| Tartronic acid | 3.1 | | 0.17 | 12 | 0.04 | 13 |
| L-Malic acid | 0.9 | 1.9 | 0.16 | 6 | 0.04 | 17 |
| DL-Malic acid | 5.8 | 7.0 | 0.20 | 15 | 0.04 | 25 |
| D-Tartaric acid | 3.5 | 3.0 | 0.11 | 15 | 0.07 | 11 |
| Mesotartaric acid | 6.8 | 0.9 | 0.14 | 10 | 0.10 | 11 |

$\Delta L$ and $\Delta\phi$ are the root-mean-square shifts in the cell axes and in the cell angles, respectively. $\Delta X$ is the translation of the molecule (calculated via fractional coordinates) and $\Delta\tau$ is the rotation. $\Delta R$ is the root-mean-square shift in the hydrogen bond lengths (defined as O . . . O distances between 2.3 and 2.9 Å). $\Delta\omega$ is the largest shift in a dihedral angle (C and O atoms only).

significantly larger than 20. Earlier experience[3] has shown that this is the limit of applicability of the random search method. Rather disconcertingly, the global minimum was also missed for the more simple compounds formic acid and butyric acid.

Our default search procedure for $G = 1$ is considered complete when 5000 structures are generated. One possible check on the degree of completeness is looking for additional $Z'' = 1$ structures generated in the $G = 2$ search. Within a window of 7 kJ/mol there

**Table 6.** Details of the Search in the OPLS-AC Force Field.

| Substance | $N_7(1)$ | $N_7(1 + 2)$ | $M(2)$ | $\Delta E_M(2)$ | $M(1)$ | $\Delta E_M(1)$ |
|---|---|---|---|---|---|---|
| Formic acid | 57 | 323 | 8 | 0.0 | 0 | 9.5 |
| Acetic acid | 276 | 2323 | 81 | 1.3 | 3 | 2.3 |
| Propionic acid | 367 | 2352 | 132 | 2.3 | 13 | 2.8 |
| *n*-Butyric acid | 254 | 384 | 182 | 0.0 | 6 | 2.9 |
| Glycolic acid | 77 | 179 | 37 | 0.9 | 1 | 7.7 |
| L-Lactic acid | 22 | 144 | 1 | 6.6 | 0 | 8.0 |
| DL-Lactic acid | 43 | 150 | 10 | 3.0 | 2 | 4.5 |
| $\alpha$-Hydroxyisobutyric acid | 35 | 100 | 8 | 4.4 | 0 | 15.1 |
| Oxalic acid | 68 | 151 | 11 | 2.3 | 0 | 13.8 |
| Malonic acid | 35 | 75 | 9 | 3.5 | 0 | 8.2 |
| Methylmalonic acid | 95 | 274 | 42 | 0.4 | 2 | 5.6 |
| Succinic acid | 33 | 77 | 2 | 4.7 | 2 | 5.9 |
| Tartronic acid | 10 | 28 | 2 | 2.9 | 0 | 7.9 |
| L-Malic acid | 18 | 37 | 5 | 2.2 | 0 | 7.6 |
| DL-Malic acid | 38 | 43 | 23 | 1.3 | 0 | 7.7 |
| D-Tartaric acid | 22 | 39 | 3 | 1.2 | 1 | 2.1 |
| DL-Tartaric acid | 13 | 16 | 6 | 0.6 | 0 | 16.7 |
| Mesotartaric acid | 12 | 15 | 11 | 0.0 | 0 | 15.5 |

$N_7(1)$ is the number of structures within 7 kJ/mol of the global minimum, as calculated for $G = 1$; $N_7(1 + 2)$ refers to the combination of the lists from $G = 1$ and $G = 2$. $M(2)$ indicates how many structures out of the set $N_7(1)$ were missed for $G = 2$, and $\Delta E_M(2)$ gives the lowest relative energy of the missing structures (kJ/mol). Likewise, $M(1)$ and $\Delta E_M(1)$ refer to the $Z'' = 1$ structures in a window of 7 kJ/mol that were found for $G = 2$ but missed for $G = 1$.

**Table 7.** Rankings and Energy Differences in the OPLS-AC Force Field.

| Substance | $\Delta E(1 + 2)$ | $R(1 + 2)$ | $R(1)$ | $F(1)$ | $F(2)$ |
|---|---|---|---|---|---|
| Formic acid | 6.5 | 218 | 48 | 313 | 38 |
| *o*-Acetic acid | 0.6 | 17 | 9 | 202 | 14 |
| *m*-Acetic acid | 0.6 | 14 | 6 | 267 | 284 |
| Propionic acid | 3.1 | 142 | 32 | 46 | 27 |
| *n*-Butyric acid | 3.9 | 85 | 63 | 132 | 4 |
| Glycollic acid | 3.8 | 21 | — | — | 8 |
| L-Lactic acid | 3.3 | 17 | 9 | 232 | 11 |
| α-Hydroxyisobutyric acid | 0 | 1 | 1 | 59 | 20 |
| *o*-Oxalic acid | 3.7 | 34 | 20 | 1251 | 72 |
| *m*-Oxalic acid | 3.8 | 37 | 23 | 700 | 171 |
| *t*-Malonic acid | 6.3 | 50 | 25 | 554 | 46 |
| *o*-Malonic acid | 0 | 1 | 1 | 133 | 5 |
| Methylmalonic acid | 0.9 | 6 | 6 | 154 | 23 |
| Succinic acid | 1.1 | 5 | 3 | 46 | 62 |
| Tartronic acid | 1.2 | 3 | 3 | 236 | 34 |
| L-Malic acid | 14.4 | 555 | — | — | 1 |
| DL-Malic acid | 23.0 | 7865 | 2139 | 2 | 0 |
| D-Tartaric acid | 13.4 | 369 | 97 | 103 | 105 |
| Mesotartaric acid | 12.0 | 116 | 103 | 119 | 0 |

Energy differences $\Delta E$ (kJ/mol) and rankings $R$ refer to the experimental structure with respect to the global energy minimum. $F$ is the number of times the experimental structure was found. Labels (1) and (2) refer to the lists from $G = 1$ and $G = 2$, respectively; $(1 + 2)$ indicates the combination of these two lists. For the ranking of DL-malic acid the structures of L-malic acid were also considered.

were in total 103 such structures, but 73 of them did not belong to one of the 10 space groups studied for $G = 1$. The numbers of really missed structures are given as $M(1)$ in Table 6; they are satisfactorily small with respect to $N_7(1)$.

In a successful structure generation the energy-minimized experimental structure(s) should be present in the list. The numbers of times that they were found (before clustering) are reported in Table 7 as $F(1)$ and $F(2)$ for the structure generations with $G = 1$ and $G = 2$, respectively. For reliable results $F$ should be at least about 5. It can be seen that $F$ depends on the space group; its values are especially large for molecules on special positions that were found in several subgroups. In retrospect, it would be more efficient to choose the number of trial structures as a function of the ability of each space group to create distinct structures.

For L-malic acid, where several possibilities for the hydrogen positions in the experimental structure exist, the entries refer to the one with lowest energy. This hydrogen bonding scheme corresponds to the one proposed by van der Sluis and Kroon.[25]

In accordance with the observations of Price and Beyer,[13] for unsubstituted monocarboxylic acids dimers as well as catemers were encountered in all energy ranges. For compounds with additional hydroxyl groups no pure dimer structures were found, hydrogen-bonded chains connecting molecules in one, two, or three dimensions. Here, as well as for the dicarboxylic acids, all kinds of hydrogen-bond schemes are possible; as far as we could ascertain, no preference for one special type was present.

## Conclusion

The relative energies of the observed polymorph(s) ($\Delta E$ in Table 7) fall within the expected range of accuracy for empirical force fields, except for the malic acids and the tartaric acids. For these complex hydroxy-dicarboxylic acids the force field is obviously inadequate. At the other extreme, $\Delta E$ is also disappointingly large for the very small molecule of formic acid.

Obviously, more sophisticated calculations are needed to arrive at a reliable structure prediction. For carbohydrates, we have found that a quantum-chemical approach, preferably extended with corrections from energy to free energy, can be quite successful.[15] For carboxylic acids we have not yet succeeded in finding a sufficiently accurate parameterization for the intermolecular energy. Apart from that, such calculations are too time-consuming to be applied to hundreds of structures. Even if we assume, perhaps somewhat optimistically, that the set of structures within an energy window of 7 kJ/mol will contain the experimentally observed polymorph(s), Table 6 shows that the number of structures needed for further consideration is still way too large. Thus, it will be necessary to work in two directions on better force fields: simple ones for fast structure generation as well as extremely accurate ones for final ranking.

In the early phase of this work we found that some crystal structure predictions failed for the simple reason that the force field favored physically unrealistic conformations excessively. Only after adjustment of some essential dihedral angle parameters could we expect to find the observed structures at all. This emphasizes that the correct reproduction of an experimental crystal structure

(as a local energy minimum) is a necessary but not a sufficient condition for the general reliability of a force field.

Returning to the main subject, Table 7 shows that a search with two independent molecules ($G = 2$) in only five space groups was in most cases sufficient to produce the experimental structures. As expected, this procedure leads mostly to smaller numbers of hits ($F$) and always to higher rankings ($R$). The problematic cases are, again, found in the malic acids and tartaric acids. The difficulty here is probably not just the failure of the force field but mainly the large number of internal degrees of freedom (4 and 5, respectively). Thus, these failures are not surprising. Even if the experimental structures had been found, the investment in computer time for such flexible molecules tends to become prohibitive. In more favorable cases, a search with $G = 2$ in only five space groups may suffice to investigate many supergroups with $Z'' = 1$ implicitly, covering more than 95% of the possibilities for homomolecular crystal structures.

Lists of hypothetical crystal structures are available on request.

## References


1. Gdanitz, R. J. In Theoretical Aspects and Computer Modeling of the Molecular Solid State; Gavezzotti, A., Ed.; John Wiley & Sons, Chicester, 1997, p. 185.
2. Verwer, P.; Leusen, F. J. J. In Reviews in Computational Chemistry; Lipkowitz, K. B.; Boyd, D. B., Eds.; Wiley-VCH, New York, 1998, p. 327, Vol. 12.
3. van Eijck, B. P.; Kroon, J. Acta Crystallogr 2000, B56, 535.
4. Allen, F. H.; Kennard, O. Chem Design Automat News 1993, 8, 31.
5. Görbitz, C. H.; Hersleth, H.-P. Acta Crystallogr 2000, B56, 526.
6. Gao, D.; Williams, D. E. Acta Crystallogr 1999, A55, 621.
7. Padmaja, N.; Ramakumar, S.; Viswamitra, M. A. Acta Crystallogr 1990, A46, 725.
8. Cole, J. C. PhD thesis, University of Bristol, England, 1995.
9. Belsky, V. K.; Zorkaya, O. N.; Zorky, P. M. Acta Crystallogr 1995, A51, 473.
10. Hahn, T. International Tables for X-ray Crystallography, Reidel, New York, 1983, Vol. A.
11. Brock, C. P.; Schweizer, W. B.; Dunitz, J. D. J Am Chem Soc 1991, 113, 9811.
12. Nobeli, I.; Price, S. L. J Phys Chem 1999, A103, 6448.
13. Beyer, T.; Price, S. L. J Phys Chem 2000, B104, 2647.
14. van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. J Comput Chem 2001, 22, 805.
15. van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. J Phys Chem 2001, B105, 10573.
16. Rieck, G. D. Rec Trav Chim Pays-Bas 1944, 63, 170.
17. Petropavlov, N. N.; Yarantsev, S. B. Sov Phys Crystallogr 1983, 28, 666.
18. van Loock, J. F. J.; van Havere, W.; Lenstra, A. T. H. Bull Soc Chim Belg 1981, 90, 161.
19. Jorgensen, W. L.; Maxwell, D. S.; Tirado–Rives, J. J Am Chem Soc 1996, 118, 11225.
20. Damm, W.; Frontera, A.; Tirado–Rives, J.; Jorgensen, W. L. J Comput Chem 1997, 18, 1955.
21. Price, D. J.; Roberts, J. D.; Jorgensen, W. L. J Am Chem Soc 1998, 120, 9672.
22. van Eijck, B. P.; Kroon, J. J Comput Chem 1999, 20, 799.
23. Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. J Am Chem Soc 2000, 122, 3500.
24. Bootsma, G. A.; Schoone, J. C. Acta Crystallogr 1967, 22, 522.
25. van der Sluis, P.; Kroon, J. J Cryst Growth 1989, 97, 645.
26. Allan, D. R.; Clark, S. J Phys Rev 1999, B60, 6328.