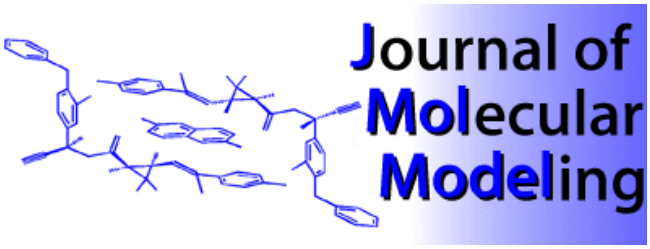
J. Mol. Model. **1998**, *4*, 132 – 144



FULL PAPER

Crystal Structure Prediction based on Statistical Potentials

Detlef W. M. Hofmann[1] and Thomas Lengauer[2,3]

[1]Institute for Algorithms and Scientific Computing, German National Research Center for Information Technology (GMD–SCAI), Schloß Birlinghoven, D-53754 Sankt Augustin, Germany. E-mail: detlef.hofmann@gmd.de

[2]Institute for Algorithms and Scientific Computing, German National Research Center for Information Technology (GMD–SCAI), Schloß Birlinghoven, D-53754 Sankt Augustin, Germany. E-mail: thomas.lengauer@gmd.de

[3]Department of Computer Science, University of Bonn, Römerstrasse 164, D-53117 Bonn, Germany.

Received: 1 October 1997 / Accepted: 4 March 1998 / Published: 23 March 1998

Abstract Organic molecule crystals are becoming more and more important in applications like piezoelectricity, ferroelectricity and pigments. These properties depend on the molecule and on the crystal structure. For this reason much effort is being made to predict the crystal structure of organic molecules. We have developed a new algorithm differing mainly in three features from other approaches (simulated annealing, Monte Carlo etc.). First, we analyze just one molecule for proper symmetry operations building up the crystal; second, the program works in a discrete space; and finally the scoring function (energy function) is derived statistically from known crystal structures and tabulated. Our program computes a list of crystal structures weighted according to our scoring function.
The new algorithm FlexCryst is currently implemented for the four space groups $P1$, $P\bar{1}$, $P2_1$, and $P2_12_12_1$. The three latter space groups are widespread in nature. The algorithm computes structural models of acceptable quality and shows excellent time performance. During our validation we found the experimental structure among the structures proposed by the algorithm in 123 of 129 cases for $P1$, in 66 of 95 cases for $P\bar{1}$, in 73 of 100 cases for $P2_1$, and in 94 of 98 cases for $P2_12_12_1$. The performance depends on the space group. In the case of $P1$ the run time per molecule is about two minutes and increases up to roughly one hour for the space group $P2_1$.

Keywords Crystal structure prediction, Statistical potentials, Similarity of crystals, Discrete molecular modeling

Introduction

One of the most fundamental unsolved problems in chemistry is predicting how a molecule will pack in the solid state solely on the basis of its molecular structure [1].

Ab-initio crystal structure prediction is still considered a long-term goal [2]. A formidable obstacle to such a prediction is the existence of a large number of local minima in the high-dimensional potential energy surface of the crystal, which makes it extremely difficult to locate the most stable structure [3]. Sometimes thousands of quite different local minima can fall within a narrow energy range (40 $kJ \cdot mol^{-1}$), as has been witnessed for monosaccharides [4].

Correspondence to: D. Hofmann

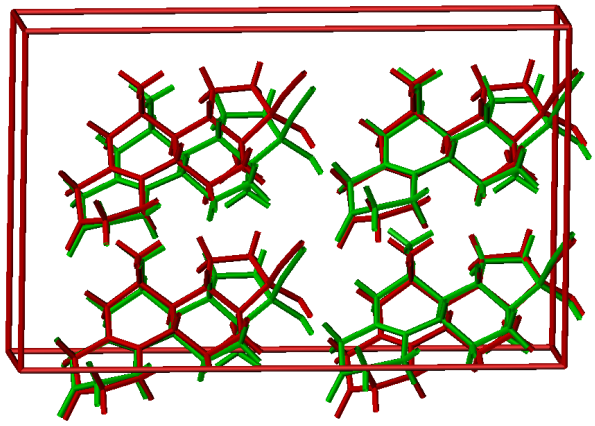


Figure 1 *Recomputation of the structure of an organic molecule (red=observed, green=computed)*

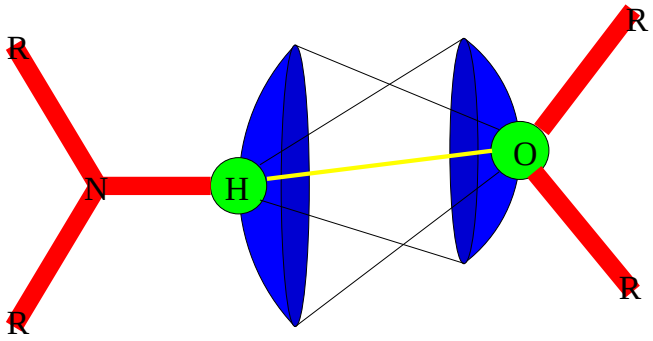


Figure 2 *Model of a hydrogen bond. The two interaction centers (green) are lie on the interaction surface (blue) of the other unit forming an interaction (yellow)*

This effect is also observed in nature, where many different crystal structures can often be found for one molecule.

Even so, sometimes the problem can be rendered feasible. Sometimes, The X-ray powder spectrum is available, but the structure cannot be solved from the spectrum due to the phase problem. The phase problem arises, when the cell parameters (the length of the translation vectors, the angles between them, and the space group) can be determined but the orientation of the molecule inside the cell cannot be calculated directly from the spectrum. In this case the spectra of proposed crystal structures can be compared with the experimental data. Thus crystal prediction is satisfactory for this purpose, if the observed crystal structure is among the highest-ranking solutions found by the algorithm. These structures can be used as good starting points for refinement procedures based on spectra comparison. An overview about various approaches for this procedures can be found in various textbooks [5, 6].

We have developed a new program, *FlexCryst*, to solve the *ab-initio* problem, but the program can be used also for interpreting powder spectra. In this case the additional information can be exploited to improve the performance and the quality of the results.

The algorithm of *FlexCryst*

Currently, the program can handle the four space groups $P1$, $P\bar{1}$, $P2_1$, and $P2_12_12_1$ and can be easily extended to further space groups. Some adaption to each space group has to take place, since different space groups are determined by different sets of symmetry operations. In the worst case of space group $P\bar{1}$, four independent symmetry elements (three translations and one inversion center) define the crystal structure. The algorithm presupposes that the molecule is rigid. This assumption is justified for pigments (Figure 4), which are often fixed by the enlarged π-systems, and for steroids (Figure 1), which are fixed by the high connectivity of the ring-systems. At the moment the program handles only crystal structures with one molecule per asymmetric unit. Fortunately most crystals observed in nature fulfill this condition. An extension of the algorithm to several molecules per asymmetric unit, increases the search space by six degrees of freedom per molecule. The corresponding variables determine the translation of the molecule and its orientation in the asymmetric unit.

In the following we give a high-level description of the FlexCryst algorithm. The input is a 3-D conformation of an organic molecule.

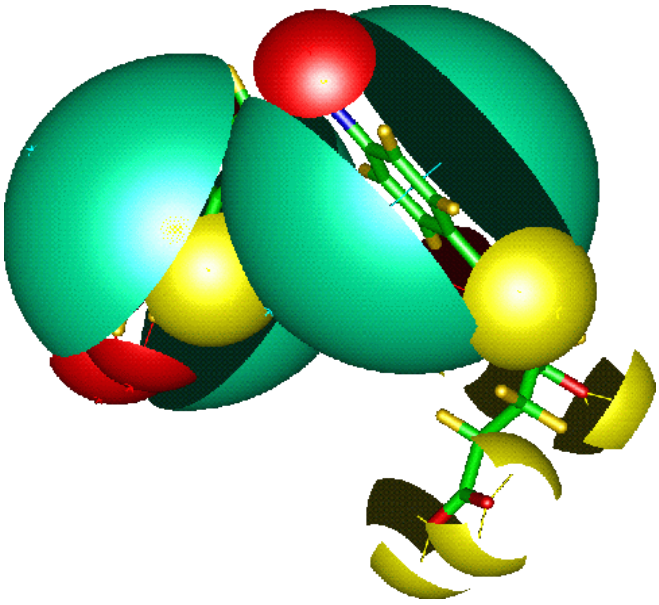


Figure 3 *Result of the analysis of a molecule by Flex. The interaction surfaces are connected by lines to the interaction centers*

Table 1 *Flowchart for the construction of the crystals in the various space groups*

$P1$	$P\bar{1}$	$P2_1$	$P2_12_12_1$
search translation	search inversion	search screw axis	search screw axis
energy constraint	energy constraint	energy constraint	angle constraint
add translation	search translation	search translation	add screw axis
Figure 5			
add translation	energy constraint	angle constraint	
	add translation	add translation	
	add translation	energy constraint	
	add translation	add translation	

Step 1: The molecule is automatically supplemented with hydrogen atoms. For this step we use SYBYL [7].

Step 2: The molecule is searched for active centers by using the program *Flex* [8-11]. Around these centers we calculate potential interaction surfaces. If an interaction between two groups is formed, an interaction center of the first group has to lie on the interaction surfaces of the second group and vice versa (Fgure 2). The result of the analysis of an molecule by *Flex* is shown in Figure 3. The different surfaces are colored depending on their functionality. An complete description and an on-line version is available via Internet http://www.gmd.de/SCAI/alg/reliwe.

Step 3: The interaction centers and the interaction surfaces are discretized (Figure 4), in order to trade off calculation time and accuracy our mesh size is 1 Å. Larger mesh sizes significantly reduce the number of interaction points, which reduces the runtime in the subsequent steps. At the same time, the accuracy of the prediction deteriorates.

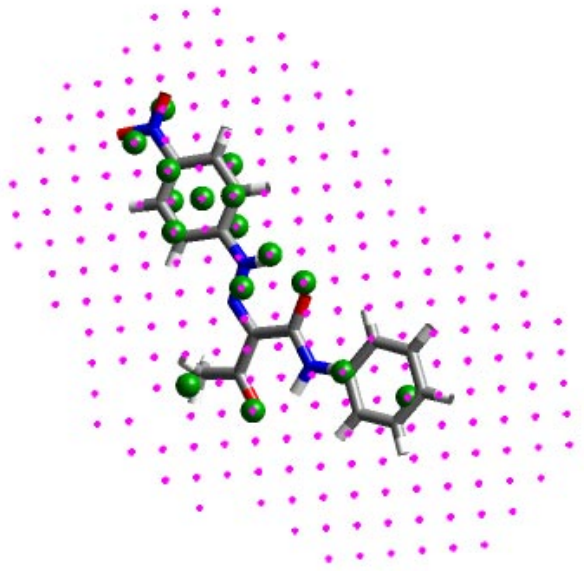


Figure 4 *Molecule with interaction points (purple) and interaction centers (green) after discretization*

Step 4: Possible crystal structures are generated. This step differs for each space group. This step is a small set of computation modules, each of which perform a certain task as searching a certain symmetry operation, adding an additional symmetry element, applying the energy constraint, and applying the angle constraint. The detailed procedure for each space group can be taken from the flowchart in Table 1.

a) search symmetry operation: Proper symmetry operations (including possible unit cell vectors) for crystal structures are determined, analyzing interaction centers and interaction points found in step 1. Currently this step is implemented for the translation, the inversion center, and the twofold screw-axis. It is very crucial for the velocity of the program. Solving the following equations gives proper symmetry elements without scanning. Each symmetry element has to map one or more points **p** onto interaction centers **c**. Each symmetry element can be described by a rotation **W** and a translation **w**. In general our condition can be written as:

$$\begin{pmatrix} W_{xx} & W_{xy} & W_{xz} \\ W_{yx} & W_{yy} & W_{yz} \\ W_{zx} & W_{zy} & W_{zz} \end{pmatrix} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + \begin{pmatrix} w_x \\ w_y \\ w_z \end{pmatrix} \approx \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix} \quad (1)$$

In the case of a pure translation, the rotational part reduces to the unit matrix, and the formula simplifies to:

$$\begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + \begin{pmatrix} w_x \\ w_y \\ w_z \end{pmatrix} \approx \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix} \quad (2)$$

The translations calculated by this equation proper unit cell vectors. Selecting three of them gives a crystal structure of space group $P1$.

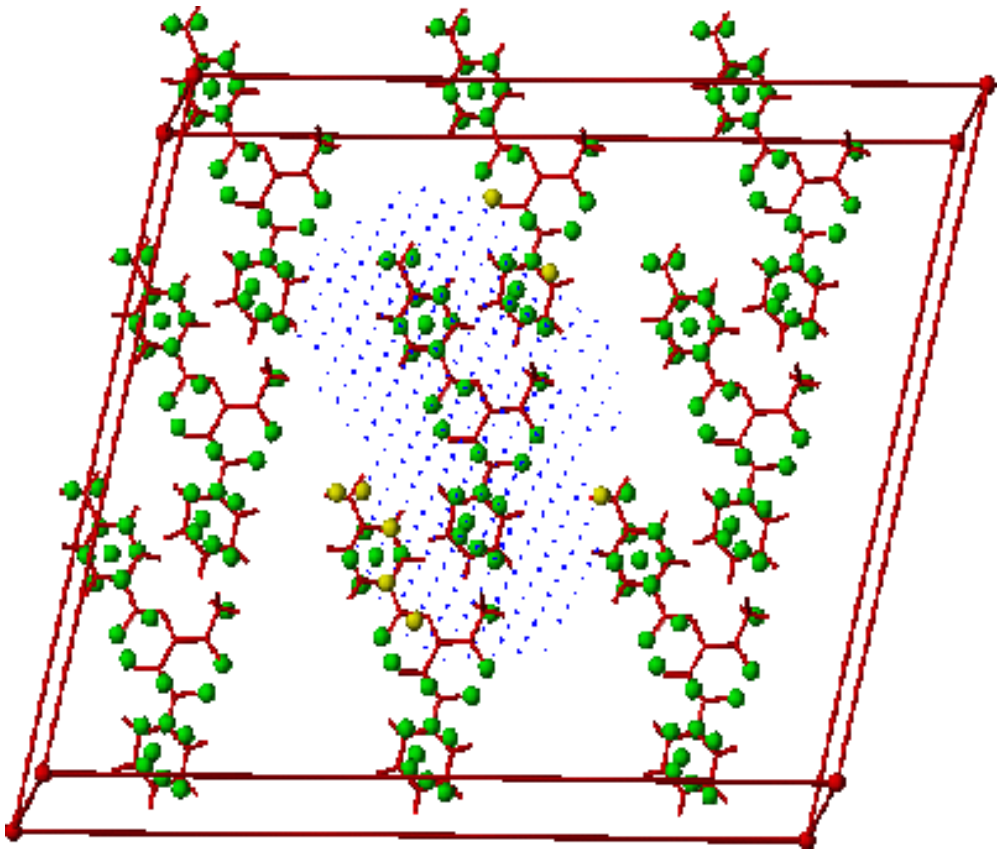


Figure 5 *A plane defined by two translation vectors. Interaction centers are depicted by green spheres. Those centers, that form interactions by being located on interaction points (blue points), are colored yellow*

The inversion has three free parameters. To determine proper inversions the following equation has to be solved.

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + \begin{pmatrix} w_x \\ w_y \\ w_z \end{pmatrix} \approx \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix} \tag{3}$$

The two-fold screw-axis M_{rot21} has five degrees of freedom. To find proper axis for the crystal structure two pairs of centers and points have to be considered simultaneously ($i \in \{1,2\}$). In general a rotation about the axis **l** is described by the equation

$$\begin{pmatrix} 2l_x^2 - 1 & 2l_x l_y & 2l_x l_z \\ 2l_x l_y & 2l_y^2 - 1 & 2l_y l_z \\ 2l_x l_z & 2l_y l_z & 2l_z^2 - 1 \end{pmatrix} \begin{pmatrix} p_{ix} \\ p_{iy} \\ p_{iz} \end{pmatrix} + \begin{pmatrix} w_x \\ w_y \\ w_z \end{pmatrix} = \begin{pmatrix} c_{ix} \\ c_{iy} \\ c_{iz} \end{pmatrix} \tag{4}$$

The symmetry operation has to be an unitary transformation, which requires det(**W**)=1. For the solution of the equation we get:

$$\mathbf{l} = \frac{\mathbf{p}_1 - \mathbf{p}_2 + \mathbf{c}_1 - \mathbf{c}_2}{|\mathbf{p}_1 - \mathbf{p}_2 + \mathbf{c}_1 - \mathbf{c}_2|} \tag{5}$$

The translation expressed gives:

$$\mathbf{w} = \mathbf{c}_1 - \mathbf{W}\mathbf{p}_1 \tag{6}$$

The condition that the transformation is unitary can be rewritten as:

$$|\mathbf{p}_1 - \mathbf{p}_2| = |\mathbf{c}_1 - \mathbf{c}_2| \tag{7}$$

b) add symmetry operation: This module adds an additional symmetry element to a structure, that is already partially defined by a number of symmetry elements. In this way the crystals are constructed step by step. This procedure is very similar to the molecular nuclei concept applied in PROMET [12]. In addition to the symmetry operations screw-axis, inversion centers, and glide planes (in work), we consider translations as symmetry operations. Selecting a first symmetry element creates a dimer. If this symmetry element is important for the crystal structure the created dimer will be energetically favorable. Thus we have to retain only a few numbers of dimers for further processing. The importance of a symmetry element for the different space groups can be derived statistically by considering the distribution of molecular centers in the unit cells [13]. Adding a second symmetry element, creates a tetramer and so on. The number of symmetry elements necessary to define a crystal uniquely depends on the space group. In the case of $P2_12_12_1$ just two symmetry elements, two two-fold screw-axis 2_{1a} and 2_{1b}, define the crystal structure uniquely. The third screw-axis 2_{1c} is the product of the other screw-axes.

$$2_{1c} = 2_{1b} \otimes 2_{1a} \tag{8}$$

The translations t are the square of the corresponding axis.

$$t_a = 2_{1a} \otimes 2_{1a} \tag{9}$$

An other example is the space group $P\bar{1}$. In this space group first a centrosymmetric dimer is constructed, selecting an inversion center calculated from eq. 3. Second a translation is added, creating a tetramer. Afterwards a second translation and a third translation are added. These four elements form a complete basis of the space group. An overview of the selected elements can be seen in Table 2.

c) energy constraint: To apply the energy constraint first the symmetry elements are applied to the molecule mapping it onto images. Then the interaction energies between the images and the reference molecule are evaluated and summed up. The structures are sorted according to their energy and only the highest-ranking structures are retained.

For scoring structures we use the widespread atom-pair energy approach. The interaction energy $E_{dimer}(I,J)$ between two molecules *I* and *J* is assumed to be the sum of atom-pair energies $E_{atom}(r,i,j)$.

$$E_{dimer}(I,J) = \sum_{i=1}^{n_I} \sum_{j=1}^{n_J} E_{atom}(r,i,j) \tag{10}$$

In contrast to most other force fields, the atom-pair potentials used were derived by analyzing known crystal structures. From these, we obtain probabilities for the contacts between the different atoms. The energy function is derived from these probabilities by the inverse Boltzmann equation. Applying the inverse Boltzmann equation, the potential energy between two interacting atoms A_{iI} and A_{jJ} of different molecules *I* and *J* (only intermolecular interactions are of interest here) can be written as:

$$E_{atom}(r_0,i,I,j,J) = E_{ij} + N_L kT \log \lim_{r_\infty \to \infty} \frac{P_{ij}(r_\infty) r_0^2}{P_{ij}(r_0) r_\infty^2} \tag{11}$$

with

$$r_0 = \left| \vec{r}(A_{iI}) - \vec{r}(A_{jJ}) \right| \tag{12}$$

$P_{ij}(r_0)$ is the probability that the shell at distance r_0 around an atom of type *i* contains an atom of type *j* and vice versa. $P_{ij}(r_\infty)$ is the probability of finding two atoms independently of each other, as in the case of an infinite distance between the two atoms. This probability can also be expressed by the average densities ρ_i and ρ_j of the atom types in the crystals.

$$\lim_{r_\infty \to \infty} \frac{P(r_\infty)}{r_\infty^2} \propto \rho_i * \rho_j \tag{13}$$

We estimated the value of the integration constant E_{ij} and the decoupled probability $P_{ij}(r_\infty)$ by the following procedure. We statistically derive the pair potential function with undetermined shift E_{ij}, applying equation (11) to the atom-pair correlation function. In order to have enough data to evaluate the atom-pair correlation function we used the Cambridge Structure Database [14]. We parameterized the most relevant

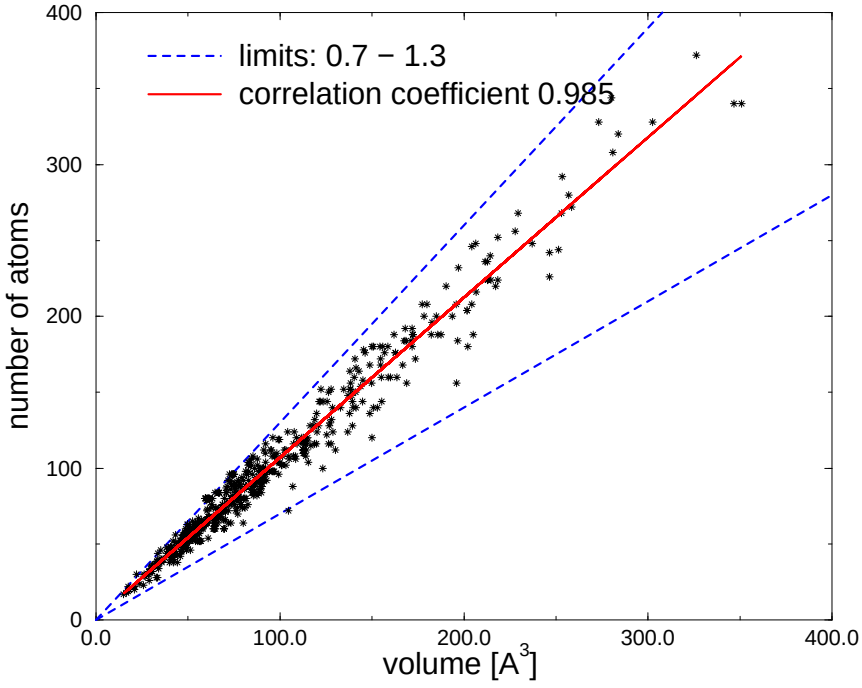


Figure 6 *Correlation between number of atoms and volume in one unit cell*

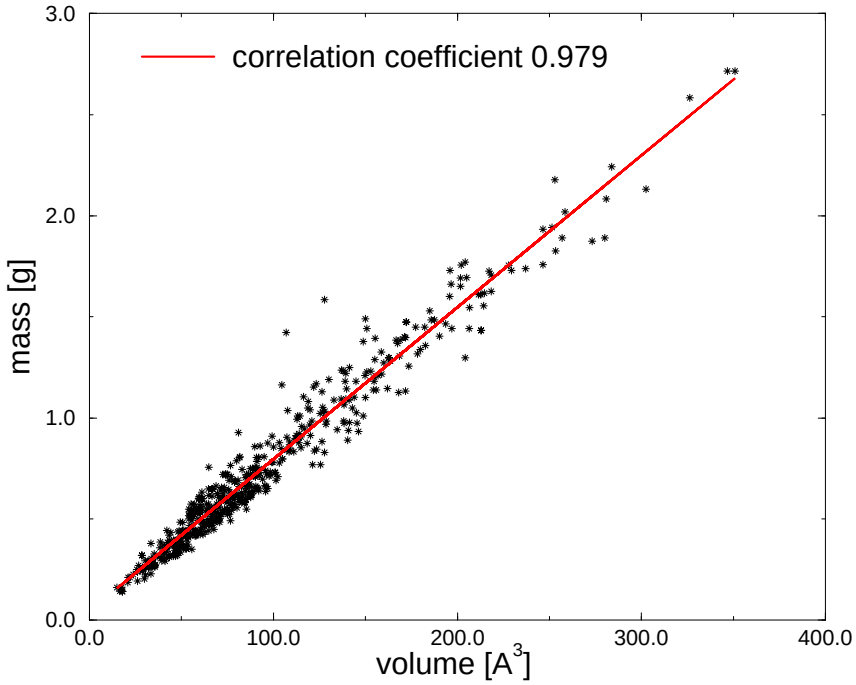


Figure 7 *Correlation between mass and volume in one unit cell*

interactions, and disregarded the contributions of other interactions. An extension to other chemical elements by providing the additional pair correlation functions of these elements is straightforward. The only limitation is the sparsity of available data for several interaction pairs. For each interaction, we evaluated the alphabetically first 1000 different crystals containing the corresponding interaction. This number of structures is sufficiently large for the calibration, as can be argued from the fact that the pair potential functions become almost constant for distances above 4.0 Å. This is to be expected for decoupled probabilities. For this reason, we replace $P(r_\infty)$ by the value of P(4.0 Å) and disregard energy contributions for atom pairs with larger distances than 4.0 Å. To determine E_{ij}, we made use of the fact, that the volume of predicted crystals depends on E_{ij}. For increasing E_{ij} the volume of the predicted crystals increases, as well. This is caused by the mostly monotonically declining pair energy functions in the range of the van-der-Waals contacts. Calibrating an average shift E_{ij} for all pair interactions such that the predicted and experimental volumes of crystals considered are equal, gives us a reasonable value for E_{ij}. For our training set we get a value of -0.68 kcal/mole. Replacing E_{ij} and $P(r_\infty)$, the inverse Boltzmann equation can be rewritten as shown in equation 14.

The cutoff of the energy function at 4 Å introduces an error to our scoring function. This error has to be balanced against the discretization. Due to the discretization all unit cell vectors and origins are located on grid points. The derivation of the atom-pair function (eq. 14) is described in detail in a previous publication [15]. This deductive approach has been introduced first for protein structure prediction [16 – 18] and, later, has been theoretically justified [19]. The atom-pair energies for the distances r are tabulated to avoid time-consuming recalculation.

d) angle constraint: If the space group is not triclinic the angles between the axis are not arbitrary. They have to be 90° or 60°.

Step 5: The atom density constraint is applied to the crystal structure. This restraint requires the density of atoms per Å^3 to be in the range from 0.7 to 1.3. It is well known that the mass density of organic molecule crystals varies only little. We found, that using the number of atoms per unit volume (Figure 6) rather than the mass (Figure7) leads to an even higher correlation between density and volume. The figures contain the complete set of structures (roughly 500), which were used for our calculations. These structures are statistically extracted form the CSD database.

Step 6: The crystal structures are sorted according to our scoring function.

Step 7: The energy constraint is applied to the crystal structures. The 2000 highest-ranking structures are retained.

$$E_{atom}(r_0,i,I,j,I)=\begin{cases}-0.68\text{kcal}\cdot\text{mol}^{-1}+N_L kT\log\dfrac{p(4.0\text{Å})r_0^2}{P(r_0)4.0\text{Å}^2} & \text{if } r_0\le 4\text{Å}\\ 0\ \text{kcal}\cdot\text{mol}^{-1} & \text{if } r_0<4\text{Å}\end{cases} \qquad (14)$$

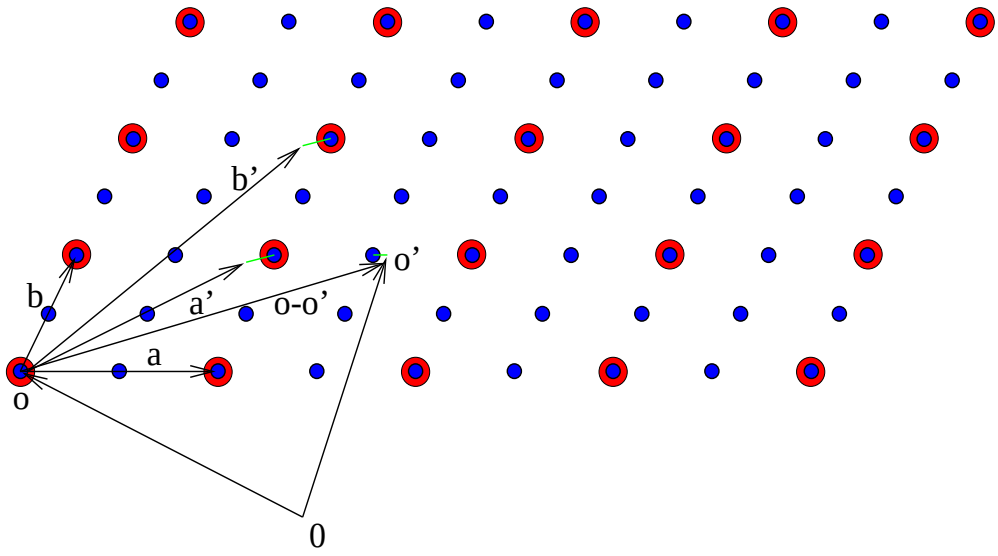


Figure 8 *Our proposed similarity index is the distance (green) of the unit cell vectors a' and b' to the nearest grid point P(B) (red) and the distance of the origins o-o' to the nearest point of the superset S (blue)*

Step 8: The structures are clustered. All structures with a similarity index (see step 9) s=0 are grouped to one cluster. For each of the resulting clusters only the highest ranking structure is retained. All other structures are screened out. These structures are physically identical, but might be different in the choice of the unit cell. Increasing the value of s reduces the number of clusters and improves the qualitative results (see *Qualitative result*). But sometimes even the experimental structure is shredded and the quantitative results (see *Quantitative results*) deteriorate. The main reason are the inadequate positioned hydrogens by our automatic supplementing procedure or/and by the experimental difficulty to determine the position of hydrogens exact.

Step 9: The crystal structures are compared with the experimental structure. Comparison of crystal structures is difficult, because an infinite number of representations for the unit cell is possible for each crystal structure. Various approaches to checking the similarity of two cells have been published. Some of them are based on the comparison of the simulated spectra [20]. In others first the two unit cells are normalized [21] and the square deviation between the atoms in the two unit cells is calculated [22, 23]. For our purposes, we propose a third method that exploits the fact, that we are always dealing with the same molecule, which is rigid and fixed in space. First the translation vectors are compared, and next the origins of the cell are compared. The similarity of the translation vectors is checked in the same way for all space groups. Assuming one base **B** defined by three vectors $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$

$$\mathbf{B} = \left(\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3}\right) \tag{15}$$

the three vectors $\mathbf{b'}_1$, $\mathbf{b'}_2$, and $\mathbf{b'}_3$ of the second cell are expressed as linear combinations $\mathbf{t}_i$ of the vectors of the first base **B**.

$$\mathbf{b}'_i = \mathbf{B}\mathbf{t}_i \tag{16}$$

The distance $\mathbf{r}_i$ between the vectors $\mathbf{b'}_i$ and the nearest grid point P_i of the grid $\{P(\mathbf{B})\}$ defined by **B** is given by

$$\mathbf{r}_i = \left|\mathbf{B}\left(\left[\mathbf{t}'_i + 0.5\right] - \mathbf{t}'_i\right)\right| \tag{17}$$

To compare the origins we have to distinguish between different space groups. The grid points $\{P(\mathbf{B})\}$ defined by the translation vectors are always a subset of the grid $\{S\}$ spanned by the possible origins,

$$\left\{P(\mathbf{B})\right\} \subseteq \left\{\mathbf{S}\right\} \tag{18}$$

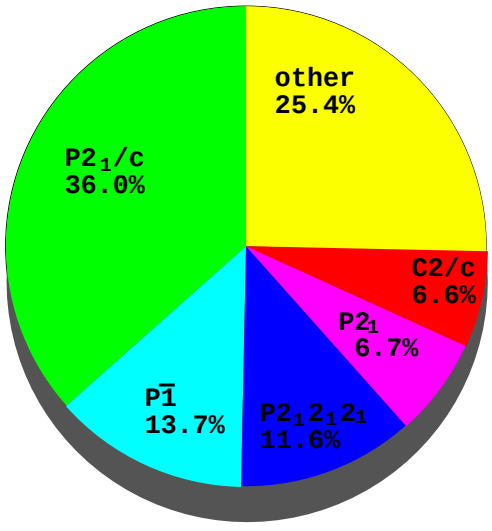


Figure 9 *Distribution of crystal structures in the various space groups*

Table 2 *The results for the different space groups*

space group	free variables	construction	tests	hits	percentage	time [min]
$P1$	9	*t,t,t*	129	123	95%	1.13
$P2_12_12_1$	9	$2_1,2_1$	98	94	96%	34.5
$P2_1$	10	2_1,*t,t*	100	73	73%	51.8
$P\bar{1}$	12	*i,t,t,t*	95	66	69%	12.7

e.g. in space group *P*1 the crystal structure is not influenced at all by the choice of the origin, in space group $P\bar{1}$ and $P2_12_12_1$ the basis vectors of the grid {**S**} are just half of the translation vectors **b**. In the same way as before the distance of the origin to the next possible grid point of the superset is calculated.

$$\mathbf{r}_{origin} = \left|\mathbf{S}\left(\left[\mathbf{t}'_{origin} - \mathbf{t}_{origin} + 0.5\right] - \mathbf{t}'_{origin} + \mathbf{t}_{origin}\right)\right| \qquad (19)$$

As similarity index we choose the maximum of the four distances.

$$s = \max\left\{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_{origin}\right\} \qquad (20)$$

For our calculations we did choose s < 1.8 Å.

In Figure 8 we show a two dimensional projection of the space group $P\bar{1}$. The grid P(B) is defined by the experimentally known vectors a and b. Following eq. 17 first we calculate the distance of the vectors **a'** and **b'** of the simulated structure to the nearest grid point P(B). In the space group $P\bar{1}$ the origin has to be an inversion center (blue). All inversion centers define a supergrid S (blue) spanned up by the vectors $\frac{\mathbf{a}'}{2}$ and $\frac{\mathbf{b}'}{2}$. The distance of the difference between the origins to the closest point of the supergrid yields our measure for similarity.

Results

For validation we extracted about 100 experimental structures from the Cambridge Structure Database [14] for each implemented space group. We selected the alphabetically first organic crystals containing only the elements H, C, N, O, F, P, S, and Cl. The crystals were required to contain only one molecule per asymmetric unit. The molecular data were input to the program *FlexCryst*. The output of the program, 2000 crystal structures for each of the 100 molecules, was compared with the experimental structure stored in the CSD, as well.

We were interested in the quantitative and qualitative aspect of our results.

Quantitative results

We first investigated whether the experimental structure was among the proposed crystal structures at all. The results are presented in Table 2. The first column shows the space group. The space group P1 is the simplest, and the one most extensively used for further developments of the program. The other three space groups are often found in nature and are most important for practical applications. Fortunately 75% of all observed crystals are described by just five space groups as can be seen in Figure 9 [24]. This suggests to restrict future extensions of the program to a limited number of space groups.

The second column contains the number of free variables that uniquely determine the unit cell. This number ranges from nine (*P*1) to twelve ($P\bar{1}$), if we consider only one rigid molecule in the asymmetric unit. The third column contains the ordered symmetry elements, which were used to construct the crystal step by step. Translations are abbreviated with t, inversion centers with *i*, and two-fold screw-axes with 2_1. In the column "tests" we report the number of structures extracted from the CSD. In case of *P*1 we used all available

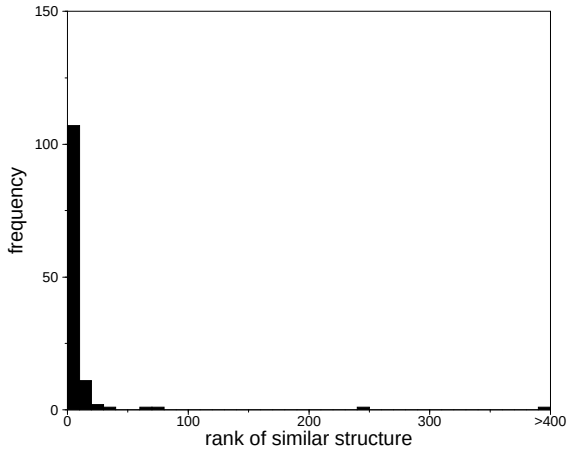


Figure 10 *Rank of the experimental structure among the proposed structures for space group* P1

Table 3 *The results for the different space groups*

rank	refcode	#clusters	#atoms	energy	time
1	*ADGSMF	360	55	-190.24	45
1	*ADGSMH	1235	55	-199.04	74
1	*BADVAD10	50	64	-163.64	45
2	BAKHOK	1510	79	-320.42	242
8	BDORLA10	1338	29	-113.36	42
1	*BEKHUU	747	74	-249.32	81
1	BERVEZ	1381	78	-226.38	141
1	BETJEP	1049	43	-131.08	50
25	BIPPEV	1624	17	-87.04	40
1	BIXHOF	860	71	-219.56	108
5	*BOTSAE	1373	36	-146.22	64
1	BXCPAF	120	48	-174.32	45
13	*CEGLCA	1629	32	-205.46	41
13	CEGLCA01	1590	32	-196.80	43
–	CERPAQ	50	40	0.00	18
37	CETROI	1349	32	-121.34	41
2	CETROI01	1423	32	-142.00	40
1	CIFYOF	1397	89	-322.30	158
1	CIFYOF10	1434	89	-323.66	163
11	CILWOJ	43	46	-193.56	20
2	*CIYRIL01	1343	51	-183.14	60
1	COMCIQ	825	64	-223.10	92
1	COTCIX	1746	52	-203.82	90
1	CUVFOO	1419	24	-133.38	44
1	DAKSAJ	830	53	-179.58	56
1	DARNUF	1160	57	-148.26	93
1	*DEBLOL	1008	90	-338.80	142
1	*DERCIM	202	56	-222.48	31
1	*DIGOXN	306	119	-404.98	176
1	DIGOXN10	461	119	-445.86	178
1	DIWXIQ	1796	61	-236.64	105
1	DOHHIR	1822	52	-337.60	72
1	DOZMIO	1223	44	-232.74	48
4	DUMCET	390	71	-235.32	53
1	EACJEX	1288	92	-377.52	190
10	*ECPRPR01	1194	45	-166.96	54
1	FADGEW	1693	79	-332.70	189
9	*FAKGAZ01	831	47	-247.18	39
3	*FALKAE	840	56	-271.68	53
1	FAMDUS	1387	39	-210.82	69
10	FATXUT	1188	39	-196.28	36
1	FAVSUQ	639	53	-193.60	63
7	*FEPZOP	1172	102	-356.78	164
2	FETWOQ	1541	30	-112.26	61
1	*FEXCOA	631	96	-295.72	110
2	FITVOT	1515	80	-314.66	240
1	FIYJIG	1118	60	-226.28	72
1	*FOMANN	1646	64	-248.38	124

Table 4 *The results for the different space groups*

rank	refcode	#clusters	#atoms	energy	time
1	FUNVUF	1322	94	-310.20	252
3	FUPVAN	952	61	-177.90	62
–	FURCOU	536	20	0.00	20
2	FUXBIJ	1010	40	-175.86	36
1	FUXBIJ01	832	40	-178.04	34
1	GEYMEC	1317	82	-229.66	157
1	GIPJEU	1255	75	-269.20	103
9	*GOJHIW	925	45	-187.44	41
3	*HAGFAW	376	52	-178.98	35
4	*HCARDO	1571	63	-243.20	125
1	HCARDO01	1549	63	-260.80	135
1	*HOLOTM	1218	84	-317.26	128
1	HPICRB	1879	56	-208.18	84
1	HTENTX10	1064	62	-215.94	75
1	*JANDUX	724	66	-193.44	70
1	JECYIZ	1495	40	-182.88	55
1	JIHREX	586	38	-170.88	31
–	JIJXEF	413	59	0.00	43
3	*JIPBIT	503	43	-139.32	42
2	*JOVZAV	1820	83	-265.92	228
1	JUFTUZ	1240	35	-180.44	40
1	*KANDUY	1306	77	-208.84	134
1	KANTOI	582	62	-351.26	48
4	*KEGBAZ	1532	39	-149.12	72
1	*KERSIJ	748	110	-363.96	211
1	*KIJCAH	515	74	-285.20	71
1	KITLUU	1550	83	-218.86	151
1	KOCHIT	356	45	-139.10	35
2	KOHNAW	1360	87	-382.02	138
16	KOPROW	1858	52	-201.00	82
1130	*LAWKUP	1220	22	-79.92	43
78	LCDMPP01	1426	20	-107.32	38
4	LCDMPP10	1314	20	-115.94	38
1	LEDNUD	918	60	-163.06	65
1	*LEKVIG	186	115	-338.58	209
1	*LEMZAE	1666	56	-226.70	85
1	LETBOB	1396	54	-231.20	104
1	LYSDOL	483	59	-219.30	49
1	MAMNAC	1270	63	-277.70	71
1	NALCYS02	1043	19	-107.98	33
1	OACGAP	638	85	-313.30	123
1	*OHWTHN	1218	73	-257.68	103
1	OMAPBD	1582	48	-251.40	95
1	PAJSOI	257	66	-193.08	53
1	PATCUI	918	65	-252.84	70
1	PATPYS	769	49	-172.92	63
4	*PEVLOR	1680	91	-378.20	280
17	*PICSEZ	1904	67	-334.04	110

Table 5 *The results for the different space groups*

rank	refcode	#clusters	#atoms	energy	time
16	PIKYIR	851	44	-159.06	48
7	*PMNTBZ	1011	29	-160.98	47
–	*PROGLE20	151	57	0.00	25
2	RPPYPY20	1396	35	-165.88	51
–	*SESHUT11	136	26	0.00	18
15	SEZLUE	1780	37	-130.50	56
27	TEOXDE01	910	22	-146.68	39
–	*THPGFA	109	59	0.00	24
12	VARHUR	1851	63	-199.24	116
1	VARWUG	1807	58	-249.98	90
1	*VEGJOG	1363	64	-218.80	108
1	VEKZAM	828	90	-332.72	117
5	VITREV	1541	31	-131.50	54
1	*VOBHEZ	1170	49	-164.72	63
11	VOFFAX	1589	41	-143.86	56
1	*VOXXUB	1495	30	-202.90	40
1	VOYVEK	1687	55	-179.26	94
240	WATCID	1438	44	-201.46	73
1	WICVUZ	1384	61	-230.66	71
1	WIKSEO	902	86	-269.04	110
1	WINWEV	1691	51	-174.60	76
1	YABVUS	339	52	-184.48	34
1	*YAMBET	1081	52	-214.68	63
60	YEBGIV	309	23	-65.50	30
1	*YEHRIM	919	41	-148.98	49
1	*YIJBUO	1510	51	-170.28	74
4	YIPPAO	773	73	-233.72	79
1	YIPWAV	1739	52	-237.70	109
1	YOGVOF	1806	65	-260.90	103
1	*YOKGIO	526	73	-198.00	75
2	YUYHAB	1417	18	-105.14	32
1	ZAYWIJ	524	54	-188.18	53

structures, while for the other groups we limited our sets to 100 structures. Some of these structures we found dubious [a] and dropped manually before the validation. The column "hits" gives the number of experimental structures found among the proposed crystals. For $P1$ only six structures are not reproduced. One of these structures CERPAQ [25] has been redetermined [26] and assumed to be of space group $P\bar{1}$. Three other failures, PROGLE20, SESHUT11, and THPGFA, are caused by an incorrect automatic addition of hydrogens. Because it is not always possible to determine experimentally the position of the hydrogens, they are sometimes omitted in the CSD. These structures were supplemented automatically with help of the program SYBYL [7]. At this step the crystal structure is not included and, therefore, sometimes the orientation of the hydrogens is essentially random. During the construction process of the crystals these supplemented H-atoms cause bad contacts. The other two structures (JIJXEF, FURCOU) have very weak interactions for one unit cell vector. The corresponding dimer is very low in energy and the translation vector falls below the energy constraint. The next column shows the same results in terms of percentages. The last column shows the average runtime per molecule for the particular space group. This time varies from 2 minutes up to 1 hour on a SUN™ULTRA™1 workstation. The runtime rises significantly with the number of free variables.

The crystals of space group $P2_1$ produces a low rate of hits compared with the other groups. This results from the construction of the crystals. The actual implementation requires the screw-axis to be a leading symmetry element of

[a] Wrong crystal structures are reported to CSD and corrected immediately

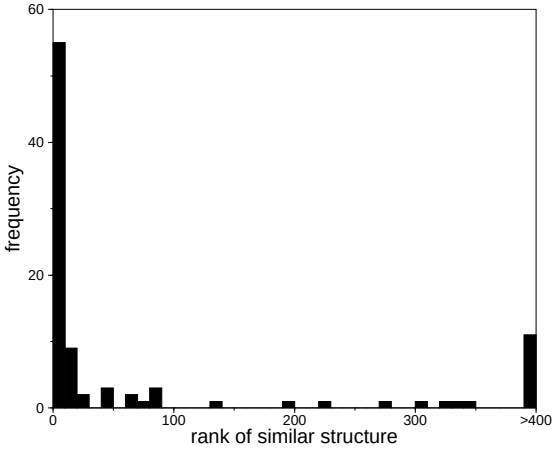


Figure 11 *Rank of the experimental structure among the proposed structures for space group* $P2_12_12_1$

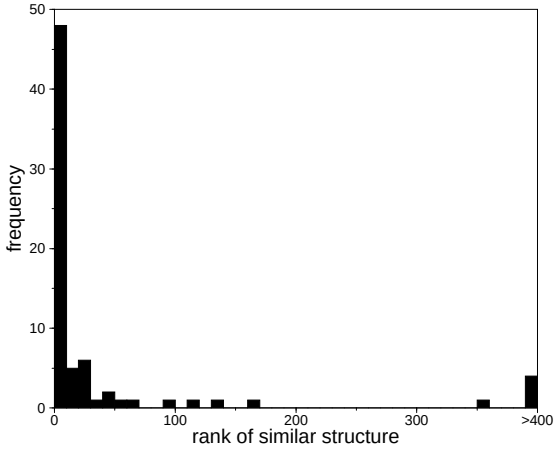


Figure 12 *Rank of the experimental structure among the proposed structures for space group* $P2_1$

the crystal but, for some crystals, the most important packing pattern incorporates a translation and our the procedure fails. To find such crystals our construction procedure has to be extended. The sequence of first determining the screw-axis and then the translation has to be reversed.

Qualitative results

Secondly we looked for the rank of the experimental structure among our ranked list of structures. As expected, the best results were achieved for space group $P1$ (Figure 10). Most of experimental structures (107 of 125) are found among the first ten structures. A complete list of the ranking is given in the Table 5. The structures with high ranks are caused by different reasons.

• Hydrogens are supplemented unfavorable (LAWKUP, PICSEZ, CEGLCA). All structures supplemented with hydrogens are marked by an asterisk.

• The structure has been misidentified (CILWOJ [27]).

• The first structure determination was imprecise. In these cases the redetermined structures have much lower ranks (CETROI 37 $\rightarrow$ 2, FUXBIJ 2 $\rightarrow$ 1, HCARDO 4 $\rightarrow$ 1, and LCDMPP 78 $\rightarrow$ 4), even if the difference between the two structures is very small. In the case of CETROI the difference between the length of unit cell vectors is less than 0.1 Å and the molecule coordinates are nearly identically (RMS = 0.15 Å).

• For some structures no obvious reason can be detected. We recalculated the structures with the TRIPOS force field [7]. The force field gives for the molecule huge energies, so we suppose bad contacts in these structures (e.g. BIPPEV +84 kJ·mol^{-1}).

With increasing number of free variables to be determined this pattern becomes more and more diffuse. For the space group $P2_12_12_1$ most of the structures (55) are still found among the ten highest ranking candidates, but a few of them (11) occupy a rank 400 or greater (Figure 11). The distribution for the space group $P2_1$ (Figure 12) is similar to that for $P2_12_12_1$.

The most diffuse pattern was obtained for space group $P\bar{1}$ (Figure 13). Many of the crystals (33) are still found among the ten highest-ranking candidates, but a remarkable number (5) has ranks above 400. This reflects the large number of degrees of freedom.

Conclusions

We have presented a discrete algorithm that detects the experimentally observed crystal structure of organic molecules among the computed candidates. Almost always, the experimental structure is found for the simple case of $P1$ with one molecule in the unit cell, and for the space groups $P\bar{1}$ with two and $P2_12_12_1$ with four molecules in the unit cell. For the space group $P2_1$ with two molecules per unit cell the structure is detected in a large percentage of the cases and the percentage might increase by further as the program develops. The program is very fast. Three ingredients are essential for the efficiency of our method:

• Analyzing the intermolecular interaction as a preprocessing step. This step makes scanning for unit cell vectors superfluous. All structures builded up exploiting this information, finishes in structures with contacts between molecules. The time consuming evaluation of the energy for structure refinement in other methods can be skipped. Only the final fine-tuning by quantum methods [28] or sophisticated force fields [29] remains.

• Using a discrete configuration space. This allow us to balance performance versus accuracy.

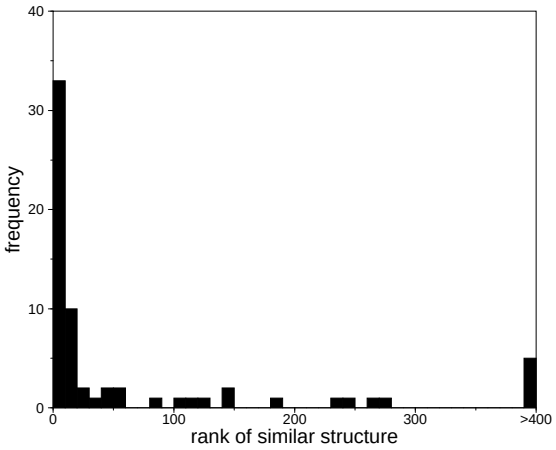


Figure 13 *Rank of the experimental structure among the proposed structures for space group* $P\bar{1}$

• Statistical potentials. The atom pair-functions are tabulated. Reducing the energy calculation to a simple table-lookup speeds up the program significantly. In addition the potentials gives a high flexibility to the program. The parameters can be easily trained for a specific group of compounds, e.g. pigments which contains mostly aromatic rings. The program compares very well in accuracy and performance to other published crystal structure predictions. Most other methods can be divided into three steps, crystal structure generation, crystal structure refinement with MD, and fine-tuning [30, 31]. The proposed algorithm unifies the first two steps. Only a few number of structures has to be considered for fine-tuning. (Structures with ranks above 400 we consider as failure caused by our procedure or worse positioned hydrogen atoms.)

Acknowledgment The authors would like to thank Dr. P. Erk (BASF) and Dr. S. Motherwell (CSD) for fruitful discussions and Dr. H. Slot (University of Nijmegen) for kindly assisting us in using the Cambridge Structure Database. Furthermore we are grateful to P. Lauwers, C. Lemmen, B. Kramer, C. Oligschleger, M. Rarey, and S. Wefing for helpful comments on the paper. This research was performed as part of GMD's contribution to HLRZ (High Performance Computing Center).

References

1. Zimmermann, S. *Science* **1997**, *276*, 543.
2. Gavezotti, A. *Accounts Chem. Res.* **1994**, *27*, 309.
3. Wawak, A.L.R.J.; Gibson, K.D.; Scheraga, H. *Proc. Natl. Acad. Sci.* USA **1996**, *93*, 1743.
4. van Eijk, B.; Mooij, W.; Kroon, K. *Acta Cryst.* **1995**, *B51*, 99.
5. Glusker, J.; Trueblood, K. Refinement of the Trial Structure. In *Crystal Structure Analysis a Primer*, 2 ed.; University: Oxford, 1985.
6. Kleber, W. In *Einführung in die Kristallographie*, 16 ed.; Technik: Berlin, 1983.
7. Associates, T. SYBYL. In ; TRIPOS Associates, Inc.: St. Lois, Missouri, USA, 1994.
8. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470.
9. Rarey, M.; Wefing, S.; Lengauer, T. *J. Comp.-Aided Mol. Des.* **1996**, *10*, 41.
10. Rarey, M.; Kramer, B.; Lengauer, T. *J. Comp.-Aided Mol. Design* **1997**, *11*, 369.
11. Lemmen, C.; Lengauer, T. *J. Comp.-Aided Mol. Des.* **1997**, *11*, 357.
12. Gavezotti, A. *J. Am. Chem. Soc.* **1991**, *113*, 4622.
13. Motherwell, W. *Acta Cryst.* **1997**, *B53*, 726.
14. Allen, F.; Kennard, O. *Chem. Des. Autom. News* 1993, *1*, 31.
15. Hofmann, D.; Lengauer, T. *Acta Cryst.* **1997**, *A53*, 225.
16. Sippl, M. *J. Mol. Biol.* **1990**, *213*, 859.
17. Sippl, M. *J. Comp.-Aided Mol. Design* **1993**, *7*, 473.
18. Sun, S. *Protein Science* **1993**, *2*, 762.
19. Gutin, A.; Badretinov, A.; Finkelstein, A. *Mol. Bio.* **1992**, *26*, 94.
20. Karfunkel, H.; Rohde, B.; Leusen, F.; Gdanitz, R.; Rihs, G. *J. Comp. Chem.* **1993**, *14*, 1125.
21. Parthe, E.; Gelato, L. *Acta Cryst.* **1984**, *A40*, 169.
22. Burzlaff, H.; Rothammel, W. *Acta Cryst.* **1992**, *A48*, 483.
23. Dzyabchenko, A. *Acta Cryst.* **1994**, *B50*, 414.
24. Mighell, A.; Himes, V. *Acta Cryst.* **1983**, *A39*, 737.
25. Bocelli, G.; Grenier-Loustalot, M.F. *Acta Cryst.* **1984**, *C40*, 679.
26. Bocelli, G.; Grenier-Loustalot, M.F. *Acta Cryst.* **1986**, *C42*, 127.
27. Bocelli, G.; Grenier-Loustalot, M.F. *Acta Cryst.* **1984**, *C40*, 1391.
28. Car, R.; Parinello, M. *Phys.Rev.Lett.* **1985**, *55*, 2471.
29. Willock, D.; Price, S.; Leslie, M.; Catlow, C. *J. Comp. Chem.* **1995**, *16*, 628.
30. Chaka, A.; Zaniewski, R.; Youngs, W.; Tessier, C.; Klopman, G. *Acta Cryst.* **1996**, *B52*, 165.
31. Shoda, T.; Yamahara, K.; Okazaki, K.; Williams, D.E. *J. Mol. Struc. (Theochem)* **1995**, *333*, 267.