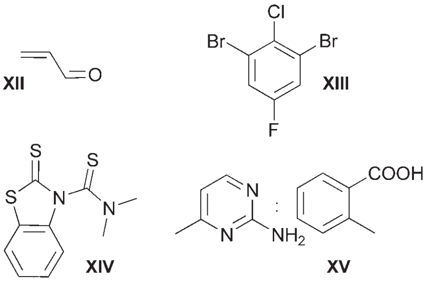
DOI: 10.1002/anie.200704247

Crystal Structure Prediction

A Major Advance in Crystal Structure Prediction[**]

Marcus A. Neumann, Frank J. J. Leusen,[*] and John Kendrick

The goal of predicting the crystal structure of an organic molecule from its molecular structure alone is of considerable industrial importance.[1] The task is complicated, owing to the number of degrees of freedom to be explored, the complexities of intermolecular and intramolecular forces, and the difficulty in choosing a suitable computational criterion for identifying those crystal structures favored by nature. The difficulty of the task is clearly demonstrated by the regular "Crystal Structure Prediction Blind Test", which is organized by the Cambridge Crystallographic Data Centre. A Blind Test has taken place in 1999,[2] 2001,[3] 2004,[4] and recently in 2007. Participants are provided with three or four molecular structures and invited to predict, within six months, up to three crystal structures which they think each compound will adopt. The experimental crystal structures have been determined but are not available until after the participants have supplied their predictions. The limited number of successful predictions reported in the previous Blind Tests reveals just how difficult crystal structure prediction (CSP) can be. "Success" in this context means that the observed, experimental crystal structure is found among the three submitted predictions of a participant. All of the previous successful predictions were based on force-field methods, in which the intermolecular and intramolecular forces are represented by analytical functions. Herein, the successful application of a new CSP approach to all four compounds of the 2007 Blind Test is presented. The four compounds chosen for the 2007 Blind Test are shown in Scheme 1.



Scheme 1. Molecular structures of the 2007 Blind Test compounds. The Roman numerals refer to the numbering scheme used in the Blind Tests. Compound **XV** is a cocrystal.

In essence, two problems need to be addressed in CSP. First, there is the physical problem of accurately describing the relative stabilities of all possible crystal packing alternatives. Second, there is the mathematical problem of finding all low-lying minima on the lattice energy hypersurface, a function with many variables, including the unit cell dimensions, the space group, the number of molecules in the asymmetric unit, their conformation(s), and their packing in the crystal lattice. The high number of degrees of freedom results in a complex global optimization problem, for which several solution strategies have been put forward.[5–11]

The central part of the approach used herein is a hybrid method, developed by one of the authors (M.A.N.), for the calculation of lattice energies[12] that combines density functional theory (DFT) simulations using the Vienna Ab initio Simulation Package (VASP) program[13–15] with an empirical van der Waals (vdW) correction expressed in terms of a sum over isotropic atom–atom pair potentials.[12] As solid-state DFT calculations are time-consuming and cannot be used directly for crystal structure generation, the hybrid method is used for the generation of reference data, from which a tailor-made force field (TMFF) is derived for every molecule under consideration. The force field involves atomic point charges calculated from bond increments, isotropic vdW potentials, and covalent bond stretch, angle bend, torsion, and inversion terms.[16] Non-equivalent atoms are attributed different force-field atom types to allow for maximum customizability. The reference data include the electrostatic potential around the molecule, as well as energies and forces at, and around, local energy minima of densely packed crystal structures and isolated molecules in large simulation boxes.[16] All force-field parameters, in particular bond increments and vdW constants, can be fitted to the reference data simultaneously. It is important that the TMFF provides a sufficiently faithful representation of the hybrid potential-energy surface, both in terms of structure and energetics. Consistency checks that enable this to be verified during the CSP will be described below.

The TMFF provides lattice energies and forces to a crystal structure generation engine. The version used for the 2007 Blind Test combines a random structure generation mechanism with an efficient lattice-energy minimizer. Molecular flexibility is treated as an integral part of the crystal structure generation process, and all 230 space groups are considered.

---

[*] Dr. F. J. J. Leusen, Dr. J. Kendrick
Institute of Pharmaceutical Innovation
University of Bradford
Bradford, BD7 1DP (UK)
Fax: (+44) 1274-236-155
E-mail: f.j.j.leusen@bradford.ac.uk
Dr. M. A. Neumann
Avant-garde Materials Simulation
30 bis, rue du vieil Abreuvoir
78100 Saint-Germain-en-Laye (France)

[**] The authors thank Dr. G. M. Day for organizing the 2007 Blind Test and for providing the COMPACK comparisons reported in Table 3. The authors also thank the Cambridge Crystallographic Data Centre for hosting the 2007 Blind Test. Dr. Neumann thanks Sanofi-Aventis and AstraZeneca for financial support during the development of the GRACE software package used for this study.

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author.

Structure generation is stopped when all structures in a user-defined energy window above the global minimum have been found several times. Typically, the 50 lowest energy structures are all found at least twice. The predictions reported herein made use of the prior information released to all participants that the experimental structures would have Z' less than or equal to one and that the cocrystal (compound **XV**) would have a stoichiometry of 1:1. For compounds **XII**, **XIII**, and **XIV** this means up to one molecule in the asymmetric unit. For compound **XV**, this means up to one of each molecule in the asymmetric unit.

The most stable structures generated with the TMFF are optimized and ranked with the hybrid method. By comparing the lattice energies obtained with the TMFF to those from the hybrid method, it is possible to work out the offset of the two energy scales and the root mean square deviation (σ) of the energy shift upon optimization with the hybrid method.[12] The energy offset and σ are used to assess the probability that the most stable structure, according to the hybrid method, has already been optimized with the hybrid method. Typically, the optimization of crystal structures with the hybrid method is continued until the probability of having reached the most stable structure is at least 99%.

The majority of calculations were performed on a 48 node Beowulf cluster at the University of Bradford, with each node comprising dual 2.8 GHz Intel Xeon processors and 2 GB of memory. The combined CPU time of all calculations carried out for compounds **XII** to **XV** corresponds to about 4 months on this hardware at full load, or approximately 280000 CPU hours.

Table 1 shows the number of crystal structures that were optimized with the hybrid DFT method and the deviations of their energies in comparison to the TMFFs. Molecules **XII**, **XIII**, and **XIV** reached a high degree of internal consistency. For molecules **XIII** and **XIV**, the main concern was that the hybrid method would prove less accurate than for molecule **XII** because of the number of halogens, in particular bromine, in molecule **XIII** and sulfur in molecule **XIV**. The poor σ value for molecule **XIII** reflects the difficulty the force field has in reproducing the electrostatic potential and the anisotropic repulsion around the halogens. It appears that the force field deals well with the sulfur atoms in molecule **XIV**. For compound **XV**, the presence of two independent molecules in the asymmetric unit means that the number of degrees of freedom roughly doubles, and hence there are a higher number of potential crystal structures in a given energy window.

Table 1: The number of crystal structures optimized with the hybrid method and the root mean square deviation (σ) between the lattice energies from the TMFF and the hybrid DFT method.

	XII	**XIII**	**XIV**	**XV**
no. structures optimized	100	50	32	96
σ [kcal mol^{-1} atom^{-1}]	0.025	0.053	0.024	0.032
no. structures within 2σ of the minimum[a]	2	14	8	>96

[a] For molecules **XII**, **XIII**, and **XIV**, the number refers to unique structures. For compound **XV**, the number refers to all structures.

The ability to estimate the faithfulness of the TMFF through its standard deviation from the hybrid calculation gives the overall methodology robustness. For those cases in which the deviation is large, the likelihood of finding the global, hybrid minimum-energy structure can be increased by either improving the TMFF or by optimizing a larger number of structures with the hybrid method. For compound **XV** it is estimated that the faithfulness of its TMFF and the number of degrees of freedom in the search problem are such that it would have required around 2000 hybrid optimizations to obtain a similar level of confidence in finding the global minimum as for the other compounds. With the 96 optimizations that were performed, there is an estimated probability of about 30% that the optimizations included the global minimum.

Table 2 presents the relative hybrid energies and the ranks according to both energy calculation methods for all structures submitted in the Blind Test. For compounds **XII**, **XIII**, and **XIV**, the first- or second-ranked TMFF structures correspond to the global minimum hybrid structures. For compound **XV**, the 37th TMFF structure corresponds to the global minimum hybrid structure. In the Supporting Information the interested reader will find further information on the top six predicted structures for each compound as well as cif files containing all the structures optimized using the hybrid method.

Table 2: The relative energies of the three lowest-energy crystal structures for each compound according to the hybrid method, their rank with the TMFF, and their space group symmetry.

Compd	Hybrid rank[a]	TMFF rank[b]	Hybrid ΔE [kcal mol^{-1} atom^{-1}]	Space group
XII	1	1, 37, 38, 69	0.000	*Pbca*
	2	7	0.036	*P21/c*
	3	13	0.059	*Pc*
XIII	1	2	0.000	*P21/c*
	2	4	0.027	*P42/n*
	3	3	0.029	*Fdd2*
XIV	1	1, 8	0.000	*P21/c*
	2	9	0.019	*Pbca*
	3	19, 25	0.042	*P21/c*
XV	1	37	0.000	*P21/c*
	2	17	0.015	$P\bar{1}$
	3	24	0.018	*P21/c*

[a] Structures which differ only in the rotation of a methyl group are regarded as equivalent. [b] This column reports all TMFF structures that minimize to the same hybrid structure.

For each compound, the three predicted crystal structures with the lowest hybrid lattice energies were submitted in the Blind Test. Once the experimental structures were disclosed, comparisons were made with the predictions. The results are shown in Table 3. The rank 1 predictions agree with the observed experimental crystal structures for all four compounds, and the deviations between predicted and experimental structures are acceptably small, as can also be seen in Figure 1. Crystallographic details of the experimental and predicted crystal structures are given in Table 4. The results

Table 3: Comparison of experimental and lowest-energy predicted crystal structures.

Compd	Space group		Density[a] [g cm^{-3}]		RMSD[b] [Å]
	Exptl	Prd	Exptl	Prd	
XII	*Pbca*	*Pbca*	1.152	1.129	0.127
XIII	*P*21/*c*	*P*21/*c*	2.528	2.548	0.082
XIV	*P*21/*c*	*P*21/*c*	1.479	1.450	0.130
XV	*P*21/*n*	*P*21/*c*	1.301	1.307	0.075

[a] Calculated using average molecular masses. [b] RMSD is the root mean square difference between the experimental and predicted non-hydrogen atomic positions as calculated by the COMPACK procedure.[17]

obtained by the other participants in this Blind Test will be published elsewhere.

The successful prediction of the crystal structures of all four Blind Test compounds represents an advance in CSP. The results suggest that accurate lattice energies are an appropriate selection criterion in CSP, at least for these four compounds. The calculations presented herein have not included zero-point energies or entropic effects. Only with their consideration will the prediction of polymorphic stability become feasible. It is for this reason that, at the present time, CSP typically aims to provide a small number of possible structures for each compound, among which the experimental polymorph(s) are likely to be found. With this in mind it could be regarded as fortuitous that all of the hybrid rank 1 predictions agreed with the experimental structures. In the case of compound **XV**, the quality of the TMFF and the large number of structures in a small energy window meant that there was an element of luck in finding the experimental structure in the list of structures optimized by the hybrid method. Further validation work on more complex and flexible compounds is required to determine the general reliability and accuracy of this new CSP approach. The bottleneck of the approach is that the hybrid DFT method is computationally demanding and can be used to optimize a limited number of crystal structures only. Hence, a highly accurate force field is required to suggest a small number of

Table 4: Crystallographic information for the experimental and predicted crystal structures for each of the four Blind Test compounds.

Compd[a]	Space group	Unit cell dimensions					
		a [Å]	*b* [Å]	*c* [Å]	α [°]	β [°]	γ [°]
XII exptl	*Pbca*	6.97	9.51	9.75	90.0	90.0	90.0
XII prd 1[b]	*Pbca*	6.97	9.49	9.98	90.0	90.0	90.0
XII prd 2	*P*21/*c*	6.41	7.56	7.73	90.0	114.1	90.0
XII prd 3	*Pc*	3.93	5.26	8.02	90.0	99.5	90.0
XIII exptl	*P*21/*c*	3.89	13.51	14.43	90.0	93.6	90.0
XIII prd 1	*P*21/*c*	3.87	13.46	14.47	90.0	95.0	90.0
XIII prd 2	*P*42/*n*	19.84	19.84	3.85	90.0	90.0	90.0
XIII prd 3	Fdd2	27.47	28.86	3.85	90.0	90.0	90.0
XIV exptl	*P*21/*c*	13.06	9.74	9.34	90.0	105.8	90.0
XIV prd 1[b]	*P*21/*c*	13.24	9.82	9.31	90.0	105.8	90.0
XIV prd 2	*Pbca*	9.77	9.48	25.61	90.0	90.0	90.0
XIV prd 3	*P*21/*c*	13.19	9.81	9.23	90.0	94.7	90.0
XV exptl	*P*21/*n*	7.28	13.67	12.67	90.0	96.6	90.0
XV prd 1[b]	*P*21/*n*	7.26	13.82	12.52	90.0	97.4	90.0
XV prd 2	$P\bar{1}$	7.36	7.68	12.26	99.9	91.6	112.5
XV prd 3	*P*21/*c*	7.60	23.56	8.27	90.0	58.1	90.0

[a] Exptl indicates experimental structure and prd means predicted structure numbered by energy rank according to the hybrid method. [b] For convenience, these predicted structures have been transformed to match the space group setting and unit cell choice of the experimental structures.

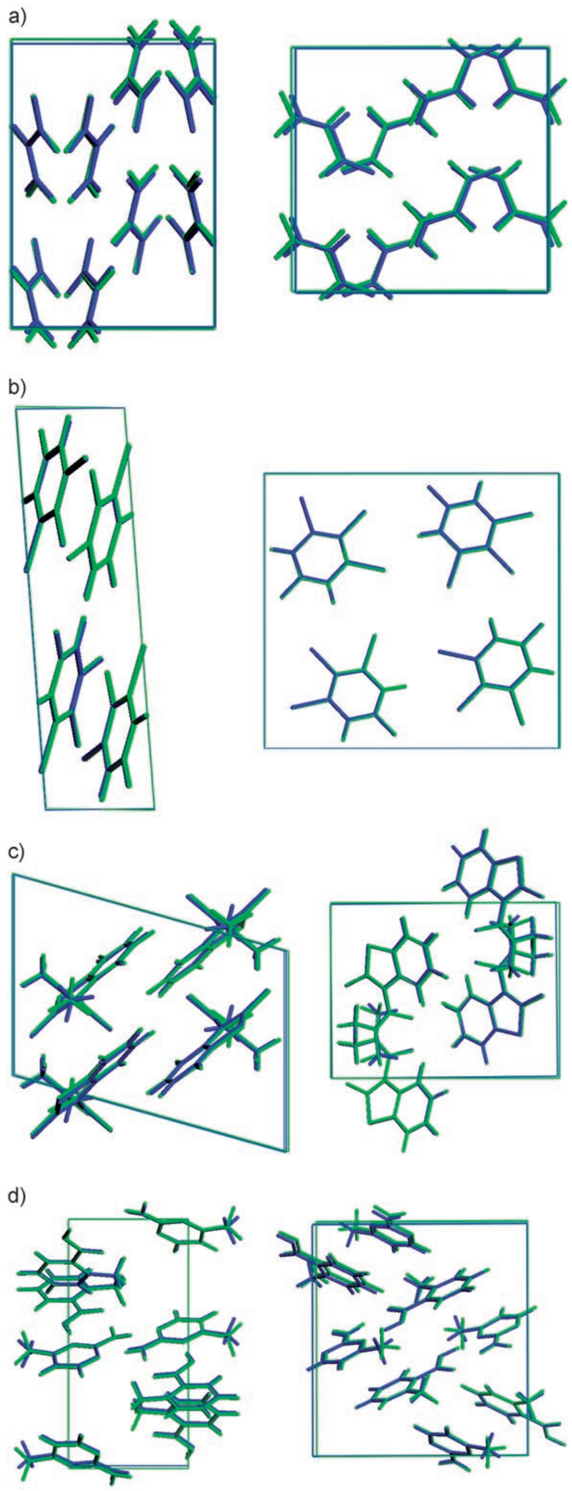


Figure 1. Superimposed experimental (blue) and predicted (green) crystal structures of the four compounds. Each comparison is viewed from two angles. a) **XII**, b) **XIII**, c) **XIV**, d) **XV**.

likely crystal structures. It is clear from this work that further improvements in the force-field descriptions of molecules in the solid state are still required. Other challenges involve highly flexible molecules, crystals with more than one molecule per asymmetric unit, solvates, salts, and cocrystals, for which the increased number of degrees of freedom typically leads to a larger number of potential structures that need to be considered. Nucleation dynamics, solvent effects, temperature, pressure, and other kinetic factors clearly play an important role in crystallization; otherwise the phenomenon of polymorphism would be less prevalent. Inclusion of these factors in CSP remains elusive.

Received: September 14, 2007

Keywords: crystal engineering · crystal structure prediction · density functional calculations · molecular mechanics · polymorphism

---

[1] S. L. Price, *Adv. Drug Delivery Rev.* **2004**, *56*, 301.

[2] J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer, D. E. Williams, *Acta Crystallogr. Sect. B* **2000**, *56*, 697.

[3] W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer, D. E. Williams, *Acta Crystallogr. Sect. B* **2002**, *58*, 647.

[4] G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, P. Verwer, *Acta Crystallogr. Sect. B* **2005**, *61*, 511.

[5] P. G. Karamertzanis, P. R. Anandamanoharan, P. Fernandes, P. W. Cains, M. Vickers, D. A. Tocher, A. J. Florence, S. L. Price, *J. Phys. Chem. B* **2007**, *111*, 5326.

[6] M. D. Gourlay, J. Kendrick, F. J. J. Leusen, *Cryst. Growth Des.* **2007**, *7*, 56.

[7] F. J. J. Leusen, *Cryst. Growth Des.* **2003**, *3*, 189.

[8] A. Gavezzotti, *J. Phys. Chem. B* **2003**, *107*, 2344.

[9] B. P. van Eijck, J. Kroon, *J. Comput. Chem.* **1999**, *20*, 799.

[10] D. W. M. Hofmann, T. Lengauer, *J. Mol. Struct.* **1999**, *474*, 13.

[11] P. Verwer, F. J. J. Leusen in *Reviews in Computational Chemistry, Vol. 12* (Eds.: K. B. Lipkowitz, D. B. Boyd), Wiley-VCH, New York, **1998**, pp. 327–365.

[12] M. A. Neumann, M. A. Perrin, *J. Phys. Chem. B* **2005**, *109*, 15531.

[13] G. Kresse, D. Joubert, *Phys. Rev. B* **1999**, *59*, 1758.

[14] G. Kresse, J. Furthmuller, *Phys. Rev. B* **1996**, *54*, 11169.

[15] G. Kresse, J. Hafner, *Phys. Rev. B* **1993**, *47*, 558.

[16] M. A. Neumann, *J. Phys. Chem. B*, accepted .

[17] J. Chisholm, W. D. S. Motherwell, *J. Appl. Crystallogr.* **2005**, *38*, 288.