Chemical Physics Letters 626 (2015) 20–24

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Crystal structure prediction from first principles: The crystal structures of glycine

Albert M. Lund [a,b], Gabriel I. Pagola [d], Anita M. Orendt [b], Marta B. Ferraro [d], Julio C. Facelli [b,c,*]

[a] Department of Chemistry, University of Utah, 155 South 1452 East Room 405, Salt Lake City, UT 84112, United States
[b] Center for High Performance Computing, University of Utah, 155 South 1452 East Room 405, Salt Lake City, UT 84112, United States
[c] Department of Biomedical Informatics, University of Utah, 421 Wakara, Salt Lake City, UT 84108, United States
[d] Departamento de and Ifiba (CONICET) (1428), Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. I, 1428 Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 27 January 2015
In final form 11 March 2015
Available online 18 March 2015

ABSTRACT

Here we present the results of our unbiased searches of glycine polymorphs obtained using the genetic algorithms search implemented in MGAC, modified genetic algorithm for crystals, coupled with the local optimization and energy evaluation provided by Quantum Espresso. We demonstrate that it is possible to predict the crystal structures of a biomedical molecule using solely first principles calculations. We were able to find all the ambient pressure stable glycine polymorphs, which are found in the same energetic ordering as observed experimentally and the agreement between the experimental and predicted structures is of such accuracy that the two are visually almost indistinguishable.

1. Introduction

More than a decade ago Professor Desiraju published [1] a critical article identifying crystal structure prediction from first principles as one of the most important unsolved problems in computational material science and questioned if this problem could ever be solved. Since 1997 there has been much effort toward the goal of being able to readily and reliably predict, by computational methods alone, the crystal structure of a molecule based only on its chemical diagram [2–5]. The process to do this is depicted in Figure 1.

The ability to accomplish this goal has far reaching implications well beyond just intellectual curiosity. On a basic science level, this can lead to an understanding of the principles that control crystal growth, by providing accurate information on the crystal energetics necessary for any further dynamical model of aggregation. More practically, the ability to successfully predict crystal structures based on computation alone will have a significant impact in many industries for which crystal structure and stability plays a critical role in product formulation and manufacturing, including pharmaceuticals, agrochemicals, pigments, dyes and explosives [6].

The current status of crystal structure prediction (CSP) can be evaluated by the performance of the participants in the periodic blind tests that have been organized by the Cambridge Crystallographic Data Centre (CCDC) [4,7,8]. The results of the last two blind tests showed the advantage of using dispersion corrected density functional theory (DFT-D) [9–11] to create a tailored molecule specific force field that is used to generate trial structures and to reorder a subset of the trial structures in search of the lowest energy crystal structures [12–16]. The software used in this approach is proprietary. There has also been an attempt to utilize first principle calculations to predict the crystal structure of organic crystals [17], in which the authors used a two-step approach optimizing the crystal structures with constrained molecular geometries in the initial stages and allowing full relaxation in the final stages. In this approach the authors used a combination of open source and proprietary software tools for the constrained and fully relaxed optimizations.

These results lend promise to using DFT-D methods to completely replace molecular mechanics and/or multistep optimization approaches as the method of choice for the evaluation of the energies of the trial crystal structures in CSP. To the authors knowledge there are no reports of any open source software capable of successfully predict crystal structures of molecules of biomedical interest directly from first principles without using either common or tailored potentials as intermediate steps and/or multistep optimization strategies.

[*] Corresponding author at: Department of Biomedical Informatics, University of Utah, 421 Wakara, Salt Lake City, UT 84108 84108, United States
E-mail address: julio.facelli@utah.edu (J.C. Facelli).

http://dx.doi.org/10.1016/j.cplett.2015.03.015
0009-2614/© 2015 Elsevier B.V. All rights reserved.

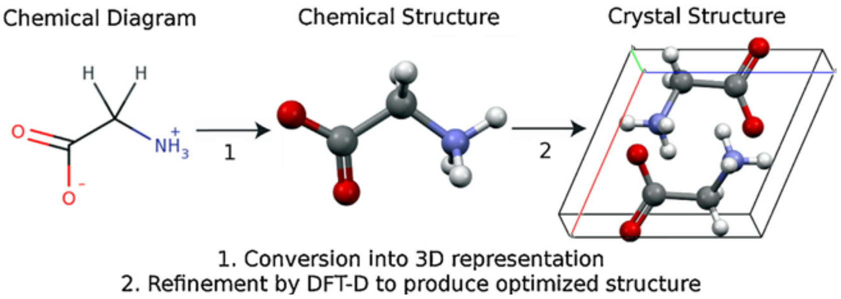
Chemical Diagram
Chemical Structure
Crystal Structure
H
H
O
NH3
O
1
2
1. Conversion into 3D representation
2. Refinement by DFT-D to produce optimized structure

Figure 1. Overview of the crystal structure prediction (CSP) process, which attempts to predict the structure or structures (when polymorphs exist) of a molecular entity based solely in its chemical diagram. The prediction of these structures directly from first principles has been identified a one of the greatest challenges remaining in computational molecular sciences.

It is important to realize that local optimization of plausible crystal structures is not a feasible approach for CSP. We have recently demonstrated [18], using a set of drug like molecules, that local optimization using full DFT-D results in near experimental structures only when the starting point is quite close to the experimental one. Therefore global optimization with a reliable and universal energy function is necessary for accurate CSP.

The MGAC (modified genetic algorithm for crystals) package[1] has been developed in our lab over the last decade [19–23]. MGAC is capable of doing CSP for any space group, any number of molecules per asymmetric unit, and can take into account the conformational flexibility of the molecule both at the local and global optimization levels. This allows an efficient, GA (genetic algorithms) based, global exploration of the crystal energy landscape. The previously released versions of MGAC relied on the use of the CHARMM [24,25] molecular mechanics program using the generalized atomic force field (GAFF) [26] for the energy evaluation and local minimization of the GA trial structures.

Previously, we used the set of molecules present in the Karamertzanis and Price (K&P) paper [27], to demonstrate the capabilities of the MGAC-CHARMM program [23]. These results demonstrated that the implementation of the GA in MGAC was effective and was always able to find the correct experimental structures provided that the GAFF potential energy represented the experimental energy landscape with sufficient fidelity. However, the matches to the experimental structure ranged from rank 1 to rank 1182 in terms of energy, highlighting the second issue with the use of the generic force field, namely the unreliability of the energy ranking.

Our more recent work [18] has demonstrated that when using Quantum Espresso (QE) to locally optimize the experimental structures in the K&P set the calculated local minima structure compares well with the experimental structure in all 32 of the molecules, with RMS differences ranging from 0.056 to 0.459 Å. This implies that for unknown structures, an approach which couples the use of MGAC with energy evaluation using QE will be successful in finding the ‘true’ experimental structures.

In this Letter we report the results of our unbiased searches for glycine polymorphs obtained using the global GA search implemented in MGAC coupled with local optimizations and energetics provided by QE (MGAC-QE). To our knowledge here we demonstrate for the first time that it is possible to predict the crystal structure of a molecule of biomedical interest, glycine, using solely first principles calculations (DFT-D) of the crystal energetics without using any intermediate steps, such as constructing special interatomic potentials, reordering the structures found by the search algorithm and/or using multistep search strategies with non-uniform approximations for the energy calculations. The only difference in the calculations presented here and a complete blind CSP search is that we only performed searches in the known space groups of each of the three stable polymorphs of glycine at atmospheric pressure.

2. Methods

Using the existing MGAC framework we have integrated the QE calculation of the energy and local optimizations into the framework as well as reworked the way in which the initial populations are selected and how the genetic algorithms were implemented (MGAC-QE). A full account of the technical and computational details of the integration of QE into the MGAC framework will be presented in detail elsewhere, along with the documentation and instructions on how to use the software that we will make available as an open source tool.

Glycine’s biological interest, relatively small size and polymorphic characteristics make it a good case to demonstrate the ability of MGAC-QE to predict the crystal structures of biomedical relevant compounds. Glycine is a precursor to the synthesis of proteins, a building block to numerous natural products, and provides the central C_2N subunit of all purines. It is a relatively small, semi rigid molecule, for which polymorphism is well establish in the literature. The existence of polymorphism is critical to demonstrate the usefulness of MGAC-QE to successfully predict crystal structures of biomedical interest for which the existence of polymorphism is prevalent [6].

Glycine has three room temperature and atmospheric pressure polymorphs: α-glycine ($P2_{1/c}$) [28], β-glycine ($P2_1$) [29], and γ-glycine ($P3_1/P3_2$) [30] (stability order: γ-glycine > α-glycine > β-glycine), as well as two high pressure polymorphs, δ-glycine (high pressure of the β-glycine form) [29], and ε-glycine (the high pressure form of the γ-glycine form) [31]. For the purpose of comparison of our results we used the following glycine reference structures from the Cambridge Structural Database (CSD): GLYCIN98 for α-glycine [28], GLYCIN71 for β-glycine [29], and GLYCIN33 for γ-glycine [30]. These three experimental structures were locally optimized using the QE vc-relax option, which allows for optimization of the unit cell parameters along with all atomic coordinates, with the same QE parameters used in our previous work (see below) [18]. In all cases the experimental structures converged to local minima in close proximity to the experimental structures. The QE energies for these local minima structures are $E_{\alpha\text{-glycine}} = -147\,662.07$ kJ/mol, $E_{\beta\text{-glycine}} = -147\,659.78$ kJ/mol, and $E_{\gamma\text{-glycine}} = -147\,663.10$ kJ/mol, which reproduce the experimental stability order: γ-glycine > α-glycine > β-glycine.

Following these preliminary tests we conducted unbiased global searches for crystal structures in the following space groups, with a number of molecules per unit cell given in parenthesis:

[1] The source code for MGAC-QE will be made available by an open source mechanisms once it is sufficiently stable for wide distribution.

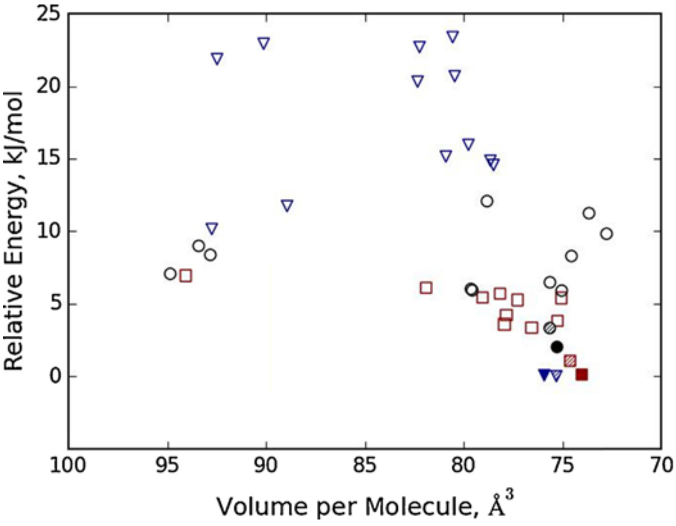

Figure 2. The distribution of the energies of crystals in the MGAC-QE populations as a function of the molecular volume. The crystals structures corresponding to the $P2_{1/c}$, $P2_1$, and $P3_1$ GA runs are represented by squares, circles and triangles, respectively. The hatched markers correspond to the experimental structures and the solid ones to the lowest energy structures found by MGAC-QE. For the experimental structures the volumes are those from the original reference and the energies correspond to the energy of the local minima obtained by QE when starting the minimization at the experimental energy.

$P2_{1/c}$ (4), $P2_1$ (2) and $P3_1$ (3). All calculations were performed using a population size of 120 individuals, a replacement rate of 1.0 per generation, and the searches were run for 50 generations. The probability of an individual being mutated was 0.01, and the probability of a crossover occurring between two individuals was 1. The selection method was a roulette wheel, using linear scaling of the energy, with the lowest energy structure having the largest selection probability. The optimization parameters for the QE optimization were again identical to those used by Lund et al. [18]. The DFT functional used was the Perdew–Burke–Ernzerhof generalized gradient approximation [32]. The dispersion correction method selected was the semi-empirical D2 method proposed by Grimme as implemented in Quantum Espresso [10]. The self-consistency threshold was set to 10^{-7} Ry and the plane wave cutoff energy was set to 55 Ry per the recommendation of the pseudopotentials creators. The pseudopotentials used for glycine were the Rappe–Rabe–Kaxiras–Joannopoulos-Ultrasoft pseudopotentials provided at the QE website, http://www.quantum-espresso.org/.

Calculations were performed on a LINUX cluster using six 16-core nodes (2×8-core Intel Xeon E5-2670 processors clocked at 2.60 GHz), with 64 GB memory per node and Mellanox FDR Infiniband for node interconnectivity. The total number of core hours for each run was: α-glycine: 10 238 core hours; β-glycine: 7174 core hours; and γ-glycine: 9518 core hours. Therefore the total number of core hours used for these three searches was 26 930, which represent a total elapsed time of approximately 12 days.

3. Results and discussion

The results of the analysis of the populations generated by the MGAC-QE runs described above are presented in Figure 2. This figure presents, as suggested by Price [33], the distribution of the energies of crystals in the MGAC populations as function of their volume. As expected when polymorphism is present, the plot shows a great deal of crowding and the volume energy pairs of the different polymorphs are not well separated [33]. This clustering of the three polymorphs reinforces that glycine is a challenging case for CSP and therefore a stringent test for the MGAC-QE method. From the figure it is also apparent that the structures found by MGAC-QE (solid symbols) for each of the symmetry groups studied here closely match the experimental ones (hatched symbols) corresponding to the most stable polymorphs, in the same space group.

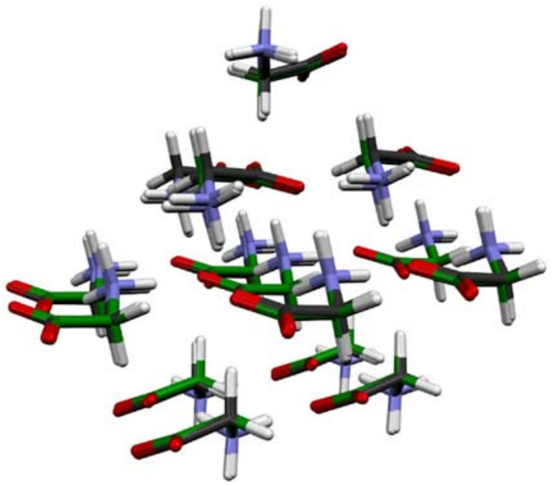

Figure 3. Comparison of the experimental structure of the α-glycine CSD structure GLYCIN98 from Ref. [28] (black) with the lowest energy structure found by MGAC-QE in the $P2_{1/c}$ space group (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Notably, in some initial generations we observed structures where the protonation state of glycine was altered and the non-zwitterionic form was adopted. This is made possible by the unconstrained optimization algorithm in QE. These structures were typically much higher in energy (by >80 kJ/mol) than structures remaining in the zwitterionic form, and were therefore eliminated rapidly from the population. The conclusion drawn from this is that one must be careful to identify low energy structures where the protonation state (and in general, bonding state) might be altered.

In Table 1 the crystallographic parameters and the calculated energies of the best structures found by MGAC-QE for each of the space groups considered here along with the RMS between them and the corresponding experimental structures are given. The results in Table 1 show an excellent agreement between the MGAC-QE predicted structures and the experimental ones; the agreement is apparent in both in the cell parameters as well as the RMS difference between the experimental and predicted structures. The RMS values can be compared to the RMS values observed when comparing different experimental structures of the same polymorphs reported in the CSD; for instance the RMS between α-glycine structures GLYCIN89 and GLYCIN17 is 0.026 Å, for β-glycine structures GLYCIN74 and GLYCIN25 is 0.114 Å and for γ-glycine structures GLYCIN65 and GLYCIN15 is 0.07 Å.

The energies of the MGAC-QE predicted structures follow the experimental stability order: $E_{\gamma\text{-glycine}} < E_{\alpha\text{-glycine}} < E_{\beta\text{-glycine}}$, with α-glycine and β-glycine 70 J/mol and 1950 J/mol, respectively, less stable than γ-glycine. These values can be compared with recent values from the literature [34] of 962 J/mol and 1506 J/mol, respectively, obtained using the DFT method plus many body dispersion correction and zero point energy corrections (PBeh + MBD + ZPE) [34].

The graphical comparison between the experimental and the best MGAC-QE structures is presented in Figs. 3–5. This comparison does not require additional discussion, as it is apparent that the agreement is of such quality that the two structures are almost indistinguishable.

In conclusion, using MGAC-QE we were able to find each of the ambient pressure stable polymorphs of glycine when searching in

Table 1
Comparison of the energies and geometries of the α-glycine, β-glycine and γ-glycine structures found by MGAC-QE with the reference experimental structures.

Polymorph	SPG	Energy[d]	Cell parameters[e]						RMS[f]
			a	*b*	*c*	*α*	*β*	*γ*	
α-Glycine									
MGAC-QE	$P2_{1/c}$	−147 663.00	5.0517	11.7146	5.7965	90.0	120.3102	90.0	0.097
Exp[a]			5.0874	11.7817	5.4635	90.0	112.0530	90.0	
β-Glycine									
MGAC-QE	$P2_1$	−147 661.12	5.6840	6.0727	5.0305	90.0	119.8711	90.0	0.199
Exp[b]			5.3880	6.2760	5.0905	90.0	113.1200	90.0	
γ-Glycine									
MGAC-QE	$P3_1$	−147 663.07	6.9166	6.9166	5.4983	90.0	90.0	120.0	0.087
Exp[c]			7.0383	7.0383	5.4813	90.0	90.0	120.0	

[a] Structure GLYCIN98 (10 K) from Ref. [28].
[b] Structure GLYCIN71 (room temperature) from Ref. [29].
[c] Structure GLYCIN33 (room temperature) from Ref. [30].
[d] Energy in kJ/mol for the lowest energy found by MGAC-QE in the corresponding space group.
[e] Crystallographic axis in Å, cell angles in degrees.
[f] Computed using the Solid Form Crystal Packing Similarity method in Mercury CSD with 15 molecules for comparison and ignoring hydrogen atoms [35].

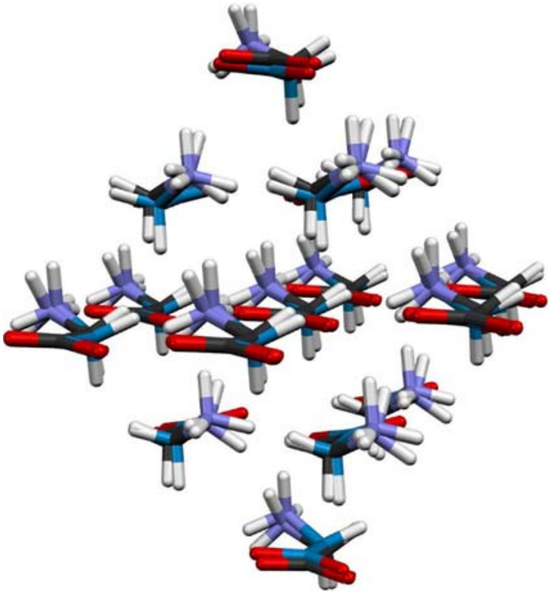

Figure 4. Comparison of the experimental structure of the β-glycine CSD structure GLYCIN71 from Ref. [29] (black) with the lowest energy structure found by MGAC-QE in the $P2_1$ space group (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

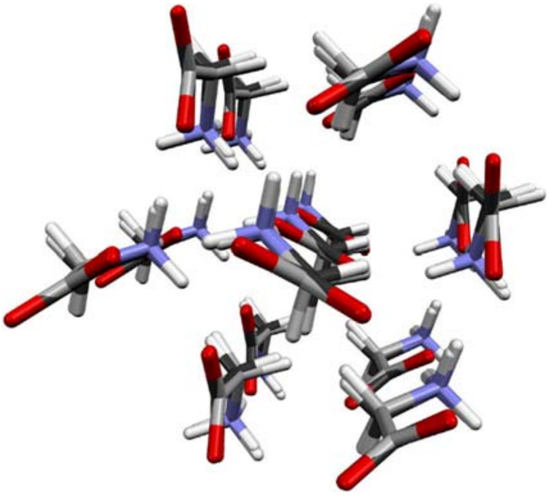

Figure 5. Comparison of the experimental structure of the γ-glycine CSD structure GLYCIN33 from Ref. [30] (black) with the lowest energy structure found by MGAC-QE in the $P3_1$ space group (white).

their corresponding space group. The match to the experimental structure was the lowest energy structure found in each of the three searches. The polymorphs encountered by MGAC-QE are energetically ordered in agreement with experimental results and the comparison of the experimental and predicted structure is of such accuracy that the two are visually almost indistinguishable. When the success of MGAC-QE is compared with the results for glycine in Ref. [17], it becomes apparent that allowing the full relaxation of both molecular and crystal structural parameters as well as using a single approach for the calculation of the crystal energies at all stages of the global optimization is critical for successful CSP. However, there is already enough evidence in the literature that current functional and dispersion correction lattice energies may not be adequate for all crystals, particularly disordered ones, therefore new DFT approaches may be needed to address those systems.

The computer times required by the calculations reported here are significant, but manageable. Computer times for larger system will be a significant challenge, but we are confident that we will be able to greatly improve performance once we better understand the optimal GA parameters like population size, replacement and number of generations and area able to make use of emerging computer technologies like GPU accelerators. A truly blind test of the method, exploring most common space groups and/or using searches in *P1* with different number of molecules per unit cell is the next goal. The exact search strategies will be defined by studies that are underway in our laboratory to establish the most efficient search protocols for blind test CSP. The results of this exploration will be used to participate in the current sixth CSP blind test, and our results will be presented at the 2015 CCDC meeting in the fall of 2015.

4. Conclusions

The results presented here show that it is possible to predict the crystal structures of molecules of biomedical interest from first principles without using any intermediate potentials, energy reordering strategies and/or step wise optimization strategies. With these results we believe that we can answer Professor Desiraju's question with an unquestionable yes! Crystal structures can be predicted from first principles and with existing computational resources and appropriate optimization of our methods, CSP can become a standard tool for material design.

Author contributions

All authors contributed to the design of the project, analysis of the data, and to the preparation of the manuscript. AML and GIP contributed to the MGAC-QE code development. AML conducted the calculations reported on glycine.

Conflict of interest

None of the authors declare any competing financial interest regarding the contents of this Letter.

Acknowledgements

Computer resources were provided by the Center for High Performance Computing at the University of Utah and the Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF grant number ACI-1053575. JCF has been partially funded by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number 1ULTR001067. MBF and GIP acknowledge the support from the University of Buenos Aires and the Argentinean Research Council. For information on the MGAC software, please contact the corresponding author.

References

[1] G.R. Desiraju, Science 278 (1997) 404.
[2] G.M. Day, Supramolecular Chemistry: From Molecules to Nanomaterials, John Wiley & Sons, Ltd., 2012.
[3] C.W. Lehmann, Angew. Chem. Int. Ed. 50 (2011) 5616.
[4] D.A. Bardwell, et al., Acta Crystallogr. B67 (2011) 535.
[5] J. Kendrick, F.J.J. Leusen, M.A. Neumann, J. Van De Streek, Chem. Eur. J. 17 (2014) 10736.
[6] S. Datta, D.J.W. Grant, Nat. Rev. Drug Discov. 3 (2004) 42.
[7] G.M. Day, et al., Acta Crystallogr. B: Struct. Sci. 61 (2005) 511.
[8] G.M. Day, et al., Acta Crystallogr. B: Struct. Sci. 65 (2009) 107.
[9] S. Grimme, J. Comput. Chem. 25 (2004) 1463.
[10] S. Grimme, J. Comput. Chem. 27 (2006) 1787.
[11] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, J. Chem. Phys. 132 (2010) 154104.
[12] M.A. Neumann, M.-A. Perrin, J. Phys. Chem. B 109 (2005) 15531.
[13] M.A. Neumann, 24th European Christallographic Meeting, Micro Symposium 14, Advanced Computational Methods in Structural Chemistry, Marrakech, Morocco, 2007, 11H00–11H20.
[14] M.A. Neumann, F.J.J. Leusen, J. Kendrick, Angew. Chem. Int. Ed. 47 (2008) 2427.
[15] J. Kendrick, F.J.J. Leusen, M.A. Neumann, J. Van De Streek, Chem. Eur. J. 17 (2011) 10736.
[16] M.A. Neumann, J. Phys. Chem. B 112 (2008) 9810.
[17] Q. Zhu, A.R. Oganov, C.W. Glass, H.T. Stokes, Acta Crystallogr. (B68) (2012) 215.
[18] A.M. Lund, A.M. Orendt, G.I. Pagola, M.B. Ferraro, J.C. Facelli, Cryst. Growth Des. 13 (2013) 2181.
[19] V.E. Bazterra, M.B. Ferraro, J.C. Facelli, J. Chem. Phys. 116 (2002) 5984.
[20] V.E. Bazterra, M.B. Ferraro, J.C. Facelli, J. Chem. Phys. 116 (2002) 5992.
[21] V.E. Bazterra, M.B. Ferraro, J.C. Facelli, Int. J. Quantum Chem. 96 (2004) 312.
[22] V.E. Bazterra, M. Thorley, M.B. Ferraro, J.C. Facelli, J. Chem. Theory Comput. 3 (2007) 201.
[23] S. Kim, A.M. Orendt, M.B. Ferraro, J.C. Facelli, J. Comput. Chem. 30 (2009) 1973.
[24] B.R. Brooks, et al., J. Comput. Chem. 4 (1983) 187.
[25] A.D. MacKerell, et al., in: P.v.R. Schleyer, et al. (Eds.), The Encyclopedia of Computational Chemistry, John Wiley & Sons, Ltd., 1998, pp. 271–277.
[26] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, J. Comput. Chem. 25 (2004) 1157.
[27] P.G. Karamertzanis, S.L. Price, J. Chem. Theory Comput. 2 (2006) 1184.
[28] T. Aree, H.-B. Bürgi, J. Phys. Chem. A 116 (2012) 8092.
[29] N.A. Tumanov, E.V. Boldyreva, H. Ahsbahs, Powder Diffr. 23 (2008) 307.
[30] E.V. Boldyreva, T.N. Drebushchak, E.S. Shutova, Z. Kristallogr. 218 (2003) 366.
[31] E.V. Boldyreva, S.N. Ivashevskaya, H. Sowa, H. Ahsbahs, H.-P. Weber, Z. Kristallogr. 220 (2005) 50–57.
[32] J.P. Perdew, M. Ernzerhof, K. Burke, J. Chem. Phys. 105 (1996) 9982.
[33] S.L. Price, Acc. Chem. Res. 42 (2009) 117.
[34] N. Marom, et al., Angew. Chem. Int. Ed. 52 (2013) 6629.
[35] J.A. Chisholm, S. Motherwell, J. Appl. Crystallogr. 38 (2005) 228.