Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 18 July 2011
Accepted 12 October 2011

Dedicated to Professor Philip Coppens on the occasion of his 80th birthday

New version of the theoretical databank of transferable aspherical pseudoatoms, UBDB2011 – towards nucleic acid modelling

Katarzyna N. Jarzembska* and Paulina M. Dominiak*

Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland. Correspondence e-mail: katarzyna.jarzembska@gmail.com, pdomin@chem.uw.edu.pl

The theoretical databank of aspherical pseudoatoms (UBDB) was recently extended with over 100 new atom types present in RNA, DNA and in some other molecules of great importance in biology and pharmacy. The atom-type definitions were modified and new atom keys added to provide a more precise description of the atomic charge-density distribution. X—H bond lengths were updated according to recent neutron diffraction studies and implemented in the *LSDB* program as well as used for modelling the appropriate atom types. The UBDB2011 databank was extensively tested. Electrostatic interaction energies calculated on the basis of the databank of aspherical atom models were compared with the corresponding results obtained directly from wavefunctions at the same level of theory (SPDFG/B3LYP/6-31G** and SPDFG/B3LYP/aug-cc-pVDZ). Various small complexes were analysed to cover most of the different interaction types, *i.e.* adenine–thymine and guanine–cytosine with hydrogen bonding, guanine–adenine with stacking contacts, and a group of neutral and charged species of nucleic acid bases interacting with amino acid side chains. The energy trends are well preserved ($R^2 > 0.9$); however the energy values differ between the two methods by about 4 kcal mol^{-1} (1 kcal mol^{-1} = 4.184 kJ mol^{-1}) on average. What is noticeable is that the replacement of one basis set by another in a purely quantum chemical approach leads to the same electrostatic energy difference, *i.e.* of about 4 kcal mol^{-1} in magnitude. The present work opens up the possibility of applying the UBDB2011 for macromolecules that contain DNA/RNA fragments. This study shows that on the basis of the UBDB2011 databank electrostatic interaction energies can be estimated and structure refinements carried out. However, some method limitations are apparent.

1. Introduction

The experimental determination of charge-density distribution is a difficult and complex task. What is more, diffraction data quality is often not good enough to obtain reliable charge-density results. The confidence in experimental charge density might be compromised either by experimental errors, multipole pseudoatom model limitations, or a lack of accurate phases and large uncertainties in the hydrogen-atom positions and thermal motion. Brock *et al.* (1991) introduced the new idea of transferability of pseudoatom parameters between different molecules, initiating the creation of databanks of aspherical atom models. To date, there are three well established databanks: the experimental databank ELMAM/ELMAM2 (Pichon-Pesme *et al.*, 1995, 2004; Domagała & Jelsch, 2008), the theoretical Invariom database (Dittrich *et al.*, 2004; Dittrich, Hübschle *et al.*, 2006) and the University at Buffalo Pseudoatom Databank (UBDB) (Volkov, Li *et al.*, 2004; Dominiak *et al.*, 2007).

The existing pseudoatom databases offer the possibility of performing structure refinement with the use of aspherical scattering factors computed from the transferable aspherical atom model (TAAM). This constitutes an improvement over the extensively applied independent atom model (IAM) refinement which does not include atomic charge-density deformations due to bond formation or lone pairs. It is, therefore, possible to model electron-density distribution and then to deconvolute thermal motion more accurately for typical X-ray data ($\sin\theta/\lambda < 0.7$ Å^{-1}) (Volkov *et al.*, 2007; Dittrich *et al.*, 2008; Pichon-Pesme *et al.*, 1995). Such atomic displacement parameter values are consequently closer to those obtained from multipole refinements of high-resolution X-ray data (Volkov *et al.*, 2007; Dittrich *et al.*, 2008; Bąk *et al.*, 2011). It has already been shown that the TAAM refinement

doi:10.1107/S0108767311042176

significantly improves the discrepancy R factors, molecular geometry (Volkov *et al.*, 2007; Dittrich, Strumpel *et al.*, 2006; Dittrich *et al.*, 2007; Jelsch *et al.*, 2005; Bąk *et al.*, 2011) and precision of the Flack parameter determination (Dittrich, Strumpel *et al.*, 2006) with respect to the standard method based on the IAM. Additionally, the aforementioned databanks can be employed to reconstruct the electron-density distribution of macromolecules and further to estimate the electrostatic properties of such complex systems (Volkov, King, Coppens & Farrugia, 2006; Li *et al.*, 2006; Dominiak *et al.*, 2009; Lecomte *et al.*, 2005; Zarychta *et al.*, 2007). According to Bąk *et al.* (2011), electrostatic interaction energies ($E_{\rm es}$'s) computed on the basis of each database model are closer to the results obtained theoretically for isolated molecules than to those derived from periodic calculations. The smallest differences in the $E_{\rm es}$ values with respect to *ab initio* results and the highest correlations were found for the UBDB database.

As described in the previous studies, the UBDB is a databank of aspherical pseudoatoms derived by pseudoatoms Fourier-space fitting to molecular electron-density distributions obtained from *ab initio* calculations. Apart from its application to the refinement of X-ray data, the UBDB is designed for the evaluation of the electrostatic properties of large molecular complexes from the reconstituted molecular electron density. It has already been shown that, in the case of amino acids, the UBDB predicts local and integrated properties of the electron density and electrostatic interaction energies with chemical accuracy (Volkov, Koritsanszky & Coppens, 2004; Volkov, Li *et al.*, 2004; Volkov, King, Coppens & Farrugia, 2006). The older version included all atom types encountered in amino-acid residues.

Hence, in this paper we describe the extension of the databank by systematic application of the spawning procedure. Recently, the databank was extended with a set of over 100 new atom types with an eye towards RNA and DNA molecules. The presence of substituted heteroaromatic rings and the importance of $\pi\cdots\pi$ interactions for such systems necessitate careful testing of the already established algorithm for the atom-type definitions. In this paper we present the new version of the University at Buffalo Databank and the related *LSDB* program (Volkov, Li *et al.*, 2004) together with its application potential and method limitations. UBDB2011 so far contains about 215 atom types with more precise atom-type definitions. The *LSDB* code was additionally supplemented with the updated X—H bond lengths according to the latest publication of Allen & Bruno (2010).

2. Methodology

2.1. Databank construction

Following the procedure of constructing the previous version of the databank (Dominiak *et al.*, 2007), a number of single-point calculations on a set of small molecules was performed with the *GAUSSIAN03* program (Frisch *et al.*, 2004) using density functional theory (DFT) with a 6-31G** basis set (Krishnan *et al.*, 1980) and B3LYP functional (Perdew, 1986; Becke, 1988; Lee *et al.*, 1988). The aforementioned treatment was applied to the selected good-quality experimental molecular geometries according to the Cambridge Structural Database (CSD) (Allen, 2002). Corresponding refcodes are listed in the supplementary material.[1] Hydrogen-atom positions were obtained by extending X—H distances to their standard neutron diffraction values with the use of the new version of the *LSDB* program. As the UBDB2011 version of the databank contains modified atomic keys describing the atom type's closest chemical environment more precisely with respect to its parent version, all of the calculations were carried out for both previously used molecular geometries and newly added ones. Also, much more attention was dedicated to the X—H distances. Hydrogen atoms are of great importance for estimating any electrostatic properties of an organic molecular system, especially interaction energy values, which are among the principal applications of the developed databank. All the different C—H, N—H, O—H, S—H and P—H bond lengths were defined according to the newest article of Allen & Bruno (2010). Complex static valence-only structure factors in the range $0 < \sin\theta/\lambda < 1.1$ Å^{-1} were derived by analytic Fourier transform of the molecular charge densities for reciprocal-lattice points corresponding to a pseudo-cubic cell with 30 Å edges (Koritsanszky *et al.*, 2002). Subsequently, they were fitted with the Hansen–Coppens pseudoatom formalism (Hansen & Coppens, 1978), using the *XD* program suite (Volkov, Macchi *et al.*, 2006). Derived multipole parameters together with κ and κ' values were averaged over a family of chemically similar atoms with an eye on their statistical consistence leading to a particular atom type stored in the databank.

2.2. Atom-type definitions

The construction of a databank requires selection of atoms that are similar enough to be averaged and, therefore, suitable to represent a particular atom type. Such a set of atom types should constitute the smallest possible number of pseudoatoms accurately reproducing the charge-density distribution of many molecules. The procedure of introducing the new atom type is the same as previously described. The following general criteria for the definition of an atom type are defined: (1) element type, (2) the number of attached atoms (atom valence, number of nearest neighbours), (3) nearest-neighbour type, which may be affected by the next-nearest neighbours, (4) aromaticity (ring planarity) and (5) local symmetry. In the new version of the databank point (4) is subdivided into the number of aromatic rings to which a given atom belongs and a summary ring member number. Consequently, atoms that are planar ring members are given a new aromaticity flag consisting of two key words: *RING* and

[1] Supplementary material is available from the IUCr electronic archives (Reference: SH5136). Services for accessing these data are described at the back of the journal.

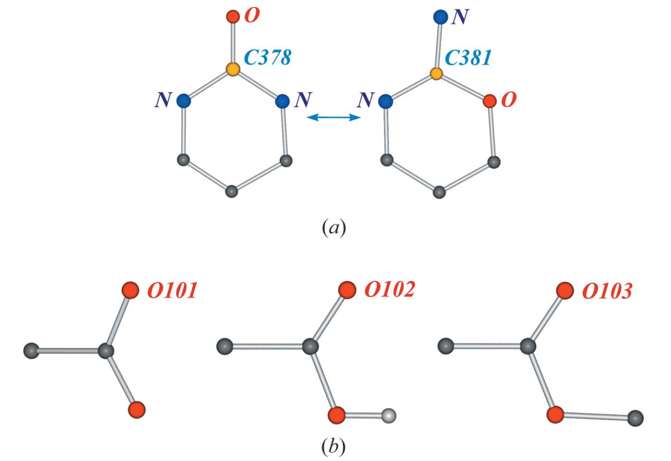


Figure 1
Examples of more precise neighbour-atom definition requirements: (*a*) carbon atoms C378 and C381; (*b*) oxygen atom types: carboxylate O101, carboxylic O102 and ester O103. Atom colour coding: oxygen – red, nitrogen – blue, carbon – dark grey, hydrogen – light grey.

MEMB in the databank file (db2011.db, see supplementary material). The first one stands for the number of planar rings while the second describes the total number of the selected ring members. Such a definition distinguishes atoms belonging to five- and six-membered rings and also enables a better description of those atoms that join fused aromatic rings together. This is of particular importance for DNA and RNA bases.

Generally, in order to characterize any neighbouring atom, only element type and hybridization are taken into account. Hybridization states of neighbouring atoms are derived solely on the basis of the number of atoms attached. However, nitrogen atoms are treated more specifically, as they are additionally split into sp^3(4), sp^3(3), sp^2(3) and sp^2(2) types (hybridization state followed by the number of closest neighbours in parentheses). Such an approach also considers the planarity of the nitrogen-containing group.

Whenever required, the described criteria are modified by providing more precise neighbour-atom definition, defining whether it belongs to a planar ring or not, or by including the effect of next-nearest neighbours. An example of the first situation is represented by statistically distinguishable carbon atoms C378 and C381 (Fig. 1*a*), while the second occurs for oxygen atoms such as O101, O102 and O103 (Fig. 1*b*). Whereas oxygen atoms O102 and O103 are statistically equal, O101 is significantly different. A simpler definition of an atom type is sometimes used when the number of occurrences of an intended atom type in the molecular sample is too small to get statistically meaningful average values of the deformation-density parameters or when it is difficult to define a certain atom type. These atoms are marked with a *TEMP* flag in the db2011.db file. When the number of the averaged atoms is equal to or lower than 5 the *TEMP* flag is followed by the number indicating the sum of the averaged atoms. These atom-type definitions are going to be improved together in the next version of the databank.

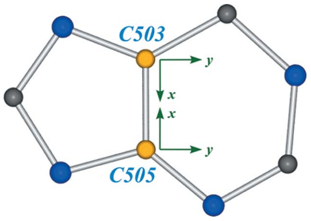


Figure 2
Local coordinate system schema for atoms belonging to two planar rings, C503 and C505. Atom colour coding: nitrogen – blue, carbon – dark grey.

2.3. Local coordinate system assignment

The *LSDB* program analyses the coordination environment and assigns the appropriate atom type for each atom present in a studied molecule. The corresponding charge-density parameters are then transferred from the databank for the subsequent procedures. It is an essential tool in the charge-density analysis of any large molecule, for which the manual assignment of coordinate systems becomes prohibitively cumbersome. It is also crucial when constructing the pseudo-atom databank based on a large number of small molecules.

The main features of the program are preserved with respect to its earlier version (Volkov, Li *et al.*, 2004). The conducted modifications concern the atom-type definition module, the aromatic ring definition and the local coordinate system determination in the case of atoms at the intersection of fused aromatic rings. Additionally, new X—H lengths were incorporated into the code.

The local coordinate system associated with an atom is oriented such as to allow local symmetry constraints. When there is more than one possibility of selecting a particular local coordinate axis within a given local symmetry, a previously established procedure is applied. It is based on a set of criteria defining the analysed atom environment, *i.e.* atomic number, hybridization state, valency and, if it is still inconclusive, distance to the central atom. In some cases, it is necessary to add a dummy atom to make use of the symmetry allowing minimization of the number of multipole parameters. In the case of atoms belonging to one planar ring, the x axis is always oriented towards the centre of the ring and mirror-plane symmetry (m) is imposed even if higher symmetry is possible. When an atom is common to two fused planar rings, the x axis is directed towards the second atom of the kind, the y axis towards an atom from the six-membered ring keeping it at a right angle and the m point-group symmetry (Fig. 2). Usually, a right-handed coordinate system is defined, except in the case of chiral atoms for which both right-and left-handed systems are allowed. Chirality is defined locally, only by the character of the nearest neighbours.

2.4. Atom types in the databank

Currently, the databank contains 207 atom types, among which there are 17 hydrogen, 109 carbon, 26 nitrogen, 35 oxygen, 8 sulfur, 8 phosphorus, 2 chlorine and 2 fluorine atom types. It includes all atom types encountered in peptides, proteins, nucleic acid bases and some other biologically

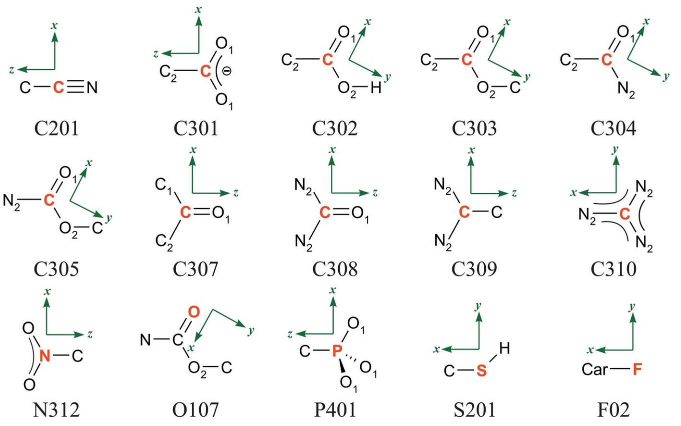


Figure 3
Example atom types stored in the UBDB2011 databank.

interesting molecules. The old atom types were recomputed, some were modified, others added. All atom types incorporated in the databank are listed in Tables 1S–2S in the supplementary material while the atom-type charge-density distribution parameters are collected in the db2011.db file. Sample atom types stored in the UBDB2011 databank are presented in Fig. 3. The current version of the databank, UBDB2011, and the corresponding *LSDB* program, are available free of charge from http://crystal.chem.uw.edu.pl or directly from the authors on request.

3. Results and discussion

3.1. Databank verification

On the basis of atom-type multipole parameters and corresponding κ values one can model the charge-density distribution of a macromolecular system for which high-resolution X-ray data are unavailable. Having such a reconstructed charge density, a number of its electrostatic properties can be computed, *e.g.* bond properties, molecular electrostatic moments, electrostatic potential. Such an analysis may constitute a good foundation for deriving essential information of the binding properties and electrostatic interactions present in the biological systems or pharmaceutical complexes. However, it is desirable to check whether it is justified to use a particular pseudoatom databank for such a purpose and, if so, to what extent.

Electrostatic interaction energy is most sensitive to any imperfection in the modelled charge-density distribution; therefore, we decided to use this electrostatic property in order to verify UBDB2011. The databank was extended with an eye to model atom types present in nucleic acid chains, RNA and DNA, and in some other species of interest. The majority of newly calculated parameters concern atoms belonging to pyrimidine and purine bases. Thus, obtained pseudoatom charge-density models were tested on a set of dimers formed by nucleic acid bases, for which *ab initio* calculations are possible.

The UBDB2011 databank, together with the *LSDB* program, was used to reconstruct the electron-density distributions of the adenine–thymine (A:T), guanine–cytosine (G:C) and also guanine–adenine (G:A) complexes (Fig. 4). Additionally, a set of nucleic acid bases (NABs) interacting with amino acid (AA) fragments, both neutral and charged, was considered (Fig. 4*d*, Fig. 3S in the supplementary material). The monomer geometries used in this study were optimized using the *GAMESS-US* package (Schmidt *et al.*, 1993) at the MP2/aug-cc-pVDZ level of theory imposing *m* point-group symmetry (Møller & Plesset, 1934; Dunning, 1989). The relative position and orientation of either guanine and cytosine, or adenine and thymine, were uniquely defined by the six base-pair parameters (shear, stagger, stretch, opening, buckle and propeller). The dimers were generated with the aid of the *3DNA* program based on the experimentally determined values of the base-pair parameters taken from crystallographic data (Czyżnikowska *et al.*, 2009, 2010). In the case of both A:T and G:C complexes the most representative 'crystallographic' mutual configurations were chosen. They were subsequently used for potential energy surface scans with respect to a given base-pair parameter, with the remaining five kept fixed. The ranges of the scanned parameters were selected regarding the experimental data and were divided into equal intervals. Altogether we tested 55 A:T and 60 G:C dimers, 48 G:A geometries (stacked complexes appearing in B-DNA crystals) and 13 NAB:AA side-chain benchmark complexes (Czyżnikowska, 2009; Czyżnikowska *et al.*, 2009). To ensure electroneutrality, all monomers were adjusted *a posteriori* to their net charges by scaling the pseudoatoms according to the Faerman and Price scaling algorithm (Faerman & Price, 1990) implemented in *LSDB*. The exact potential multipole method (EPMM) (Volkov, Koritsanszky & Coppens, 2004; Volkov, King, Coppens & Farrugia, 2006), implemented in the *XDPROP* module of the *XD* package, was employed to compute electrostatic interaction energies from the derived densities. It combines numerical evaluation of the exact Coulomb integral in the inner region ($\leq$ 4.5 Å) with the Buckingham-type multipole approximation for the long-range interatomic interactions (Buckingham, 1967).

Such electrostatic interaction energy ($E_{\rm es}$) values were then compared with the corresponding reference results obtained directly from the molecular wavefunctions at the same level of theory (B3LYP) with two different basis sets: 6-31G** and aug-cc-pVDZ. The *SPDFG* program (Volkov, King & Coppens, 2006) was used for the evaluation of $E_{\rm es}$ from monomer charge distributions expressed in terms of Gaussian-type basis functions. The program requires wave function input files which were calculated within the *GAUSSIAN03* package. All *GAUSSIAN03* calculations were conducted with the SCF = Tight option, which requests tight self-consistent field convergence criteria. The *SPDFG* program uses the numerical Rys quadrature method for the estimation of one- and two-electron Coulomb integrals (Dupuis *et al.*, 1976; Rys *et al.*, 1983). The *SPDFG* program $E_{\rm es}$ results might be taken as a reference point as it has already been shown that they are in excellent agreement with those obtained with the Moro-

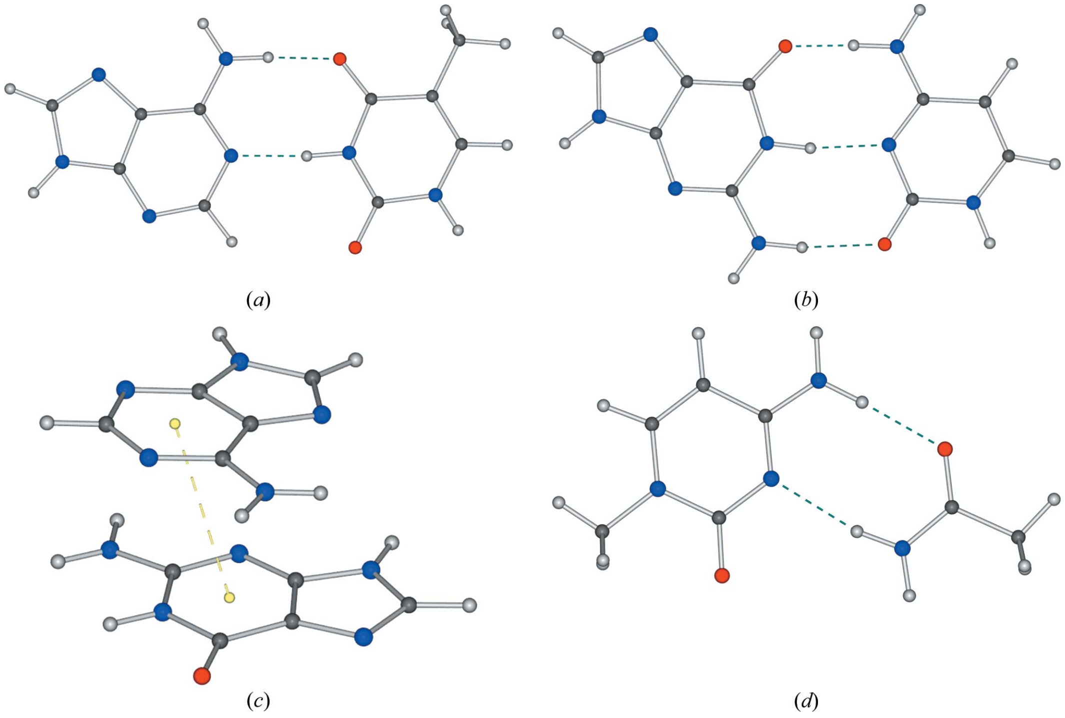


Figure 4
Selected chemical systems: (*a*) A:T; (*b*) G:C; (*c*) G:A; (*d*) uracil interacting with methylamide residue (NAB:AA). Atom colour coding: nitrogen – blue, oxygen – red, carbon – dark grey, hydrogen – light grey.

kuma–Ziegler energy decomposition scheme (Morokuma, 1971; Ziegler & Rauk, 1977) implemented in *GAMESS-US*.

The chosen systems cover the two most important interaction types present in DNA and RNA molecules, *i.e.* hydrogen bonding and $\pi\cdots\pi$ stacking interactions. However, so far, it is only possible to derive the electrostatic component of the total interaction energy on the basis of the molecular charge-density distribution obtained from the databank. The correlations between the combined UBDB2011 + EPMM approach and the *SPDFG* data are quite rewarding, $R^2 > 0.90$ in the case of A:T, G:C and NAB:AA complexes, and $R^2 = 0.75$ for G:A (Table 1, and Figs. 1S–2S and Tables 3S–6S in the supplementary material).

The overall root mean square deviation (RMSD) for all the data sets taken together amounts to 3.7 kcal mol^{-1} (1 kcal mol^{-1} = 4.184 kJ mol^{-1}), while the linear constant is close to unity and the correlation coefficient reaches the value of 0.99. The distribution of energy points between the UBDB2011 + EPMM and SPDFG methods for the analysed data series is shown in Fig. 5.

RMSDs between SPDFG/B3LYP/6-31G** *ab initio* values of electrostatic energy and the UBDB2011 + EPMM amounts to 2.2 kcal mol^{-1} for the set comprising 55 A:T Watson–Crick base pairs and to about 5.0 kcal mol^{-1} if we consider the set of G:C complexes (Table 2). What is interesting is that G:C UBDB2011 + EPMM results are systematically closer to SPDFG/B3LYP/aug-cc-pVDZ-derived energy values (RMSD = 1.4 kcal mol^{-1}) than to corresponding SPDFG/B3LYP/

Table 1
Determination coefficient (R^2) between UBDB2011 + EPMM and SPDFG $E_{\rm es}$ values.

Complex type	R^2
A:T	0.91
G:C	0.97
G:A	0.75
NAB:AA	0.99

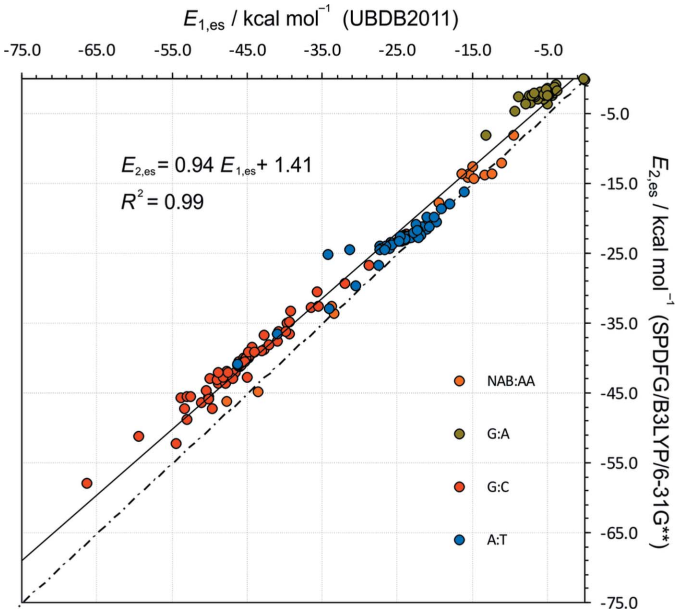
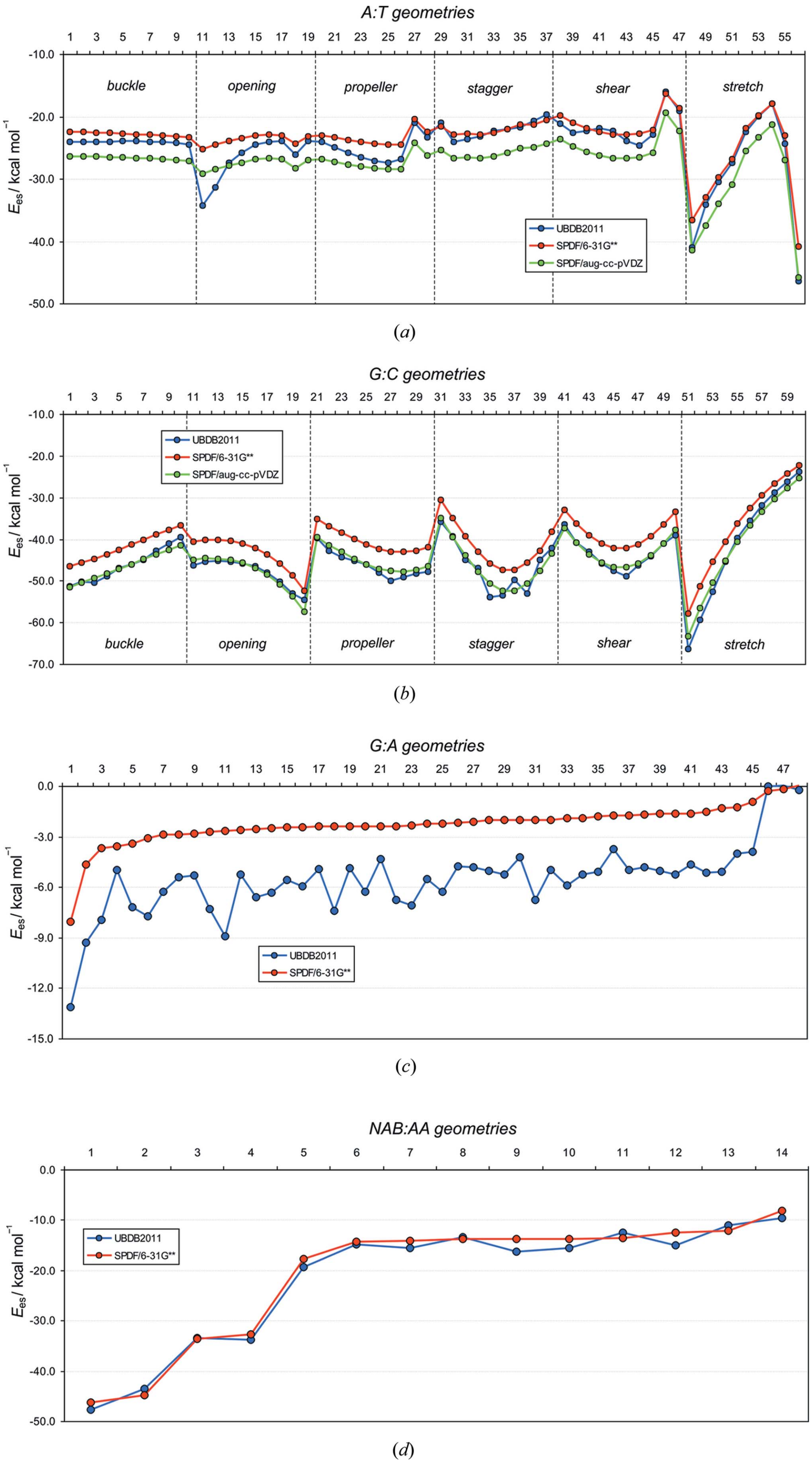


Figure 5
Correlation between SPDFG/B3LYP/6-31G** $E_{\rm es}$ results and UBDB2011 + EPMM method. Total RMSD = 3.7 kcal mol^{-1}. Solid line represents linear least-squares fit (R^2 = 0.99). Dashed–dotted line represents $E_{1,\rm es}$ = $E_{2,\rm es}$ diagonal.



(*a*)

(*b*)

(*c*)

(*d*)

Figure 6
Energy trends. SPDFG and UBDB + EPMM comparison: (*a*) A:T; (*b*) G:C; (*c*) G:A; (*d*) NAB:AA.

6-31G** results. Such an effect might be caused by the superposition of errors in determining the electrostatic interaction energy by the UBDB2011 + EPMM method, which is related to the molecular mutual geometry and interacting atom types. The overestimation of the E_{es} might result from the number of atoms in close contact calculated by the exact potential (EP) method on the basis of not perfectly accurate pseudoatom models. There are also particular random mutual configurations of the interacting nucleic acid bases where the E_{es} significantly differs from the SPDFG value. This can be due to some specific orientations and covering of the multipoles which affect the proper E_{es} estimation (*e.g.* peak at 12th position, Fig. 6*a*). It should be noted that the multipole model might not be flexible enough to properly describe some of the electron-density distribution features, *e.g.* electronic lone pairs, which influences the derived energy values. On the other hand, it is worth stressing that the RMSD values between UBDB2011 + EPMM and SPDFG are comparable to the electrostatic interaction energy differences derived by SPDFG with the use of 6-31G** and aug-cc-pVDZ basis sets (3.9 and 4.6 kcal mol^{-1} for A:T and G:C complexes, respectively).

Even though Figs. 5 and 6 clearly show that the UBDB2011 + EPMM method tends to overestimate the interaction energy when compared to the *ab initio* results, the overall performance of the UBDB2011 databank is rather satisfactory in predicting the electrostatic energy of hydrogen-bonded nucleic acid base pairs. It should be emphasized here that both methods preserve the general tendencies in electrostatic interaction energy values according to mutual geometry variation of the

Table 2
Root mean square deviations (RMSDs) between UBDB2011 + EPMM and SPDFG/B3LYP/6-31G** or SPDFG/B3LYP/aug-cc-pVDZ E_{es} values, respectively.

	RMSD (kcal mol^{-1})	
Complex type	SPDFG/B3LYP/6-31G**	SPDFG/B3LYP/aug-cc-pVDZ
A:T	2.2	4.3
G:C	5.0	1.4
G:A	3.5	
NAB:AA	1.5	

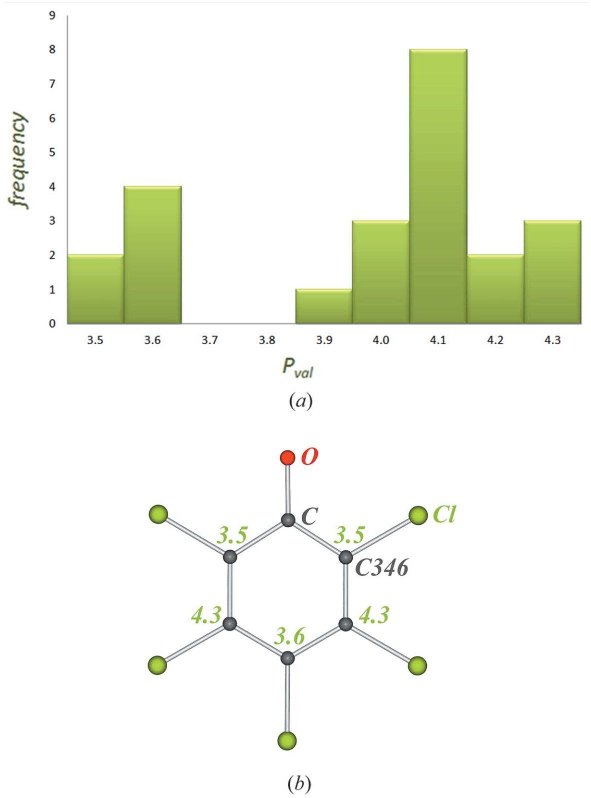
analysed complexes (Fig. 6). The study also showed that in the case of G:A complexes the energetic variability is compatible for both methods (UBDB2011 + EPMM and SPDFG). However, the magnitude of the UBDB2011 + EPMM-derived energy value and its deviation from the theoretical results are comparable. This means that such low E_{es} values obtained for dispersively interacting G:A dimers are statistically meaningless. The UBDB2011 + EPMM method works better for stronger electrostatic interactions around 50–70 kcal mol^{-1}. For the amino-acid species the UBDB2011 + EPMM method exhibits the same accuracy as the previous version of the databank (R^2 = 0.99, RMSD = 1.5 kcal mol^{-1}) (Volkov, Koritsanszky & Coppens, 2004).

To summarize this section, the presented results confirm the sufficiently good quality of the pseudoatom parameters of the newly added atom types. The UBDB2011 + EPMM method can be applied to quantitative electrostatic interaction energy evaluation in the case of macromolecules but only if an accuracy of 4 kcal mol^{-1} will not discredit the results. If, for instance, the interaction energy values of a protein with different inhibitors are similar (*i.e.* differ by about 4 kcal mol^{-1} or less), no meaningful conclusions can be drawn.

3.2. Remarks and limitations

Most of the atom types are satisfactorily transferable and described by means of the Hansen–Coppens formalism. However, in some cases, it is very difficult to define an atom type properly or even impossible to do it in a simple way. There are several factors that influence the atom-type description, much of which depends on the CSD sample. If an atom is usually present in a group of similar molecules, its definition is more consistent and parameter deviations smaller. When a particular atom type is averaged over structurally various and sometimes quite exotic molecules, its definition is statistically less accurate or it requires a special treatment. Some of the atoms are more sensitive to their chemical environment, *i.e.* easier polarized, than others. This is especially true for carbon atoms bonded to nitrogen atoms. Here, the valency and nitrogen-atom neighbours are of great importance.

What is quite straightforward for a chemist, aromatic systems or those containing coupled or alternated double bonds, can be particularly problematic as they have labile π-electrons. Such molecules mark out the limits of atomic transferability. This is easily visible in Fig. 7. The simple example of pentachlorophenolate shows that having equal first, second and even third neighbours does not guarantee the same set of multipole parameters. There is a significant alternation of P_{val} parameter values and all the multipole populations on the carbon atoms with chlorine substituents. The same situation is observed for the chlorine atoms themselves [P_{val}(*o*-Cl) = 7.23, P_{val}(*m*-Cl) = 7.19, P_{val}(*p*-Cl) = 7.31]. This is a result of a chemically known phenomenon explained on the basis of resonance electronic structures. Therefore, even though the chemical environments of aromatic carbon atoms are practically equal, their multipole populations differ significantly as atoms in the *meta* position usually exhibit more negative charges than their *para* and *ortho* equivalents. To precisely describe those atom types, a very complex atom-type definition should be applied. Fig. 8 presents a well described atom deposited in the UBDB2011 databank in contrast to the previously mentioned example.



(*a*)

(*b*)

Figure 7
(*a*) C346 atom type P_{val} values distribution [e.s.d.(P_{val}) = 0.27]; (*b*) P_{val} alternation depending on carbon-atom position in the aromatic ring of pentachlorophenolate. Atom colour coding: oxygen – red, carbon – dark grey, chlorine – green.

Aromatic systems may be very sensitive to substituent effects and therefore very difficult to model. However, the differences in multipole parameters are usually less pronounced than in the case of pentachlorophenolate. The effect of charge-density distribution influenced by some further neighbours can be observed in the case of five- and six-membered rings.

The heterocyclic character of a ring or different substituents among the calculated structures cause the ambiguity in atom-

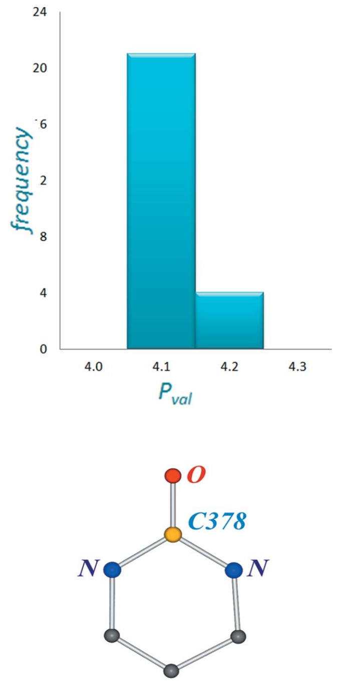


Figure 8
C378 atom type $P_{\rm val}$ values distribution [e.s.d.($P_{\rm val}$) = 0.011]. Atom colour coding: nitrogen – blue, oxygen – red, carbon – dark grey.

type description reflected in higher standard deviations. Carbon atoms C311–C313 also need an improved atom-type definition, being more sensitive to electronic effects due to double bonds. Chemical elements that are easily polarized may pose some problems in the future when further extending the databank. Here, especially in the case of the possible addition of some metal atoms or ions, different approaches should be concerned to model these atoms properly such as, for example, the modification of the atomic core description proposed by Fischer *et al.* (2011). Such phenomena narrow the applicability of the databank to electrostatic energy estimation for more complicated or exotic systems. However, it can still be satisfactorily applied as a source of aspherical atom scattering factors in the TAAM refinement.

4. Conclusions

The UBDB databank was recalculated and extended with atoms required to model RNA and DNA molecules. It currently contains over 200 atom types present in the most relevant biomolecules. New atomic keys were implemented to distinguish the geometry of multipole populations centred on atoms belonging to five- and six-membered planar rings. Atoms common to two fused planar rings, crucial in the case of purine bases, were properly defined. The *LSDB* program was modified so as to provide adequate atom-type definitions, local coordinate systems for atoms joining two aromatic rings together and updated X—H bond distances. It was shown that the UBDB2011 + EPMM method satisfactorily reproduces electrostatic interaction energies for a set of nucleic acid base complexes with respect to *ab initio* results ($R^2 > 0.9$, RMSD = 3.7 kcal mol^{-1}). Correlations are high while energy trends are preserved. The UBDB2011 can therefore be applied to estimate electrostatic interaction energies of macromolecular systems. However, one should be aware of the method's limitations. The databank does not describe conformational variety and does not take into account the crystal field influence and other subtle effects. Consequently, when applied to electrostatic interaction energy evaluation UBDB2011 + EPMM results should be interpreted rather qualitatively than quantitatively. Meaningful comparisons may only be drawn when the electrostatic interaction energy difference for two chemical systems exceeds 5 kcal mol^{-1}. For more precise interaction energy computations and derivation of other electrostatic properties (usually slightly less sensitive to the charge-density model than electrostatic energy values), a new, more sophisticated and powerful model is required (*e.g.* Koritsanszky *et al.*, 2010).

In view of the above, considering the currently achievable experimental data resolution and its quality, the presented approach is still accurate enough for the purpose of structure refinement. Despite its restrictions, the UBDB2011 constitutes a good source of aspherical atomic scattering factors, which may be used for TAAM refinement to enhance the quality of the final molecular structure.

The authors would like to thank Dr Anatoliy Volkov for providing the *SPDFG* program code and Dr Żaneta Czyżnikowska for distributing the analysed chemical system coordinates. Support of this work by the Polish Ministry of Science and Higher Education through the grant Iuventus Plus No. IP2010 0076/70 (PMD) and grant No. NN204 129138 (KNJ) is gratefully acknowledged.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H. & Bruno, I. J. (2010). *Acta Cryst.* B**66**, 380–386.
Bąk, J. M., Domagała, S., Hübschle, C., Jelsch, C., Dittrich, B. & Dominiak, P. M. (2011). *Acta Cryst.* A**67**, 141–153.
Becke, A. D. (1988). *Phys. Rev. A*, **38**, 3098–3100.
Brock, C. P., Dunitz, J. D. & Hirshfeld, F. L. (1991). *Acta Cryst.* B**47**, 789–797.
Buckingham, A. D. (1967). *Adv. Chem. Phys.* **12**, 107–142.
Czyżnikowska, Ż. (2009). *J. Mol. Struct. (Theochem)*, **895**, 161–167.
Czyżnikowska, Z., Góra, R. W., Zaleśny, R., Lipkowski, P., Jarzembska, K. N., Dominiak, P. M. & Leszczynski, J. (2010). *J. Phys. Chem. B*, **114**, 9629–9644.
Czyżnikowska, Z., Lipkowski, P., Góra, R. W., Zaleśny, R. & Cheng, A. C. (2009). *J. Phys. Chem. B*, **113**, 11511–11520.
Dittrich, B., Hübschle, C. B., Luger, P. & Spackman, M. A. (2006). *Acta Cryst.* D**62**, 1325–1335.
Dittrich, B., Koritsanszky, T. & Luger, P. (2004). *Angew. Chem. Int. Ed.* **43**, 2718–2721.
Dittrich, B., Koritsanszky, T., Volkov, A., Mebs, S. & Luger, P. (2007). *Angew. Chem. Int. Ed.* **46**, 2935–2938.
Dittrich, B., McKinnon, J. J. & Warren, J. E. (2008). *Acta Cryst.* B**64**, 750–759.
Dittrich, B., Strumpel, M., Schäfer, M., Spackman, M. A. & Koritsánszky, T. (2006). *Acta Cryst.* A**62**, 217–223.
Domagała, S. & Jelsch, C. (2008). *J. Appl. Cryst.* **41**, 1140–1149.

Dominiak, P. M., Volkov, A., Dominiak, A. P., Jarzembska, K. N. & Coppens, P. (2009). *Acta Cryst.* D**65**, 485–499.

Dominiak, P. M., Volkov, A., Li, X., Messerschmidt, M. & Coppens, P. (2007). *J. Chem. Theory Comput.* **3**, 232–247.

Dunning, T. H. (1989). *J. Chem. Phys.* **90**, 1007–1023.

Dupuis, M., Rys, J. & King, H. F. (1976). *J. Chem. Phys.* **65**, 111–116.

Faerman, C. H. & Price, S. L. (1990). *J. Am. Chem. Soc.* **112**, 4915–4926.

Fischer, A., Tiana, D., Scherer, W., Batke, K., Eickerling, G., Svendsen, H., Bindzus, N. & Iversen, B. B. (2011). *J. Phys. Chem. A*, doi:10.1021/jp2050405.

Frisch, M. J. *et al.* (2004). *GAUSSIAN03.* Version C. 02. Gaussian Inc., Pittsburgh, Pennsylvania, USA.

Hansen, N. K. & Coppens, P. (1978). *Acta Cryst.* A**34**, 909–921.

Jelsch, C., Guillot, B., Lagoutte, A. & Lecomte, C. (2005). *J. Appl. Cryst.* **38**, 38–54.

Koritsanszky, T., Volkov, A. & Chodkiewicz, M. (2010). *Structure and Bonding*. Berlin, Heidelberg: Springer-Verlag. (doi:10.1007/430_2010_32.)

Koritsanszky, T., Volkov, A. & Coppens, P. (2002). *Acta Cryst.* A**58**, 464–472.

Krishnan, R., Binkley, J. S., Seeger, R. & Pople, J. A. (1980). *J. Chem. Phys.* **72**, 650.

Lecomte, C., Guillot, B., Jelsch, C. & Podjarny, A. (2005). *Int. J. Quantum Chem.* **101**, 624–634.

Lee, C., Yang, W. & Parr, R. G. (1988). *Phys. Rev. B*, **37**, 785–789.

Li, X., Volkov, A. V., Szalewicz, K. & Coppens, P. (2006). *Acta Cryst.* D**62**, 639–647.

Møller, C. & Plesset, M. S. (1934). *Phys. Rev.* **46**, 618–622.

Morokuma, K. (1971). *J. Chem. Phys.* **55**, 1236–1244.

Perdew, J. P. (1986). *Phys. Rev. B*, **33**, 8822–8824.

Pichon-Pesme, V., Jelsch, C., Guillot, B. & Lecomte, C. (2004). *Acta Cryst.* A**60**, 204–208.

Pichon-Pesme, V., Lecomte, C. & Lachekar, H. (1995). *J. Phys. Chem.* **99**, 6242–6250.

Rys, J., Dupuis, M. & King, H. F. (1983). *J. Comput. Chem.* **4**, 154–157.

Schmidt, M. W., Baldridge, K. K., Boatz, J. A., Elbert, S. T., Gordon, M. S., Jensen, J. H., Koseki, S., Matsunaga, N., Nguyen, K. A., Su, S. J., Windus, T. L., Dupuis, M. & Montgomery, J. A. (1993). *J. Comput. Chem.* **14**, 1347–1363.

Volkov, A., King, H. F. & Coppens, P. (2006). *J. Chem. Theory Comput.* **2**, 81–89.

Volkov, A., King, H. F., Coppens, P. & Farrugia, L. J. (2006). *Acta Cryst.* A**62**, 400–408.

Volkov, A., Koritsanszky, T. & Coppens, P. (2004). *Chem. Phys. Lett.* **391**, 170–175.

Volkov, A., Li, X., Koritsanszky, T. & Coppens, P. (2004). *J. Phys. Chem. A*, **108**, 4283–4300.

Volkov, A., Macchi, P., Farrugia, L. J., Gatti, C., Mallinson, P., Richter, T. & Koritsanszky, T. (2006). *XD2006 – A Computer Program for Multipole Refinement, Topological Analysis of Charge Densities and Evaluation of Intermolecular Energies from Experimental or Theoretical Structure Factors*.

Volkov, A., Messerschmidt, M. & Coppens, P. (2007). *Acta Cryst.* D**63**, 160–170.

Zarychta, B., Pichon-Pesme, V., Guillot, B., Lecomte, C. & Jelsch, C. (2007). *Acta Cryst.* A**63**, 108–125.

Ziegler, T. & Rauk, A. (1977). *Theor. Chim. Acta*, **46**, 1–10.