Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures

Miha Andrejašič, Jure Pražnikar and Dušan Turk*

Department of Biochemistry and Molecular and Structural Biology, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

Correspondence e-mail: dusan.turk@ijs.si

Received 25 April 2008
Accepted 25 August 2008

The number and variety of macromolecular structures in complex with 'hetero' ligands is growing. The need for rapid delivery of correct geometric parameters for their refinement, which is often crucial for understanding the biological relevance of the structure, is growing correspondingly. The current standard for describing protein structures is the Engh–Huber parameter set. It is an expert data set resulting from selection and analysis of the crystal structures gathered in the Cambridge Structural Database (CSD). Clearly, such a manual approach cannot be applied to the vast and ever-growing number of chemical compounds. Therefore, a database, named PURY, of geometric parameters of chemical compounds has been developed, together with a server that accesses it. PURY is a compilation of the whole CSD. It contains lists of atom classes and bonds connecting them, as well as angle, chirality, planarity and conformation parameters. The current compilation is based on CSD 5.28 and contains 1978 atom classes and 32 702 bonding, 237 068 angle, 201 860 dihedral and 64 193 improper geometric restraints. Analysis has confirmed that the restraints from the PURY database are suitable for use in macromolecular crystal structure refinement and should be of value to the crystallographic community. The database can be accessed through the web server http://pury.ijs.si/, which creates topology and parameter files from deposited coordinates in suitable forms for the refinement programs *MAIN*, *CNS* and *REFMAC*. In the near future, the server will move to the CSD website http://pury.ccdc.cam.ac.uk/.

1. Introduction

The paper of Engh & Huber (1991), with its description of accurate geometrical parameters of amino-acid residues, has provided a foundation for the use of geometrical restraints in the refinement of protein structures. A similar step forward in the area of nucleic acids was made by Parkinson *et al.* (1996). However, these efforts have no counterpart for so-called 'hetero' compounds. As a result, the structures of small molecules found in complexes with biomacromolecules are often less reliable than those of the surrounding amino acids and nucleic acid bases. The reason is most probably the essentially boundless structural diversity of small molecules as opposed to the limited numbers of building blocks of proteins and nucleic acids (Kleywegt *et al.*, 2003). These small molecules are physiological ligands, cofactors, lead compounds, substrate analogs *etc.* and the accuracy of their structures can be of crucial importance for interpreting their (potential) biological roles.

Several attempts have been made to fill the gap between the geometric parameters for macromolecules and small mole-

cules. There are a few software packages that are capable of generating topological descriptions, with their corresponding geometric restraints, for the refinement of hetero molecules, such as *PRODRG* (Schüttelkopf & van Aalten, 2004), *SMILES2DICT* (Greaves *et al.*, 1999)/*LibCheck* (Vagin *et al.*, 2004), *HicUp*/*XPLO2D* (Kleywegt & Jones, 1998), *CORINA* from Molecular Networks GmbH and *AFITT* (Wlodek *et al.*, 2006). Common to all these is the use of parameters with a predefined set of atom classes originating from various force fields and the use of selections of published values of bond distances and angle values. The program *Hess2FF* is an attempt to construct restraints on a purely theoretical basis (Nilsson *et al.*, 2003), whereas *eLBOW*, part of *PHENIX* (Adams *et al.*, 2002), is a kind of amphibian that uses combined empirical and theoretical approaches.

Nevertheless, after the Engh and Huber parameter set for amino-acid residues had been elaborated by analysis of the crystal structures of small molecules determined at high accuracy and deposited in the Cambridge Structural Database (CSD; Allen, 2002), it became clear that a new standard had been defined for future developments of geometrical restraints for use in refining macromolecular structures. Averaged values for each parameter analyzed were defined as the target values and standard deviations were used to define the force-constant values. In this way, the expected variation in the parameters determined the force constants rather than the physical force (Engh & Huber, 1991).

The number and variety of macromolecular structures of complexes with hetero ligands is growing. Only 12% of the hetero structures deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000) have an exact match in the CSD (R. Taylor, personal communication). Therefore, it is important to generate a parameter set that will provide parameters for existing and emerging compounds with an accuracy and precision approaching those of the Engh–Huber and Parkinson parameter sets. Since such a parameter set has to cover the vast diversity of existing and emerging chemical space, it must be able to extend over thousands of atom classes and parameters connecting them. It is clear that such a set can only be reliably constructed, maintained and updated in an automated manner. With these goals in mind, the PURY database was constructed as an automatically generated library of geometric parameters for the refinement and validation of hetero compounds based on high-resolution crystal structures. PURY parameters can be used for the refinement and validation of the geometry of not only hetero compounds but also amino-acid and nucleic acid residues and whatever else comes its way. Here, we describe the current features, accuracy, limitations and an outline of the development of PURY. The accompanying server makes the database available to the crystallographic community.

2. Methods

To describe the interactions between atoms in terms of lists of target values and force constants for bond, angle and conformational restraints, atom classes were assigned to individual atoms depending on the kind of fragment to which they belong. Fragments have been derived from a common understanding of chemical structure, exploiting the CSD (Allen, 2002).

2.1. Energy terms

The terms of geometric restraints comprise bonding terms (bond distances and angles) and improper and dihedral angles. The bonding-angle, improper and dihedral terms are compatible with most refinement programs, including *X-PLOR* (Brünger *et al.*, 1987), *MAIN* (Turk, 1992), *CNS* (Brünger *et al.*, 1998), *PHENIX* (Adams *et al.*, 2002) and *REFMAC* (Murshudov *et al.*, 1997). Bond angles are also alternatively analyzed as angle distances for compatibility with the *SHELX* refinement programs (Sheldrick, 2008). Energy terms describing nonbonding interactions have not been considered. Their values have been assigned in accordance with the *X-PLOR* TOP_19 parameter set (Brünger *et al.*, 1987).

All terms, apart from the dihedral angles describing rotations about single bonds (freely rotatable bonds), use the quadratic form of the restraint function,

$$E = k(g - g_t)^2, \tag{1}$$

where E is the energy of the term, k is its force constant, g is the geometric value of the term and g_t is its target or ideal value. The force constant is specific for each specific restraint. It is derived from the σ value of the geometric distribution using the distribution law in the form used by Engh & Huber (2001),

$$k = 0.592/\sigma^2. \tag{2}$$

There is a slight difference between software packages in whether they use the σ value or the force constant as a restraint, although these are connected *via* the above equation. *SHELX*, *REFMAC* and *PHENIX* use estimated standard deviations or σ of bond lengths and bond angles in their formulation of restraints, rather than the force constant as is used in *MAIN*, *CNS* and *X-PLOR*.

For the dihedral angles, the quite commonly used periodic function the cosine term is used,

$$E = E_0[1 - \cos(n\varphi - \delta)], \tag{3}$$

where E is the energy of the term, E_0 is the energy barrier, φ is the geometric value, δ is the target value and n represents the periodicity of the dihedral term, usually 2 for *cis* or *trans* configurations and 3 for freely rotatable bonds. In the proximity of the equilibrium, the cosine function mimics the quadratic form (as used for the bonding term) quite well.

The force constant and energy barrier in the case of dihedrals and the geometrical target value are specific for each particular combination of atom classes involved.

2.2. Creation of connectivity

Generation of the atom classes starts from the list of bonding connections. It is followed by the assignment of atomic states, after which the *PURY* algorithm assigns atom

names, from which the database of geometric restraints is generated. The connectivity list is read from the structure file. In the case of its absence, the list is calculated from the overlap of van der Waals radii (http://www.ccdc.cam.ac.uk/products/csd/radii; Allen *et al.*, 1979),

$$d_{\rm bond} = R_1 + R_2 + 0.45\ \text{Å}. \quad (4)$$

2.3. Assignment of atomic states

The hybridization states of atoms are derived from the bonding parameters, taking into account neighbours and their geometrical arrangement, including bond lengths and angles. For example, four neighbours for a C atom define sp^3 hybridization. Three neighbours for a C atom in a tetrahedral arrangement define sp^3 hybridization, whereas a planar arrangement defines sp^2 hybridization. Such sp^3 atoms are checked for chirality. For C atoms with two neighbours, a nonlinear arrangement specifies sp^3 hybridization if the angle falls below the sp^2 threshold, otherwise sp^2 hybridization is assumed. A linear arrangement suggests that the C atom is involved in either a triple bond or two double bonds, thus specifying sp^1 hybridization. It should be noted that for atoms with only one covalent bond, the neighbouring atom and the bond length are the only source of information about its chemical environment.

The next state of atoms is their involvement in rings. Only ring structures with up to 12 members are considered. Ring structures are divided into planar and nonplanar systems in accordance with their aromaticity. When the maximal distance of a ring member from the LSQ plane of the rings remains below a cutoff value (0.1 Å) the ring is considered planar; otherwise it is considered nonplanar. Exceptionally, a ring is considered to be planar (aromatic) when all ring members are sp^2 hybrids. For a chain of non-ring aromatic systems, planarity is not explicitly considered. Atoms in such chains are evaluated as sp^2 hybrids.

2.4. PURY atom-class assignment

The atom-class code is meant to be human-readable. Each atom class is composed of four ASCII characters to ensure compatibility with the existing software. (The meaning of the characters is occasionally position- and context-dependent in order to compensate for the limitation of the length of the class code.) The resulting atom-class algorithm is rather branched, using numerous conditions. Below, only the basic rules are presented.

The generation of classes for organic elements (C, H, N, O, P, and S) is different from that for other elements in order to accommodate the greater diversity of the compounds. For organic elements only the first position is reserved for the element symbol, whereas for other elements the first two positions are used. In the cases of two-letter chemical element symbols, the second position is always written in lower case, whereas for inorganic elements with a single-letter symbol the second position is '_'.

The second position of an organic element class with only one bonded atom, apart from hydrogen, shows the double or triple character of the bond, which are represented by the numbers '2' and '3', respectively. A single bond is represented by '_'. Positions three and four are reserved for the bonded atoms: 'O2C_' denotes the O atom of a carbonyl group, whereas 'N3C_' represents a cyano-group nitrogen.

The second character of a class of organic atom with more than one bond usually contains the character 'H' followed by the number of attached H atoms in the third position. 'OH1<' represents a hydroxyl group, 'CH3X' a methyl group and 'CH_6' a phenyl C atom with one H atom attached.

The fourth character describes the geometric arrangement of the neighbours of an atom. Non-ring atoms are described with the characters 'X', 'Y', 'I' and '<'. 'X' denotes four valences, including sp^3 hybrids such as ammonia (NH_3), which has one free electron pair. 'Y' denotes a planar sp^2 hybrid such as that present in the amide group. 'I' and '<' indicate arrangements around an atom with two neighbours. 'I' describes the linear arrangement of a C atom involved in one triple or two double bonds, whereas '<' describes an arrangement at an angle (an sp^3 hybrid) such as that for the S atom in methionine 'S__<'.

For aromatic ring systems the fourth character describes the size of the ring, whereas in the case of non-aromatic rings the size is written in the second position. The code for the ring can accept up to 12 members, extending the code beyond '9' by the use of '0', 'A' and 'B'. Rings with more than 12 members are however not considered as rings. For example, atom class 'CC_6' describes a phenyl ring C atom attached to a C atom outside the ring. The bridge atoms between two rings are marked with an asterisk in either the second or the fourth position. The fourth character can also describe certain exceptions such as the guanidinium group, for which the character 'G' is used.

In the first row of Fig. 1, three quite common fragments are used to illustrate the coding algorithm on a non-ring system, whereas the second line illustrates three atom classes from

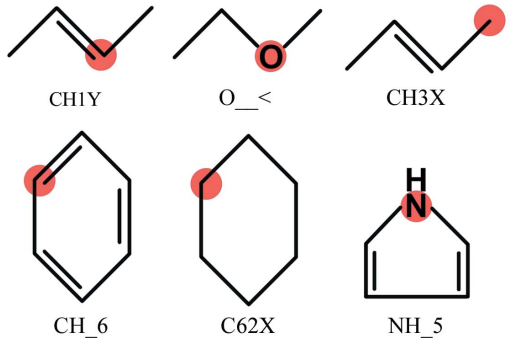

Figure 1
Demonstration of PURY atom classes for six common fragments. The atom with the assigned class is marked with a red dot. From left to right and from top to bottom: CH1Y, sp^2-hybridized C atom forming one double bond with a bonded H atom; O__<, ether O atom; CH3X, sp^3-hybridized C atom with three bonded H atoms (methyl group); CH_6, benzene C atom with a bonded H atom; C62X, cyclohexane C atom with two bonded H atoms; NH_5, pyrrole N atom with a bonded H atom.

ring systems. Fig. 2 demonstrates the procedure of atom-class assignment in the case of 4-chlorophenyl acetic acid.

In the case of metal atoms involved in coordinate bonds, metals with bonding partners of the same kind are represented by 'Ca6_', where '6' indicates that the calcium ion has six coordinated atoms. Elements from the fourth and fifth periodic rows can form metallo acids. They are denoted by 'O' at position three followed by the number of O atoms, for example 'MoO5'. Coordinate bonds are not considered as part of the atom-class assignment procedure of organic compounds.

2.5. Generation of a database of geometric restraints

Storage lists of all possible covalent and coordinate bond and angle terms are generated from the connectivity list. A storage list comprises a list of all appearances of each combination of atom classes for each particular geometric term. In addition to atom names and classes, storage-list entries also contain the CSD reference code. The storage list of bonds thus contains pairs of atoms, the storage list of angles contains combinations of all bonded atoms and the storage list of improper terms is generated from all atoms bonded to three or more neighbours, among which at least three non-H atoms should be present. The values of improper terms are always set to be positive, which means that the *PURY* analysis does not differentiate between *R* and *S* chiral centres. The storage lists of dihedral angles are generated by storing every possible combination of neighbours of all bonded atoms. Those containing H atoms at both ends are excluded.

Each restraint is derived from its own storage list. The target value is the average value of the storage-list members, whereas the force constants reflect the standard deviation of the assembled values on the list, as described above. Exceptions are bond terms with only a single repeat, for which σ is set to 0.066 Å (3 average σ), whereas for angle and improper terms with only single occurrence σ has been set to 5°.

In deriving the dihedral angle restraints, the storage list of dihedral angles is assigned to 36 bins, each with a 10° span. The planarity restraints are detected by inspection of the shells spanning from −180° to −170°, −10° to 0°, 0° to 10° and 170° to 180°. When both inner atom classes show a planar nature, the periodicity is 2 and the target value is assigned as 0° or 180° in accordance with the higher occupancy of the shells. For the freely rotatable bonds the periodicity is set to 3 and the target value to 60°, while the corresponding energies are calculated from evaluation of the shell occupancy distribution.

The parameters derived using the above equations and procedures are shown in Table 1, where examples of the selected bond, bond angle, dihedral angle and improper angle restraints are presented.

3. Results

The atomic coordinates of the structures used to create the PURY database were extracted from CSD v.5.28. The structures were selected using the CSD *ConQuest* (Bruno *et al.*, 2002) browser tool using as filters 'no errors' and a crystallographic *R* value below 5%. A total of 162 540 entries contained almost ten million atom positions, 10 529 799 bond lengths, 20 342 046 bond angles, 2 703 482 improper terms and 8 440 410 dihedral angles. From these data, 1971 different atom classes were derived and parameter lists yielding 32 634 bond lengths, 236 821 bond angles, 64 133 improper terms and 229 821 dihedral angle restraints were generated. The numbers reporting the amount of the data on the web server may differ since PURY is constantly evolving.

From the total of 1971 atom classes, 270 were assigned to carbon, which is obviously the most chemically versatile and frequently appearing chemical element. The second on the list is nitrogen with 159 classes, followed by phosphorus with 136. S and O atoms, represented by 88 and 69 classes, respectively, complete the group of 'organic' atoms. The halogens F, Cl, I and Br are represented by 51, 37, 29 and 27 atom classes.

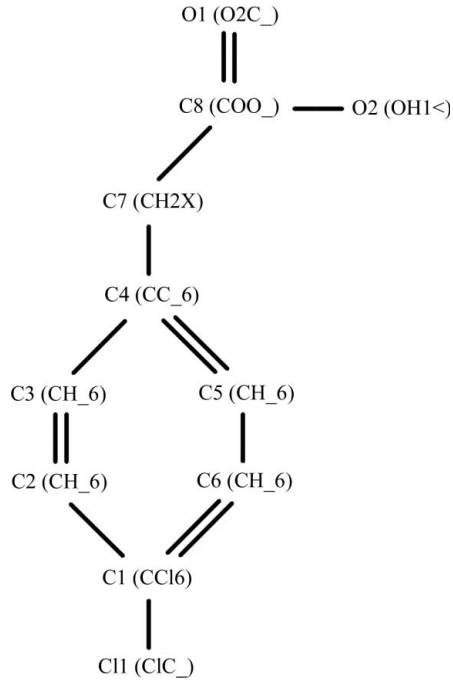

Figure 2
A sample description of how the *PURY* algorithm evaluates 4-chlorophenyl acetic acid ($C_8H_7O_2Cl$; CSD reference AHATAE). Atoms are labelled with their name and their derived atom class in parentheses. The ring consists of six atoms (C1, C2, C3, C4, C5 and C6) and is found to be planar. All ring atoms are also planar so it is considered to be aromatic. The C atoms receive a '6' at position four since they are all members of a six-membered aromatic ring. Ring atoms with bonded H atoms (C2, C3, C5 and C6) receive 'H' at position two, while other atoms are given the name of the bound atom at positions two and three: 'Cl' for chlorine (C1) and 'C_' for carbon for C4. The methylene C atom (C7) is assumed to have two H atoms bonded and since the bond distance between atoms C4 and C7 corresponds to a single bond and the angle is below the sp^2 threshold, it receives the class name 'CH2X'. The carbonyl C atom (C8) has two bonded O atoms and is assigned the special class 'COO_'. The O atom (O2) with a bonded H atom forms two single bonds and receives the class name 'OH1<', while the other O atom (O1) forms only a single bond with the C atom and its distance suggests a double bond. Since oxygen is organic, its bond type is written at position two. C_ is written at the third and fourth places. The Cl atom (Cl1) forms only one covalent bond so it is a bonding atom, which is a single-character organic element and is written at position three. The atom receives the class name 'ClC_'.

Table 1
Output example for bond, angle, improper and dihedral terms for use in macromolecular refinement.

The output shows atom classes, equilibrium values, corresponding force constants and multiplicity values where appropriate. The σ values from which force values were calculated are also shown.

Entry	Class 1	Class 2	Class 3	Class 4	Force constant	Multiplicity	Average value	σ
Bond	CH2X	S__<			1643.6		1.818 Å	0.0190 Å
Angle	CH2X	S__<	CH3X		733.8		100.9°	1.6274°
Improper	CC_6	CH_6	CH_6	CH2X	1211.3	0	0.00°	1.123°
Dihedral	CH2X	CH2X	S__<	CH3X	7.41	3	90.0°	20.000°

Table 2
Relative distribution of appearances of bond, angle and improper angle parameters generated with *PURY*.

No. of appearances	Bonds (%)	Angles (%)	Impropers (%)
1	6.90	20.86	36.88
Below 5	38.87	37.63	37.05
Below 10	14.64	14.51	10.05
Below 30	17.77	14.44	8.90
Below 100	11.23	7.60	4.34
Below 1000	8.41	4.24	2.41
Over 1000	2.18	0.72	0.39
Over 30	21.82	12.56	7.13

There are 18 Si classes and two hydrogen classes. In addition, there are 1091 atom classes representing the remainder of the periodic table.

3.1. Assessment of the data: quality and reliability

The prime measure of the reliability of a statistical parameter is its number of repetitions. Table 2 shows the number of terms represented by over 1000, 100, 30 and fewer individual values. Only a tiny proportion of the parameters (2.2% of bonds, 0.7% of angles and 0.4% of impropers) are really accurately described and even then there are a few exceptions, as shown below. In general, the parameters extracted from more than 30 representatives appear to be statistically reliable (at 5% reliability), which corresponds roughly to 15% of all parameters (21.8% of bonds, 12.6% of angles and 7.1% of impropers). A substantial number of the entries are the result of a single observation (6.9% of bonds, 21% of angles and 36.9% of impropers). This table suggests that the standard deviation and the target value of a substantial number of terms are not reliable and their σ values were therefore adjusted to reasonable values.

Fig. 3 illustrates the connection between accuracy, precision and number of repetitions. Three covalent-bond cases have been chosen, one from each population class: the first case ('CF_6–F_C_'; F atom bonded to C atom in a six-membered aromatic ring) was highly populated, the second case ('CH1X–O__<'; aliphatic carbon–oxygen bond; ether) was moderately populated and the third case ('CO_5–CCl5'; C atoms from a five-membered aromatic ring, the first with an O atom attached and the second's covalent partner outside the ring being a Cl atom) was sparsely populated. The peak value of the first case has over 1000 repetitions and corresponds to the average of 1.347 Å, while the peak value of the second case with over 90 repetitions corresponds to the average value of 1.437 Å. In the third case the highest peak has only three

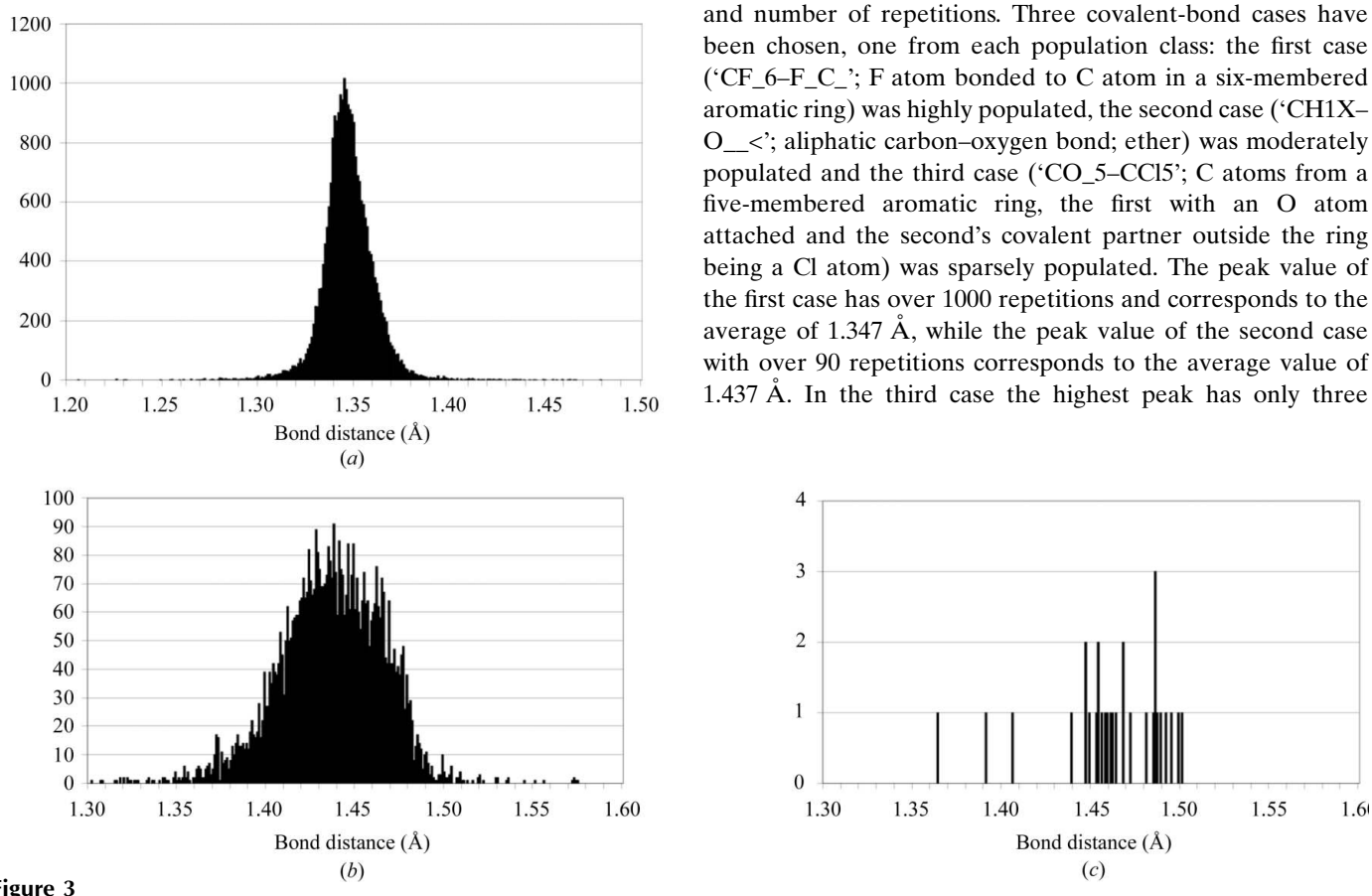

Figure 3
Histograms of bond distances between selected atom classes. (*a*) Bond between an sp^2-hybridized C atom in a six-membered aromatic ring 'CF_6' and an F atom 'F_C_'. (*b*) Bond between an sp^3-hybridized C atom with one bonded H atom 'CH1X' and an sp^3-hybridized O atom 'O__<'. (*c*) Bond between an sp^2-hybridized C atom in a five-membered aromatic ring with a bonded O atom 'CO_5' and an sp^2-hybridized C atom in a five-membered aromatic ring with a bonded Cl atom 'CCl5'.

repetitions at 1.487 Å and does not really correspond to the average at 1.462 Å. The precision of the terms is reflected in their σ values, which are 0.0147, 0.032 and 0.031 Å for the first, second and third case, respectively. The shape of the first histogram gives the impression of a highly accurate term. For the histogram of the third term it is obvious that the term is underrepresented, although its minimum and maximum bond lengths are closer than those of the other two terms; even the standard deviation is sharp, suggesting that its precision may not differ much from that of the middle term. This is a warning that indicates a more general phenomenon. Transfer of geometric restraints from similar fragments to unknown structures in the absence of the corresponding experimental structure or statistical validation of the term may be less reliable than anticipated. Only 12% of hetero compounds have a matching structural deposit in the CSD (R. Taylor, personal communication).

The cases presented in Fig. 4 illustrate the behaviour for bonding, dihedral and improper angles. The bond angle of an amide fragment ('CH1X–C__Y–NH1Y'; Fig. 4*a*) involved in a peptide bond has a clearly defined maximum peak which corresponds to the average value of 116.3° and a smaller peak positioned the other side of 120°. The double peak of the bonding angle is reminiscent of proline-residue analysis, for which coupling has been observed between the bonding and dihedral angles (Lamzin *et al.*, 1995; Engh & Huber, 2001); however, the population of the lower peak is too low to allow firm conclusions to be drawn. The amide-fragment dihedral (CH1X–NH1Y–C__Y–O2C) has the major peak at 0° and a much less populated peak at 180° (Fig. 4*c*), an indication of the appearance of *trans* and *cis* amide (peptide) bond conformations. The freely rotatable bond around the two sp^3-hybridized C atoms, dihedral CH1X–CH1X–CH1X–CH1X (Fig. 4*d*), exhibits a distinct peak at 0° in addition to a period of 3, indicating the presence of the *cis* conformation. The absence of peaks at ±120° suggests that a period of 6 is not justified, whereas the absence of any width of the 0° peak is indicative of the use of constraints during structure refinement. This case illustrates that dihedral angle terms describing conformations of freely rotatable single bonds may not be represented well by the ideal value. Alternatively, one may use an all-atom model with all H atoms included or not use any specific term for such dihedral angles (the energy barrier is set to 0) and use the 1–4 nonbonding interaction terms instead (as in the *X-PLOR* TOP_19 parameter set). PURY can deliver restraints for both types of restraint.

Improper angles, on the other hand, exhibit the least ambiguous behavior. In 'CC_6–CH2X–CH_6–CH_6' (Fig. 4*b*) the planarity of the CG atom of residues such as Tyr or Phe has a peak at 0° with a σ of 1.0°. (It needs to be emphasized, however, that the histograms of data for improper restraints are symmetric as a consequence of the way they are sampled.)

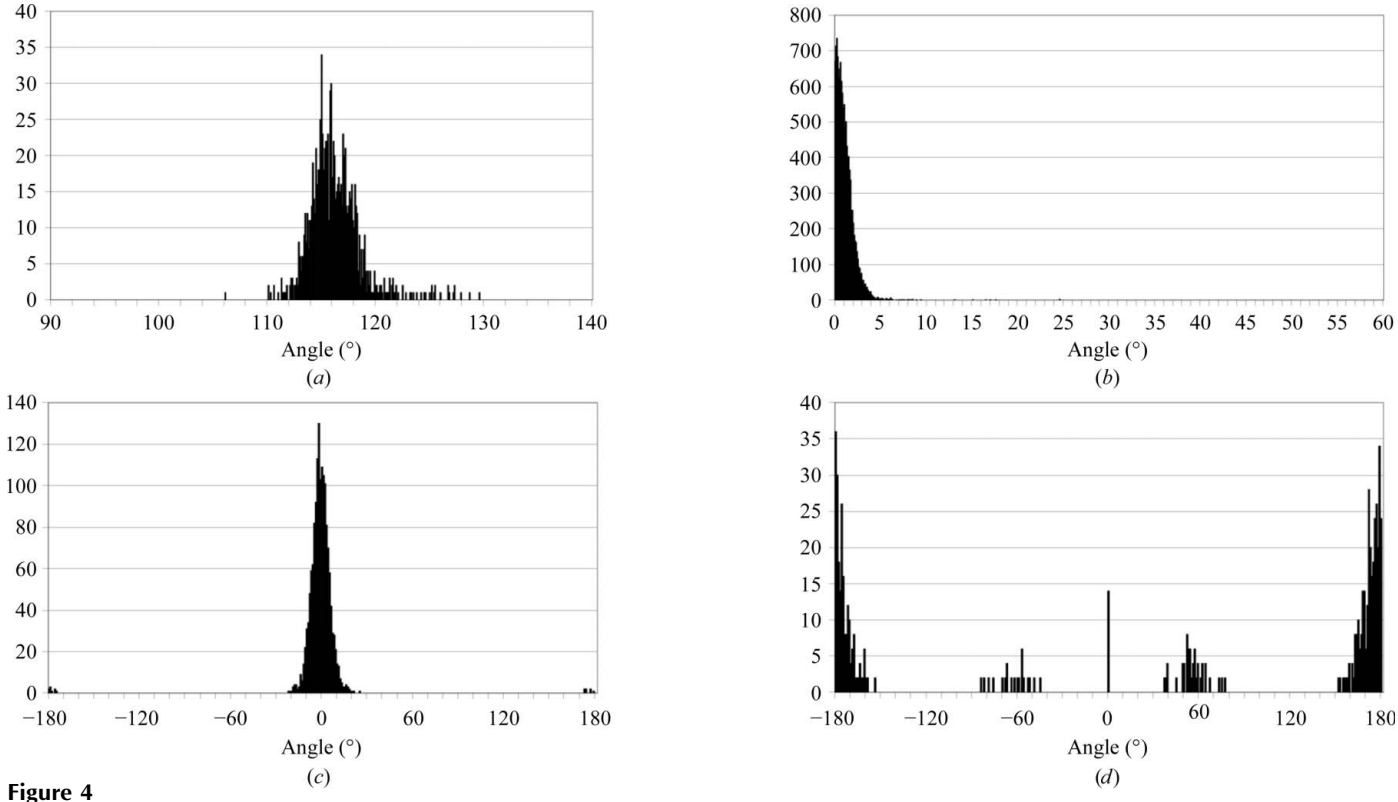

Figure 4
Histograms of selected parameters. (*a*) A peptide-bond angle including C^{α} atom 'CH1X', an sp^2-hybridized planar C atom 'C__Y' and an sp^2-hybridized planar N atom with one bonded H atom 'NH1Y'. (*b*) A tyrosine- or phenylalanine-like improper angle around an sp^2-hybridized C atom in a six-membered aromatic ring 'CC_6' including a C^{β} atom 'CH2X' and two sp^2-hybridized C atoms in a six-membered aromatic ring with bonded H atoms 'CH_6'. (*c*) A dihedral angle through a planar peptide bond having only a single peak 'CH1X–NH1Y–C__Y–O2C_'. (*d*) A dihedral including four sp^3-hybridized C atoms with a freely rotatable single middle bond which has many energy minima 'CH1X–CH1X–CH1X–CH1X'.

Table 3
Bond lengths of H atoms from neutron-derived structures.

We have selected only the terms represented by more than 100 repeats. Columns 1 and 2 show the PURY atom classes forming the bond. Column 3 shows the average bond length obtained from neutron data, whereas the values in parentheses show the average from the whole CSD. Column 4 shows the corresponding σ values, while column 5 shows the ratios between the whole CSD parameter and the neutron-derived parameter. Column 6 shows the ratios between the number of representatives from the whole CSD and the number of neutron data representatives.

PURY class 1	PURY class 1	Average (all data) (Å)	σ values (Å)	σ ratio (all/neutron)	Repeat ratio (all/neutron)
HC__	CH3X	1.073 (0.971)	0.043	0.99	309
HC__	CH_6	1.079 (0.956)	0.034	1.33	329
HC__	CH2X	1.091 (0.980)	0.040	1.07	352
HP__	OH1<	1.020 (0.889)	0.084	1.26	69
HP__	OH2<	0.961 (0.875)	0.042	2.18	113
HC__	C62X	1.089 (0.983)	0.036	1.25	473
HP__	NH3X	1.027 (0.915)	0.031	2.19	95
HC__	CH_5	1.074 (0.956)	0.029	1.65	447
HP__	NH2Y	1.000 (0.889)	0.035	2.15	68
HC__	CH1X	1.096 (0.984)	0.036	1.19	223
HC__	C*2X	1.096 (0.985)	0.012	3.86	297
HC__	C52X	1.083 (0.981)	0.026	1.72	785
HC__	C*1X	1.097 (0.985)	0.013	3.56	380
HC__	C61X	1.099 (0.987)	0.008	5.53	239
HP__	B__X	1.238 (1.100)	0.071	1.52	52
HC__	CH4X	1.070 (0.967)	0.056	0.97	308
HC__	CH1Y	1.079 (0.958)	0.055	0.96	361

3.2. Lengths of covalent bonds involving H atoms

Fig. 5 shows the distributions of bond lengths of an H atom bound to an O atom (hydroxyl group), to a phenyl ring carbon, to an amide and to an aliphatic ring carbon. The presence of several high narrow peaks on the background of an extremely broad distribution of bond lengths ranging from 0.7 Å with several outliers even beyond 1.3 Å in all four cases demonstrates that the positioning of H atoms is ambiguous. The CSD does not provide experimental data to verify in which cases the positioning and refinement of H-atom positions is supported by diffraction data; the CSD filters do not allow us to analyze the reliability of the structures to the resolution at which they were determined. However, they do allow structures to be chosen according to the radiation source. Fig. 6 shows the equivalent bond lengths of structures determined only by neutron diffraction. There were 920 such deposits in the analyzed release, which contained 33 339 atoms and 66 540 covalent bonds. In a total of 1692 bond-length parameters from the neutron data, there are 121 cases of bonds with H atoms. In contrast to the whole set of CSD structures analyzed, the most populated bond length of H atoms from neutron structures only exhibits a single peak. The positions of these peaks are marked on the corresponding figures derived from the complete CSD (Fig. 5). The marked peaks represent only a

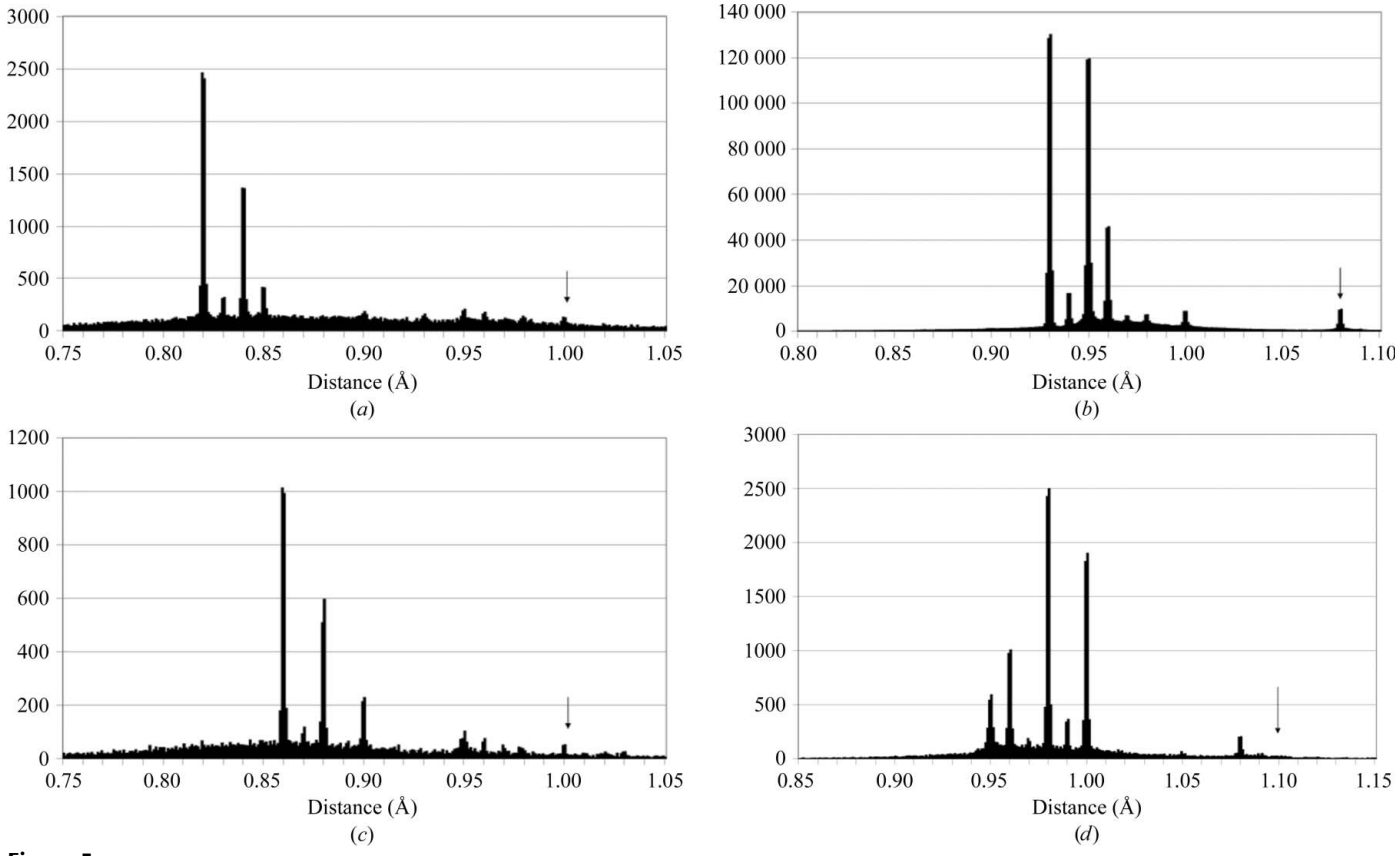

Figure 5
Histograms of bond distances between H atoms. (*a*) The bond between an sp^3-hybridized O atom with one bonded H atom 'OH1<' and an H atom 'HP__'. (*b*) The bond between an sp^2-hybridized C atom in a six-membered aromatic ring with one bonded H atom 'CH_6' and an H atom 'HC__'. (*c*) The bond between an sp^2-hybridized planar N atom with one bonded H atom 'NH1Y' and an H atom 'HP__'. (*d*) The bond between an sp^3-hybridized C atom in a six-membered non-aromatic ring with one bonded H atom 'C61X' and an H atom 'HC__'. The arrows mark the peaks obtained from structures determined by neutron radiation.

minor fraction of the whole data set. However, these peaks are much more highly populated than those obtained by analysis of the structures determined by neutron diffraction only. This reveals that only a small fraction of the H-atom bond lengths of structures determined using X-rays are consistent with those determined using neutrons. Hence, most hydrogen-bond lengths present in the CSD are the consequence of the use of preset values during refinement and these presets differ from the real values obtained by neutron diffraction. In general they are 0.1 Å too short (Table 3). Also, the distribution of bond lengths is substantially narrower for the neutron data, as seen from column 4. Unfortunately, the terms from the neutron diffraction data do not make it possible to replace most of the X-ray-derived terms for hydrogen-bond lengths, so the current state of the art of PURY takes into account all experimental data but includes cases from Table 3 which are represented by over 100 repetitions. Clearly, this is only our current solution.

3.3. The PURY web server

The database is accessible through the WWW interface (http://pury.ijs.si), which enables geometric restraint parameters for three-dimensional structures of molecules or fragments to be downloaded. A user has to upload the coordinate file of the three-dimensional model or submit a SMILES string and download the resulting geometric restraints and topology files from the server.

Currently, only the PDB format for three-dimensional molecular structures is supported for upload. Alternatively, the starting geometry of the compound may be created on the server by the interactive three-dimensional graphical program *JME* (P. Ertl, Novartis; http://www.molinspiration.com/jme/). The output topology and parameter files are in formats readable by the *MAIN*, *X-PLOR*/*CNS*, *REFMAC* and *PHENIX* macromolecular refinement programs. For *REFMAC* a modified `ener_lib.dic` special class library with PURY added classes and corresponding covalent and van der Waals radius has to be used. (*SHELX* `ins` files read-in and read-out is on the way.)

The primary purpose of the server is to provide geometric parameters for ligands in macromolecular crystal structure refinement; however, the server can also be used for the validation of hetero compounds. In small-molecule structure refinement, the server can be used either as a validation tool or as an aid in assigning initial geometric target values for initial positional refinement.

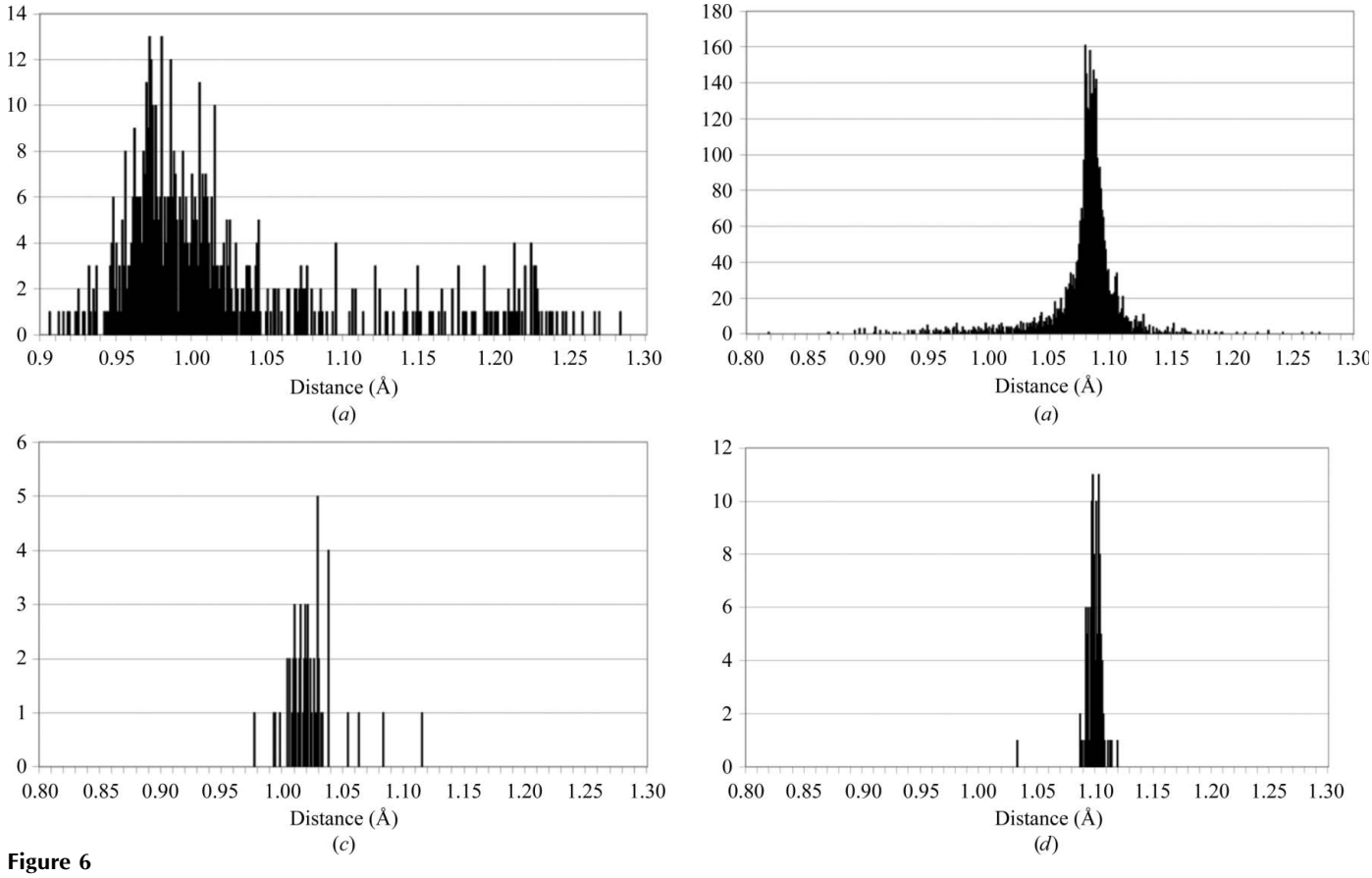

Figure 6
Histograms of bond distances between H atoms and selected atoms. The data from an analysis performed on structures determined using a neutron radiation source is presented. (*a*) The bond between an sp^3-hybridized O atom with one bonded H atom 'OH1<' and an H atom 'HP__'. (*b*) The bond between an sp^2-hybridized C atom in a six-membered aromatic ring with one bonded H atom 'CH_6' and an H atom 'HC__'. (*c*) The bond between an sp^2-hybridized planar N atom with one bonded H atom 'NH1Y' and an H atom 'HP__'. (*d*) The bond between an sp^3-hybridized C atom in six membered non-aromatic ring with one bonded H atom 'C61X' and an H atom 'HC__'.

Since the use of parameters derived from the CSD is bound to the CCDC license, internet access is restricted to CSD licensees. The current server (http://pury.ijs.si) will therefore shortly move to http://pury.ccdc.cam.ac.uk/.

4. Discussion

Validation of the PURY approach for the generation of geometric restraints for refinement has been performed from various aspects.

(i) By comparing a few experimental structures, we have checked the consistency and stability of the derived terms.

(ii) By comparing the variability of the bonding terms in different parts of chemical space, we have tried to assess the level of accuracy provided by the PURY parameters.

(iii) By cross-validation of the macromolecular crystal structures refined against PURY and Engh and Huber (EH) parameter sets, we have tried to assess the suitability of the PURY parameter set for refinement.

(iv) By comparing PURY parameterization with an expert-derived parameter set on a clearly defined subset of chemical space (EH parameters for amino-acid residues), we have tried to assess the limitations of the PURY approach and make some suggestions for EH set improvement.

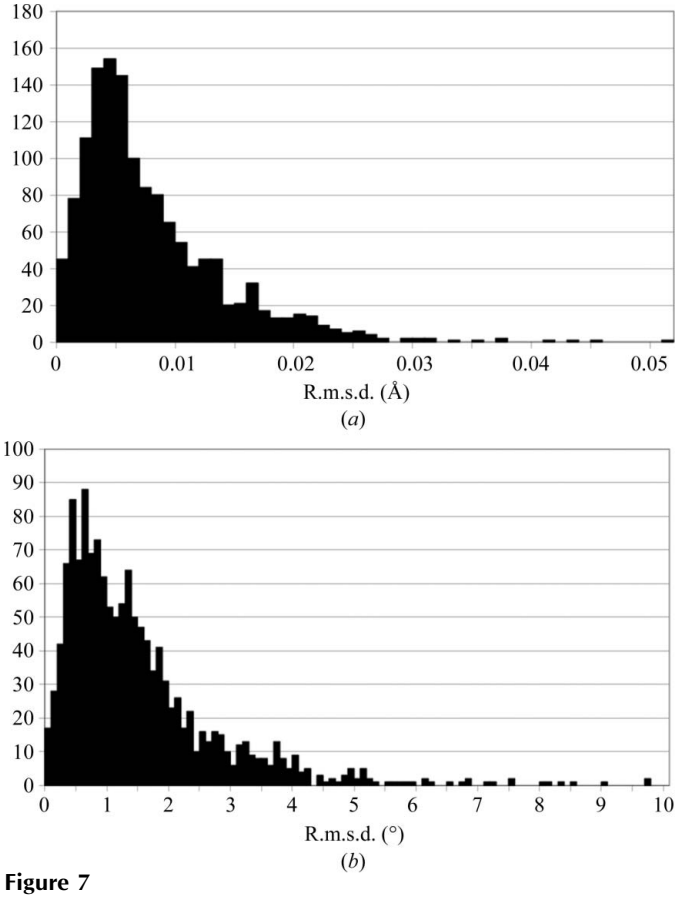

Figure 7
Histograms of bond and angle r.m.s.d. distribution for 1388 CSD structures validated with PURY parameters. (*a*) Bond r.m.s.d. distribution. (*b*) Angle r.m.s.d. distribution. R.m.s.d.s were calculated for each structure separately with *MAIN*. The bin thicknesses are 0.01 Å and 0.1° for bond and angle r.m.s.d.s, respectively.

Table 4
Validation of structures ABIYUF, CETPIA and ENAMEL.

The first two rows show the *R* factor and the temperature of experiment. The second group of rows show the bond r.m.s. values of experimental, *GAMESS*, *eLBOW* (*PHENIX*) and PURY models as validated using PURY parameters. The third group of lines shows the angle r.m.s. values of experimental, *GAMESS*, *eLBOW* and PURY models as validated with PURY parameters. The last group of rows show the coordinates r.m.s. difference between models. All PURY models were energy-minimized for 1000 steps using *MAIN*. The optimum geometric search for *GAMESS* models was performed with *ab initio* calculations at the HF/6-31 level until the density change between two consecutive runs was less than 1.0×10^{-5} using the *GAMESS* (US) package (Schmidt *et al.*, 1993; Gordon & Schmidt, 2005) from 22 November 2004 on G5, a dual 2.0 GHz with 1 GB RAM running OSX. The optimum geometric search for *eLBOW* models was performed using *eLBOW* from *PHENIX* v.1.3 RC2 using the −opt switch.

	ABIYUF	CETPIA	ENAMEL
Experimental *R* factor	4.84	6.8	4.18
Temperature (K)	283–303	283–303	105
Bond r.m.s. (Å)			
Experimental model	0.02	0.03	0.03
GAMESS model	0.02	0.002	0.02
eLBOW model		0.08	0.12
PURY model	0.003	0.002	0.002
Angle r.m.s. (°)			
Experimental model	1.38	1.65	2.22
GAMESS model	1.84	1.65	2.54
eLBOW model		3.462	8.284
PURY model	0.707	0.429	0.324
Coordinates r.m.s. (Å)			
Exp./PURY (max. value)	0.040 (0.081)	0.142 (0.361)	0.083 (0.151)
Exp./PURY (max. value)	0.040 (0.081)	0.142 (0.361)	0.083 (0.151)
GAMESS/PURY (max. value)	1.175 (2.323)	0.142 (0.361)	0.168 (0.470)
eLBOW/PURY (max. value)		2.444 (5.770)	1.729 (4.300)
eLBOW/*GAMESS* (max. value)		2.217 (5.430)	1.754 (4.383)
eLBOW/Exp. (max. value)		2.453 (5.716)	1.731 (4.362)
Exp./*GAMESS* (max. value)	1.175 (2.305)	0.142 (0.361)	0.143 (0.361)

4.1. Comparison with CSD experimental data

The minimal criterion for applicability of the generated geometric restraints is their consistency with the experimental structures from which they were derived. Deviation of bonding and angle terms of experimentally determined structures from PURY targets should lie within the limits of deviation of the database. To do this, we performed two validation tests. In the first we selected approximately 1300 deposited structures and validated them against the PURY data set, whereas in the second we selected three crystal structures for more elaborate comparisons.

The subset of 1388 CSD structures was selected using two criteria: they had to contain only C, H, N and O atoms and have B or E character on the third position in their refcode. Their bonds and angles were validated against PURY bond and angle parameters (Figs. 7*a* and 7*b*). The histograms show that the majority of bond deviations fall within 0.01 Å, with an average bond r.m.s.d. of 0.008 Å. Most of the angle deviations fall within 2°, with an average r.m.s.d. of 1.55°.

The three additional experimental structures were first validated using PURY geometry parameters. Comparison revealed that an r.m.s.d. for a bond of the ABIYUF structures of 0.02 Å and an r.m.s.d. for the angles of ABIYUF and CETPIA of 1.38° and 1.65° correspond to tight acceptance

criteria, whereas bond deviations of 0.03 Å for CETPIA and ENAMEL and an angle deviation of 2.22° for ENAMEL lie within the broadly acceptable boundaries (Table 4, Fig. 8).

When all three structures were energy-minimized in *MAIN* until convergence was reached (gradient < 4.2 kJ mol^{-1}) using PURY geometric restraints, the r.m.s. deviations dropped drastically, indicating that the PURY parameters are self-consistent. Also, the r.m.s.d.s of the energy-minimized models superimposed on the experimental structures (0.04, 0.14, 0.08 Å) revealed that the conformations of the PURY-minimized models remained essentially unchanged.

In order to assess the consistency of parameters with theoretical predictions, we have used the optimum geometric search with *ab initio* calculations using the *GAMESS* (US) package (Gordon & Schmidt, 2005). The initial models for the *ab initio* calculations for ABIYUF and ENAMEL were CSD structures, while for CETPIA the PURY-minimized model was used because the optimization using the experimental model did not converge. These structures delivered deviations of bond and angle terms within the range of PURY deviations when compared with the experimental model. The large difference in the r.m.s. of the coordinates, however, is the result of conformational differences.

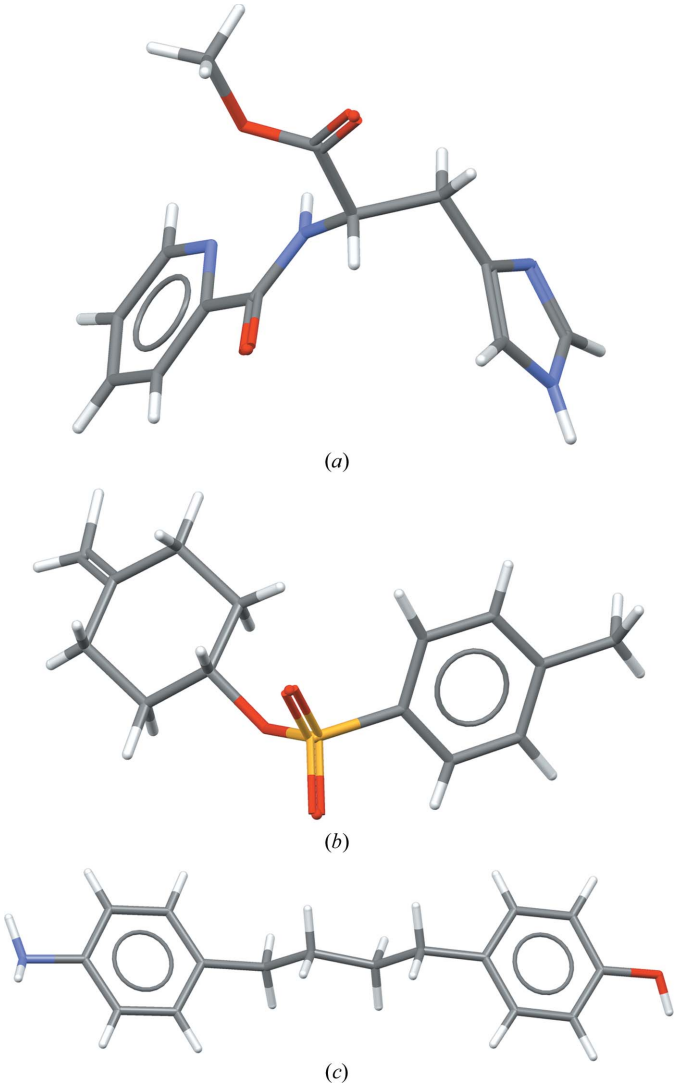

Figure 8
CSD v.5.28 selected entries ABIYUF (*a*), CETPIA (*b*) and ENAMEL (*c*) used for geometric comparison with PURY parameters. Figures were produced with *Mercury* (Macrae *et al.*, 2006).

Table 5
Average σ values for bonds and angles in PURY chemical subsets, the EH set and PDB hetero molecule (HET) derived sets.

	PURY, amino acids	PURY, nonmetal	PURY, all	EH	HET, nonmetal	HET, all
Bonds (Å)	0.027	0.043	0.064	0.022	0.073	0.091
Angles (°)	2.52	2.59	5.66	1.84	5.96	6.36

In addition, we optimized the selected structures with the program *eLBOW* from the *PHENIX* program suite (Adams *et al.*, 2002). The ENAMEL and CETPIA structures delivered deviations of bond and angle terms that were several times larger than others when compared with PURY and experimental structures. The ABIYUF minimization, however, failed to run. The large differences in the r.m.s. of the coordinates are again the result of conformational differences.

Hence, the PURY parameter set is consistent within itself, with the CSD structures as well as with *ab initio* calculations and is thus suitable for use in structure refinement.

4.2. Reliability parameters for different parts of chemical space

To demonstrate that parameters for different compounds may exhibit differing reliability, we divided the chemical space into three groups: amino acids, compounds containing metals and compounds not containing metals. We then compared their variability with those of the EH parameter set and with the hetero compounds deposited in the PDB divided into two groups: all structures containing nonmetals and those not containing nonmetals. The average σ values for bond and angle terms show that on broadening the pool of analyzed data the σ values of bonds and angles also broaden (Table 5). This is true for the three PURY portions, as well as when comparing the variability of EH parameters with those of the hetero compounds with and without metals. From Table 5 it is also evident that the PURY deviations for amino acids are higher than those of EH; however, they are still within acceptable limits, as suggested by Jaskolski *et al.* (2007), who recently analyzed PDB and CSD data for the use of geometric restraints in refinement of macromolecules. However, the deviations of PURY parameters are lower than those obtained by analyzing hetero compounds in the PDB: 0.043 Å and 2.59° for PURY *versus* 0.073 Å and 5.96° from the PDB for nonmetal hetero compounds, and 0.064 Å and 5.66° *versus* 0.091 Å and 6.36° for all hetero compounds.

This comparison suggests that by using the PURY parameter set the deviations from ideal values of hetero compounds deposited in the PDB could be significantly decreased and thereby made more accurate.

Table 6
Refinement statistics and cross-validation of the crystal structure of cathepsin B (PDB code 1sp4; Štern *et al.*, 2004), two β-lactamase crystal structures (PDB codes 2q9m and 2q9n; Plantan, 2007) and the crystal structure of the SARS coronavirus ORF7A accessory protein (PDB code 1xak; Nelson *et al.*, 2005).

All four structures were refined with the program *MAIN* using all structure factors in the available resolution range. The crystallographic refinement target was set to 0.01 Å for the r.m.s.d. bond deviations. The structures were first distorted with a 0.3 Å kick and then refined against PURY and EH target values until the gradient reached the value of 5 energy units.

PDB code	1sp4	2q9m	2q9n	1xak
Resolution (Å)	2.20	2.05	2.20	1.80
PURY *R* factor ($R_{\rm free}$)	19.5	21.3 (24.9)	25.4 (29.8)	23.9 (27.5)
EH *R* factor ($R_{\rm free}$)	19.3	21.2 (24.8)	25.1 (29.7)	23.1 (28.4)
Bond r.m.s. (Å)				
PURY/PURY	0.0112	0.0108	0.0109	0.0093
EH/EH	0.0111	0.0113	0.0110	0.0129
PURY/EH	0.0141	0.0144	0.0133	0.0125
EH/PURY	0.0149	0.0158	0.0114	0.0167
Angle r.m.s. (°)				
PURY/PURY	1.649	1.744	1.786	1.559
EH/EH	1.578	1.701	1.807	1.744
PURY/EH	1.656	1.798	1.796	1.676
EH/PURY	1.858	1.960	1.846	1.904
Coordinate r.m.s. (PURY/EH) (Å)	0.026	0.020	0.005	0.082
Max. coordinate shift (PURY/EH) (Å)	0.244	0.196	0.030	0.312

4.3. Validation of PURY restraints in refining macromolecular structures

Comparison of the parameter set generated by an algorithm, which is supposed to cover the complete space of chemical compounds, with a widely validated parameter set constructed by an expert from a selected set of chemical compounds provides yet another estimate of reliability of the derived parameters. We have therefore compared the consistency of the PURY parameters with the EH parameters. We have chosen four macromolecular crystal structures, three from the laboratory and one external structure, with different resolution ranges and refined them against EH and PURY targets. Each structure was refined using the same starting coordinates, the same target (bond r.m.s.d. = 0.01 Å to enlarge the impact of the tight geometric restraints) and the same computational tools and protocols. As seen in Table 6, the deviations from the bond target values are almost identical in all four cases. They differ by −0.0001, 0.0005, 0.0001 and 0.0036 Å, indicating that the structures have indeed been refined equivalently. Also, the r.m.s.d.s of angle deviations against the targets used in refinement are very similar, 0.077°, −0.044° and 0.021°, although in the fourth case the difference is slightly higher at 0.185°. The cross-validation, in which structures refined against PURY targets were validated against the EH parameter set, showed a slight increase of bond r.m.s.d.s (0.003, 0.003, 0.002, 0.003 Å) as well as angle r.m.s.d.s (0.007°, 0.054°, 0.01°; in the fourth case it fell) and the reverse cross-validation is in general slightly higher. Comparison of the crystallographic *R* values revealed small differences (0.1, 0.03, 0.2 and 0.8%) in favour of the EH parameter set. The conclusion is nevertheless clear; the PURY parameter set performs essentially equivalently to the EH parameter set, suggesting that the use of PURY parameters in the refinement of hetero compounds will behave equivalently to the expert-derived data set(s).

Table 7
Matches between EH and PURY atom classes and *vice versa*.

EH class	PURY class	PURY class	EH class
C	COO_, C__G, C__Y	CH3X	CH3E
CN	COO_	CH2X	CH2G, CH2E
C5	CC_5	CH1X	CH1E
C5W	CC_5	C51X	CH1E
CF	CC_6	C52X	CH2P, CH2E
CY	CC_6	CH_5	CR1H, CRH, CRHH
CY2	CO_6	CH_6	CR1W
CW	C__*	CC_5	C5, C5W
CR1E	CH_5, CH_6	CC_6	CY, CF
CR1H	CH_5	CO_6	CY2
CR1W	CH_6	C__*	CW
CRH	CH_5	COO_	C, CN
CRHH	CH_5	C__G	C
CH1E	CH1X, C51X	C__Y	C
CH2E	CH2X, C52X	NH3X	NH3
CH2G	CH2X	NH2Y	NH2
CH2P	C52X	NH1Y	NH1
CH3E	CH3X	NH1G	NH1
NH1	NH1Y, NH1G, NH_5	NH2G	NC2
N	N5_Y	N5_Y	N
NC2	NH2G	NH_5	NH1
NH2	NH2Y	N__5	NR
NH3	NH3X	OH2<	OT
NR	N__5	OH1<	OH1
O	O2C_	O2C_	O
OC	O-1_	O-1_	OC
OH1	OH1<	SH1<	SH1E
OT	OH2<	S__<	S, SM
S	S__<	HP__	H, HN4T, HT
SH1E	SH1<	HC__	HC
SM	S__<		
H	HP__		
HC	HC__		
HN4T	HP__		
HT	HP__		

4.4. Detailed comparison with an expert-derived parameter set

We have shown that the PURY parameter set performs essentially equivalently to the EH set; however, a direct comparison of individual terms would be informative as it exposes a few limitations and provides hints for future development. Only a selected set of comparisons between the two parameter sets are presented here.

The number of PURY classes (30) covering the 20 amino-acid residues differs from the number of EH classes (35). Translation from one set to the other is not 'bidirectional' since quite often a single EH class is described by several PURY classes and *vice versa* (Table 7). Although the comparison of classes is indicative, the true value of such a comparison can only be revealed by comparison of the geometrical parameters they define.

4.4.1. Tyrosine/phenylalanine CG atoms. PURY uses the same atom class 'CC_6' to describe the CG atom of phenylalanine and tyrosine residues, whereas EH uses two classes, 'CF' and 'CY'. A consequence of this is that in the EH set

there are two different parameters for the CB—CG bond which describe the single bond by which the phenyl ring is attached to the alanine base and two for the CG—CD bonds which describe the bond in the aromatic ring and the corresponding angles.

The PURY mean bond value for CB—CG is 1.508 Å (0.019 Å), which is between the EH target values of 1.502 Å (0.023 Å) and 1.512 Å (0.022 Å) for Phe and Tyr residues, respectively. Both EH values lie within the 1σ range of the PURY target: the differences are −0.006 Å and +0.004 Å. Also, the 0.01 Å difference between the two EH values lies well within 1σ. The histogram of bond lengths shown in Fig. 9(*a*) indicates the presence of a double peak in the 'CC_6–CH2X' bonds corresponding to the EH target values. The histogram of distances of CG—CD bonds shown in Fig. 9(*b*) also shows a double peak. Interestingly, although only one peak matches the EH value (1.389 Å) corresponding to the Tyr bond, the other peak appears higher (1.394 Å) rather than lower as for the EH value for the phenylalanine bond (1.384 Å), suggesting that the target for this bond-length restraint should be re-evaluated.

In contrast, the histograms for angles shown in Figs. 9(*c*) and 9(*d*), corresponding to the phenylalanine and tyrosine CB—CG—CD(1,2), CD1—CG—CD2 angle-value distributions, exhibit no double peaks. The EH parameters for the CB—CG—CD(1,2) angle differ by 0.5° and the PURY angle is in the middle of the two. Interestingly, the histogram for 'CH_6–CC_6–CH_6' (Fig. 9*d*) reveals the presence of a peak at exactly 120°, which is probably an indication of the constrained angle value used during refinement. The absence of this peak from the CB—CG—CD(1,2) histogram (Fig. 9*c*) is probably not apparent since the restraint and the mean are equal.

This comparison thus questions the need for the two different classes for the CG atom of Phe and Tyr residues in the refinement of macromolecular structures.

4.4.2. Carboxylic group. The EH parameter set (Engh & Huber, 1991) uses two different atom classes for designation of the carboxylic acid group C atom: 'C' and 'CN' are used for charged and neutral carboxylic groups, respectively. The corresponding targets for the C—OC and CN—O bond lengths are 1.249 and 1.208 Å, respectively, where the C—OC bond describes groups with a delocalized double bond in the charged group and CN—O describes the double bond of the neutral group. The CN—OH1 bonding parameter is lacking; however, an equivalent parameter for the carbon–oxygen single bond C—OH1 can be found with a target bond length of 1.304 Å. [In the updated EH parameter set published in 2001

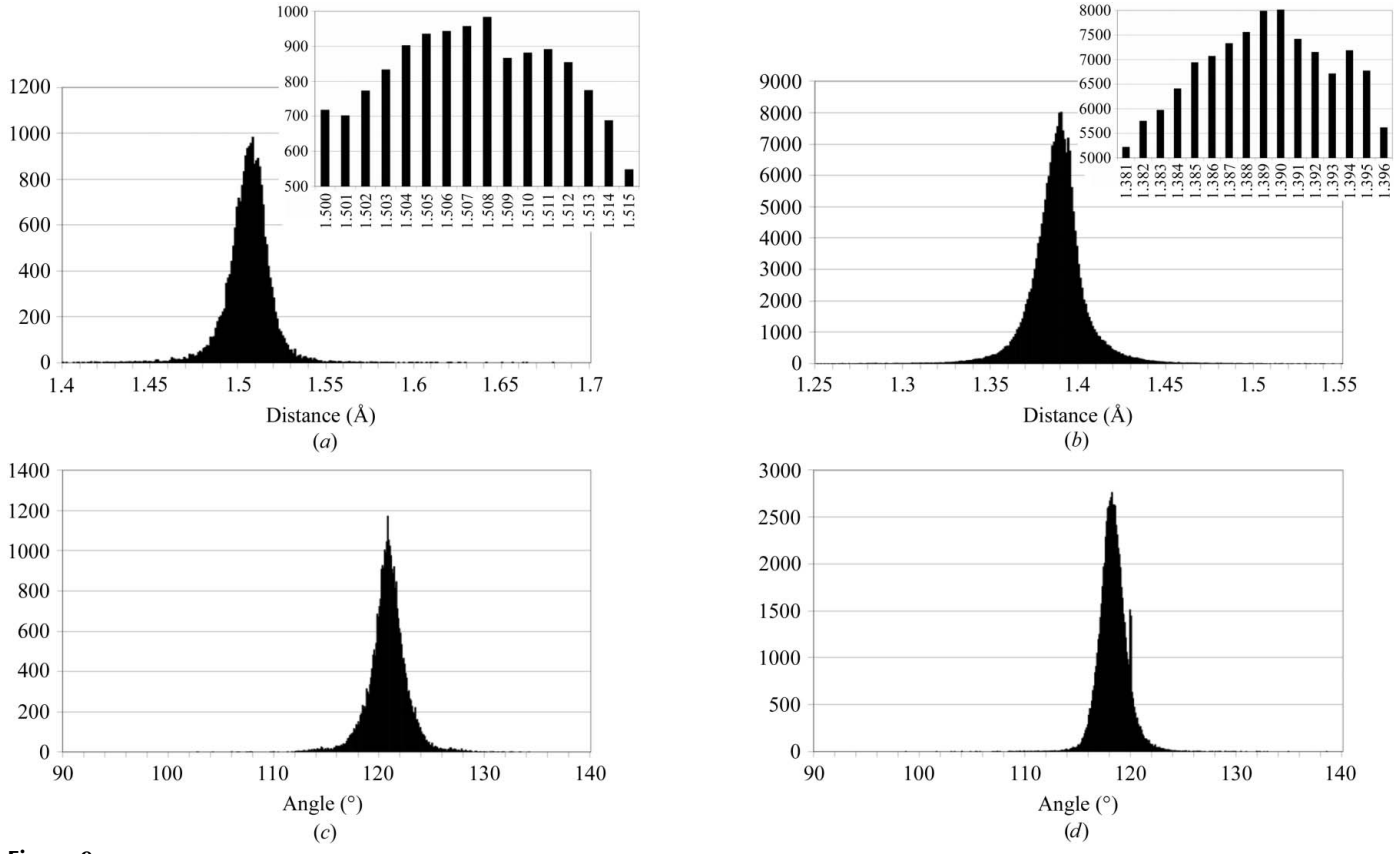

Figure 9
Histograms of bond and bond-angle parameters involving the CG atoms of phenylalanine and tyrosine residues. (*a*) A CB—CG bond, which describes the single bond by which the phenyl ring is attached to the alanine base 'CH2X–CC_6'. The insert shows an enlargement of the peak. (*b*) A CG—CD bond, which describes the bond in the aromatic ring 'CC_6—CH_6'. The insert shows an enlargement of the peak. (*c*) A CB—CG—CD angle, which describes the angle by which the phenyl ring is attached to the alanine base 'CH2X–CC_6–CH_6'. (*d*) A CD1—CG—CD2 bond, which describes the angle in the aromatic ring 'CH_6–CC_6–CH_6'.

(Engh & Huber, 2001), the fragments do not contain bond lengths for the neutral carboxylic group.] In the PURY parameter set, three bonding parameters are used to describe the charged and neutral carboxylic groups of Glu and Asp residues: 'COO_–O-1_' (1.255 Å) for the delocalized double bond of the charged groups and 'COO_–OH1<' (1.306 Å) and 'COO_–O2C_' (1.215 Å) for the double and single bonds of the neutral group. The histogram of the bond distances 'COO_–O-1_' is symmetrical (Fig. 10*a*), suggesting that two equivalent O atoms are bonded to the C atom. Its average value of 1.255 Å comes right in the middle of the averages for the double (1.215 Å) and single (1.306 Å) bond distances of the neutral carboxylic group (Figs. 10*b* and 10*c*). The corresponding EH and PURY angle terms are equivalent; however, the EH parameter set provides no angle terms for the 'CN' atom class. The angle terms shown in Figs. 10(*d*) and 10(*e*) show single peaks with equivalent target values, suggesting that for the angle parameters there is no noticeable difference between the neutral and charged carboxylic group; thus, the EH set angle values apply equally well to both charged and neutral carboxylic groups.

The analysis of the CSD is based on the assumption that the structural data are correct and that H atoms are present when the carboxylic group is not charged. However, the broad bottom of the charged-group bonding distance (Fig. 10*a*) and the nonsymmetrical distributions of the bond distances of the neutral groups (Figs. 10*b* and 10*c*) leave the impression that not all carboxylic groups are necessarily correctly assigned. This is also true for the protein structures. Our analysis suggests that the lack of consistent parameters for neutral

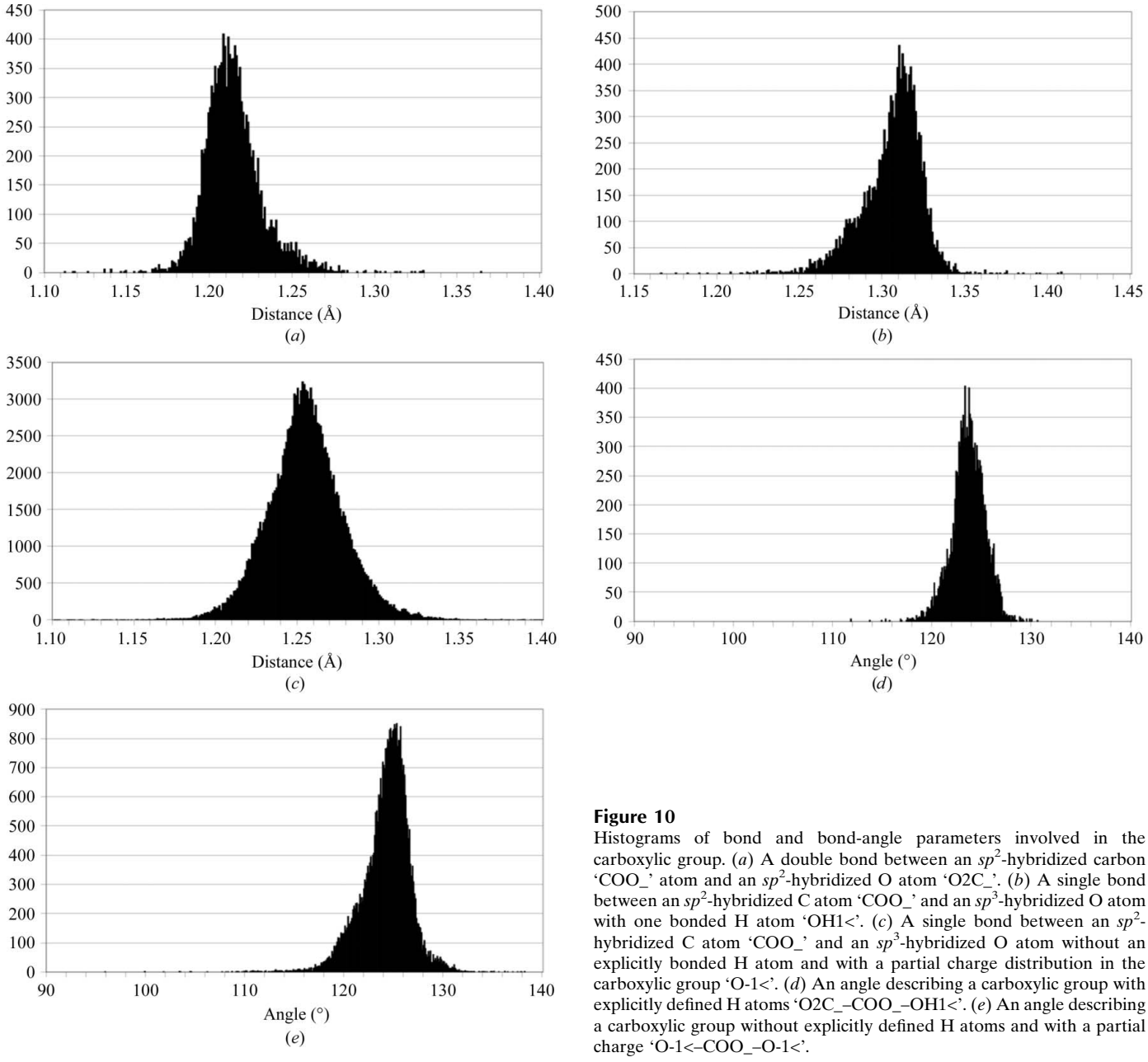

Figure 10
Histograms of bond and bond-angle parameters involved in the carboxylic group. (*a*) A double bond between an sp^2-hybridized carbon 'COO_' atom and an sp^2-hybridized O atom 'O2C_'. (*b*) A single bond between an sp^2-hybridized C atom 'COO_' and an sp^3-hybridized O atom with one bonded H atom 'OH1<'. (*c*) A single bond between an sp^2-hybridized C atom 'COO_' and an sp^3-hybridized O atom without an explicitly bonded H atom and with a partial charge distribution in the carboxylic group 'O-1<'. (*d*) An angle describing a carboxylic group with explicitly defined H atoms 'O2C_–COO_–OH1<'. (*e*) An angle describing a carboxylic group without explicitly defined H atoms and with a partial charge 'O-1<–COO_–O-1<'.

charged carboxylic groups should be corrected and the standard EH parameter set should be extended for use in the refinement of protein structures.

4.4.3. Proline. The PURY bond length for the CA—CB bond in the proline residue ('C51X–C52X', average = 1.529 Å, σ = 0.021 Å) is broader than that used in the EH set, where it is kept constant for all amino acids (average = 1.530 Å, σ = 0.020 Å); however, PURY differentiates between the proline ring CA–CB atoms ('C51X–C52X') and the non-ring CA–CB atoms ('CH1X–CH2X'). As previously noted (Engh & Huber, 2001), the proline CB—CG bond with an average length of 1.492 Å has a large σ value (σ = 0.05 Å). PURY analysis also delivered similar values (average = 1.491 Å, σ = 0.051 Å). However, the histogram of 'C52X–C52X' bonds between sp^3 atoms in a five-membered ring is skewed, with the peak at 1.518 Å, which lies away from the average value. The broadening and the skewed shape indicate that the variability is a result of ring puckering, during which the two atoms approach each other (Figs. 11*a* and 11*b*).

The double peak of the bonding angle is reminiscent of the proline-residue analysis, in which coupling was observed between the bonding and dihedral angles. The current PURY approach cannot differentiate between *cis* and *trans* prolines, as noted by Engh & Huber (2001), suggesting that an expert parameter set performs better. The bond-length assignment is actually a problem of the concept of atom-class assignment

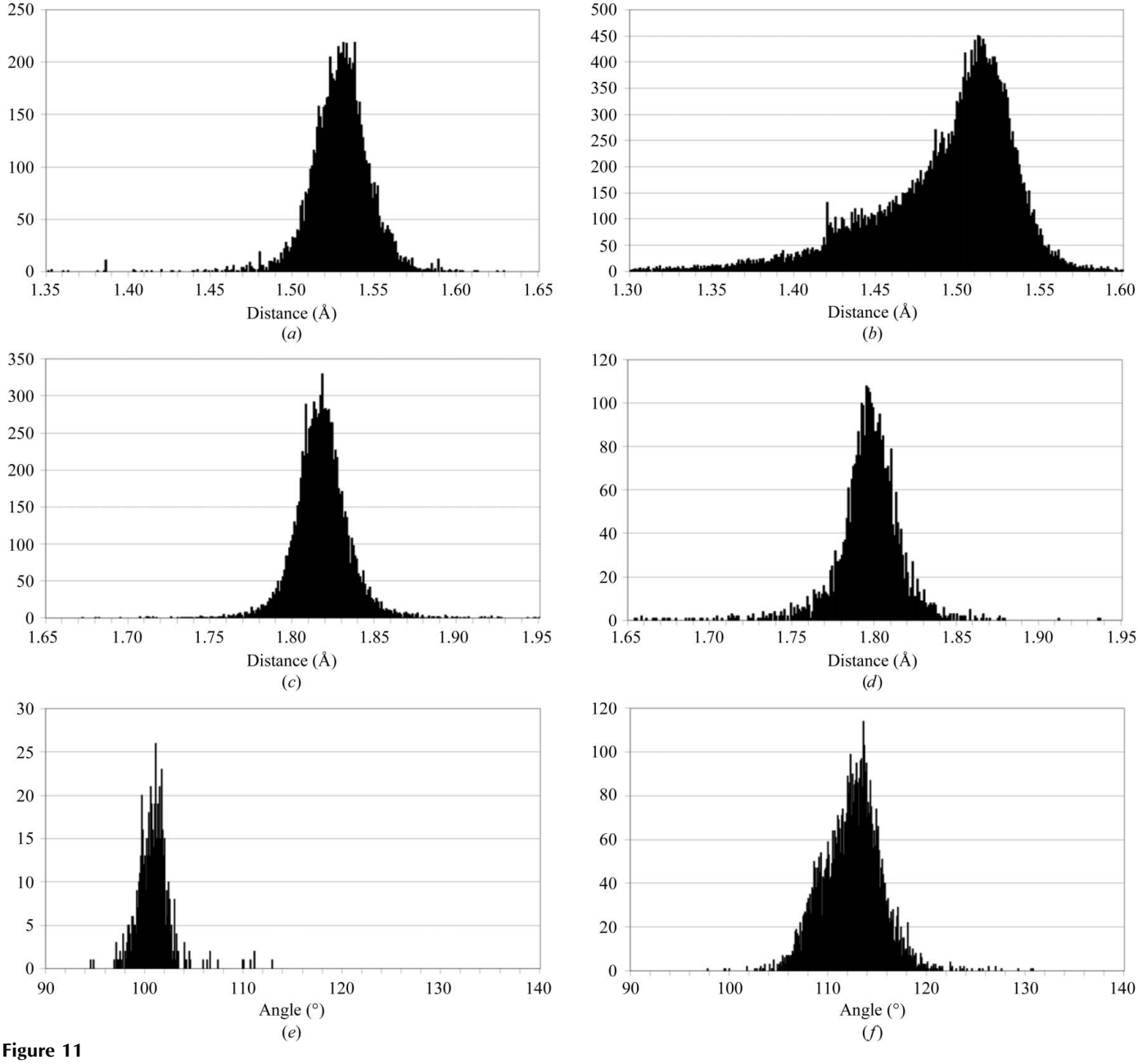

Figure 11
Histogram distributions of bond and bond-angle parameters involved in various amino-acid residues: (*a*) proline CA—CB 'C51X–C52X', (*b*) proline CB—CG 'C52X–C52X', (*c*) methionine CG—SD 'CH2X–S__<', (*d*) methionine SD—CE 'S__<–CH3X', (*e*) methionine CG—SD—CE 'CH2X–S__<–CH3X', (*f*) methionine CB—CG—SD 'CH2X–CH2X–S__<'.

Table 8
Comparison of methionine geometric parameters from PURY and EH sets.

Atom classes involved in bonds, angles and dihedral angles describing the side chain of methionine and corresponding average values and force constants are shown. Specific parameters for the selenomethionine residue are shown below.

Entry	Class 1	Class 2	Class 3	Class 4	Force constant	Multiplicity	Average value
EH methionine-specific parameters							
Bond	CH2E	SM			512.111		1.803 Å
Bond	CH3E	SM			170.066		1.791 Å
Angle	CH2E	SM	CH3E		401.534		100.900°
Angle	CH2E	CH2E	SM		215.936		112.700°
Dihedral	X	CH2E	SM	X	3.60	3	0.000°
Dihedral	X	CH2E	CH2E	X	4.80	3	180.000°
PURY methionine-specific parameters							
Bond	CH2X	S__<			1643.603		1.818 Å
Bond	CH3X	S__<			1324.387		1.796 Å
Angle	CH2X	S__<	CH3X		733.833		100.904°
Angle	CH2X	CH2X	S__<		217.735		112.487°
Dihedral	CH2X	CH2X	S___<	CH3X	7.41	3	60.000°
Dihedral	CH2X	CH2X	CH2X	S__<	4.71	3	60.000°
PURY selenomethionine-specific parameters							
Bond	CH2X	Se_<			2492.665		1.960 Å
Bond	CH3X	Se_<			502.996		1.943 Å
Angle	CH2X	Se_<	CH3X		1070.409		97.976°
Angle	CH2X	CH2X	Se_<		342.319		113.030°
Dihedral	CH2X	CH2X	Se__<	CH3X	7.27	3	60.000°
Dihedral	CH2X	CH2X	CH2X	Se_<	4.72	3	60.000°

based on chemical environment, which cannot differentiate between geometric arrangements such as *cis* and *trans* peptide bonds. A simple solution to this problem is to use different topology library entries in combination with different atom classes for describing prolines in *cis* and *trans* conformations.

4.4.4. Methionine. Parameters involving the S atom in methionine EH and PURY parameters differ. The EH CG—SD bond target of 1.803 Å is 0.015 Å shorter (almost 1σ, 0.019 Å) than the PURY value (1.819 Å), whereas the SD—CE targets are much more alike: 1.791 and 1.796 Å for EH and PURY, respectively. The parameters for the angles CB—CG—SD and CG—SD—CE are also similar. Interestingly, however, the PURY force constants are evidently higher than those for EH for all terms except the CG—SD—CE angle, where they are approximately equal. The bond and angle histograms involving methionine parameters are shown in Figs. 11(*c*), 11(*d*), 11(*e*) and 11(*f*). Owing to the larger number of structures used in the PURY analysis (when compared with the presumably smaller number used to derive EH targets), we suggest modifying the geometric

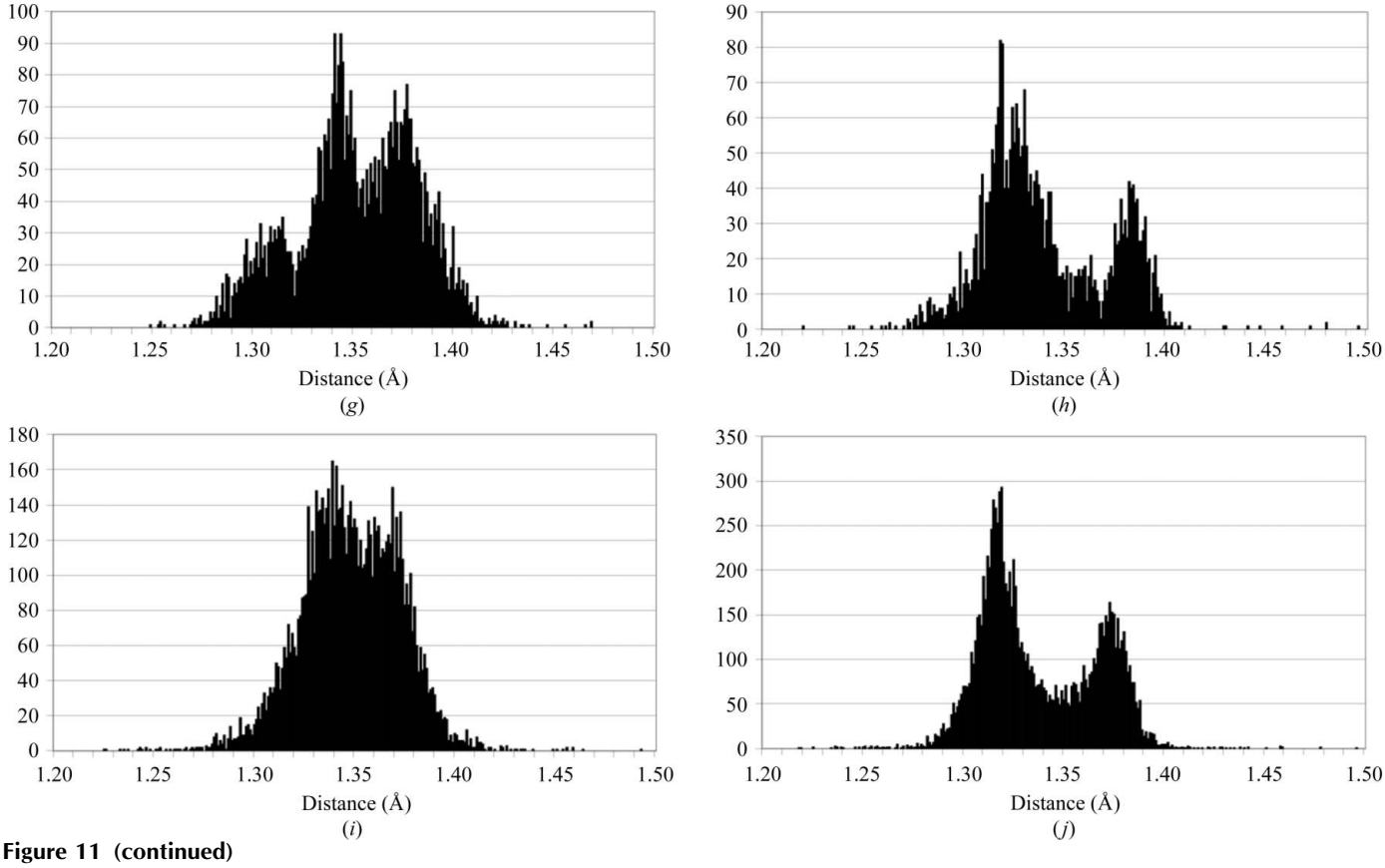

Figure 11 (continued)
(*g*) histidine CG—ND1 'CC_5–NH_5', (*h*) histidine NE2—CG 'CC_5–N__5', (*i*) histidine CE1—NE2 'CH_5–NH_5', (*j*) histidine CE1—ND1 'CH_5–N__5'.

restraints (in particular forces) for the methionine residue. In addition, PURY parameters for the restraint of selenomethionine residues are provided in Table 8.

4.4.5. Histidine. The PURY and EH parameter sets differ most in the histidine-residue terms. Both atom-class assignments differentiate between protonated and nonprotonated ND1 and NE2 atoms; however, PURY atom-class assignment does not differentiate between CD2 and CE1 atoms, which are both recognized as 'CH_5' (C atom within a five-membered planar ring with a bonded H atom). As a consequence, four different bonding targets of the EH set merge into one PURY target which lies somewhere in the middle. The situation is similar with the bond length of nonprotonated histidines, which merge into a single PURY target value. The longer and shorter bonding distances are an indication of single-bond and double-bond character of the ring bonds and cannot be appropriately elaborated with the current PURY concept (Figs. 11*g*, 11*h*, 11*i* and 11*j*). Namely, only a single target value can be derived for bonds between two atom classes, yet the bonds between two atom classes can be single or double. For example, in a putative case in which sp^2 hybridized atoms are bonded to each other in a chain, single and double bonds alternate, yet they are both bonds between the same atom classes. Clearly, the chance of such a case occurring within a small set is rather small; however, when a large pool of data such as the CSD has been analyzed, the occurrence of such cases is not that uncommon.

This suggests including the bond type in the creation of the bonding-parameter database is a task for future development. This may not directly affect the current concept of one pair of atom classes, one target value. Within a single residue this concept could still be valid since in the case of ambiguity new 'artificial' atom classes could be introduced, whereas for the future the concept of a single parameter descriptor with several target values will be introduced (as previously noted in the case of *cis* and *trans* proline residues).

5. Conclusions and future plans

The creation of the PURY database from structures deposited in the CSD and its analysis have shown that the derived geometric restraints are of sufficient accuracy for use in refining the crystal structures of macromolecules at non-atomic resolution. The use of the PURY database would probably increase the accuracy of the geometries of hetero-compound structures deposited in the PDB. Comparison with the Engh–Huber parameter set has shown that an expert-derived data set derived from a preselected set of structures has advantages over the general approach applied in PURY. The comparison also revealed that the EH parameter set can be expanded with a few PURY terms. Such a modified EH parameter set with the data presented here is made available as part of the *MAIN* distribution (http://www-bmb.ijs.si).

The analysis, in particular the multiple maxima and nonsymmetrical histograms of geometrical terms, has exposed two essential questions. Firstly, is the current atom-class assignment scheme indeed recognizing all appropriate atom classes (which important issues have we missed)? Secondly, how reliable are the data presented in the CSD? We hope that the use of PURY and the imminent validation of all structures deposited in the CSD against the PURY database will expose further potential misassignments, while the use of PURY for the validation of small-molecule structures may draw more attention to details of the structures such as charged states and protonation and thereby increase the reliability of the data being deposited. Unfortunately, the lack of structure factors prevents the remediation of structures already in the CSD. Perhaps it is time to change the policy to require deposition of these data for all published structures in all journals and make them publicly available too.

The future plan is to update and evolve the *PURY* algorithm, which will be expanded in such a way as to deal successfully with the above-mentioned problems and gain functionality as a simple web-driven tool for validation of small-molecule geometries.

R. Taylor is acknowledged for finding hetero compounds in the PDB which have a match in the CSD. Z. Štefanić and G. Gunčar are gratefully acknowledged for discussions and R. Pain is gratefully acknowledged for critical reading of the manuscript. The Slovenian Research Agency is acknowledged for funding.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 382–392. Dordrecht: Kluwer Academic Press.

Gordon, M. S. & Schmidt, M. W. (2005). *Theory and Applications of Computational Chemistry: The First Forty Years*, edited by C. E. Dykstra, G. Frenking, K. S. Kim & G. E. Scuseria, pp. 1167–1189. Amsterdam: Elsevier.

Greaves, R. B., Vagin, A. A. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 1335–1339.

Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst.* D**63**, 611–620.

Kleywegt, G. J., Henrick, K., Dodson, E. J. & van Aalten, D. M. F. (2003). *Structure*, **11**, 1051–1059.

Kleywegt, G. J. & Jones, T. A. (1998). *Acta Cryst.* D**54**, 1119–1131.
Lamzin, V. S., Dauter, Z. & Wilson, K. S. (1995). *J. Appl. Cryst.* **28**, 338–340.
Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & van de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Nelson, C. A., Pekosz, A., Lee, C. A., Diamond, M. S. & Fremont, D. H. (2005). *Structure*, **13**, 75–85.
Nilsson, K., Lecerof, D., Sigfridsson, E. & Ryde, U. (2003). *Acta Cryst.* D**59**, 274–289.
Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *Acta Cryst.* D**52**, 57–64.
Plantan, I. *et al.* (2007). *J. Med. Chem.* **50**, 4113–4121.
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., Windus, T. L., Dupuis, M. & Montgomery, J. A. (1993). *J. Comput. Chem.* **14**, 1347–1363.
Schüttelkopf, A. W. & van Aalten, D. M. F. (2004). *Acta Cryst.* D**60**, 1355–1363.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Štern, I., Schaschke, N., Moroder, L. & Turk, D. (2004). *Biochem. J.* **381**, 511–517.
Turk, D. (1992). PhD Thesis. Technische Universität, München.
Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst.* D**60**, 2184–2195.
Wlodek, S., Skillman, A. G. & Nicholls, A. (2006). *Acta Cryst.* D**62**, 741–749.