Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

Introduction and validation of an invariom database for amino-acid, peptide and protein molecules

B. Dittrich, C. B. Hübschle, P. Luger and M. A. Spackman

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Introduction and validation of an invariom database for amino-acid, peptide and protein molecules

B. Dittrich,[a]* C. B. Hübschle,[b] P. Luger[b] and M. A. Spackman[a]

[a]Chemistry M313, School of Biomedical, Biomolecular and Chemical Sciences, University of Western Australia, Crawley, WA 6009, Australia, and [b]Institut für Chemie/Kristallographie, Freie Universität Berlin, 14195 Berlin, Germany

Correspondence e-mail: birger@cyllene.uwa.edu.au

Received 13 April 2006
Accepted 25 July 2006

A database of invarioms for structural refinement of amino-acid, oligopeptide and protein molecules is presented. The spherical scattering factors of the independent atom or promolecule model are replaced by 'individual' aspherical scattering factors that take into account the chemical environment of a bonded atom. All amino acids were analysed in terms of their invariom fragments. In order to generate 73 database entries that cover this class of compounds, 37 model compounds were geometry-optimized and theoretical structure factors were calculated. Multipole refinements were then performed on these theoretical structure factors to yield the invariom database. Validation of this database on an extensive number of experimental small-molecule crystal structures of varying quality and resolution shows that invariom modelling improves various figures of merit. Differences in figures of merit between invariom and promolecule models give insight into the importance of disorder for future protein-invariom refinements. The suitability of structural data for application of invarioms can be predicted by Cruickshank's diffraction-component precision index [Cruickshank (1999), *Acta Cryst.* D**55**, 583–601].

1. Introduction

Advances in computer performance and software development today allow a general application of aspherical scattering factors. An aspherical scattering-factor formalism for modelling the intensities in X-ray single-crystal diffraction experiments was introduced as early as 1969 (Stewart, 1969). Aspherical scattering factors in this work are based on invarioms (Dittrich *et al.*, 2004; from **invari**ant at**oms**), intermolecular transfer-invariant pseudoatoms within the Hansen and Coppens multipole formalism (Hansen & Coppens, 1978). The multipole model allows a pseudoatom representation of deformations of the electron density $\rho(\mathbf{r})$ arising from chemical bonding and is described in detail by Coppens (1997). 'Individual' invariom aspherical scattering factors take into account the chemical environment of a bonded atom and replace the usual spherical independent-atom model (IAM) scattering factors.

Recently, we have shown that invariom modelling improves the accuracy of molecular geometry from X-ray single-crystal diffraction. Modelling with aspherical scattering factors was shown to be more appropriate than application of the IAM, in the process improving the description of thermal movement by anisotropic displacement parameters as quantified with the Hirshfeld test (Dittrich *et al.*, 2005). An overview of previous and related work is given in that paper.

In the present work, we introduce a library of invarioms (containing invariom name, name of the model compound,

doi:10.1107/S090744490602899X

local atomic coordinate system/site symmetry and their multipole parameters) from which aspherical scattering factors can be calculated for structure refinement of amino-acid, oligopeptide and protein molecules from single-crystal X-ray diffraction data. We use a database approach similar to that based on multipole populations obtained from ultrahigh-resolution diffraction experiments (Brock *et al.*, 1991; Pichon-Pesme *et al.*, 1995) of small-molecule compounds. A major difference from earlier work is that the entries generated are based on quantum-mechanical calculations. Recent debate has centred around the advantages and disadvantages of how to obtain such databases (Pichon-Pesme *et al.*, 2004; Volkov, Koritsánzky *et al.*, 2004). In our opinion, the obvious advantages of a theoretical database are that no experimental error is involved, no varying influence of hydrogen bonding or crystal packing occurs and an unlimited number of chemical environments consisting of all possible elements can be simulated with ease by such virtual experiments. These arguments have also been emphasized by Volkov, Koritsánzky *et al.* (2004), who have also developed a theoretical pseudoatom data bank where parameters are averaged based on a selection of organic compounds (Volkov, Li *et al.*, 2004). Apart from details in the computational procedure, avoiding averaging parameters is the major difference in the invariom database, where entries are each derived from a unique model compound.

The invariom database reported here has been tested and verified on an extensive number of amino-acid and oligopeptide small-molecule compounds for which experimental diffraction intensities were available from IUCr journals. All amino acids, in protonated and unprotonated forms where available, and some of their derivatives are contained in this set of trial structures. These and the forms absent in the trial structures were covered by our library, which we intend to place in the public domain after having applied it to protein data.

Considerable effort has been invested in protein crystallography at subatomic resolution and recent work has been reviewed by Petrova & Podjarny (2004). Work on a database of generalized scattering factors especially for amino-acid, oligopeptide and protein molecules began in 1995. Based on several oligopeptide charge-density studies, Pichon-Pesme *et al.* (1995) stated that 'parameters of the same kind of atoms in the same chemical environment are statistically equal'. The authors further reported a considerable improvement of the figures of merit for a pseudotripeptide structure after experimentally derived database scattering factors were applied. The study relied on transferred and unrefined P_{lm} parameters, κ (κ') were set to unity and P_v were kept neutral[1]. However, in a later study of an octapeptide (Jelsch *et al.*, 1998) it was found 'that P_v does not seem to be transferable', limiting transferability to P_{lm}. Subsequently, the database approach of Lecomte and coworkers has evolved further and more recent publications and findings clearly contradict the early optimism of the 1995 study. For example, in the study of crambin (Jelsch *et al.*, 2000) it was reported that the 'database deformation density is clearly overestimated', while no figures of merit were given. The authors refined P_v and were able to obtain a crude electrostatic potential. Another study published in the same year on the scorpion toxin II from *Androctonous australis* (Housset *et al.*, 2000) concluded that the database values 'did not provide a significant overall improvement in the agreement between $F_{\rm obs}$ and $F_{\rm calc}$'. More recently, in the ongoing study of the protein aldose reductase (Muzet *et al.*, 2003) their database values were seen as 'starting values' for 'charge-density refinement of proteins' and the refinement is still in progress (Jelsch *et al.*, 2005).

While an experimental database is appropriate to provide starting values for charge-density refinement of proteins, it is not entirely satisfactory for a database approach that aims at providing a generalized scattering model. For the theoretically derived database entries reported here, electron-density transferability was the underlying principle. Our database parameters are precisely transferable and application involves minimal rescaling of P_v parameters. Transferability cannot be limited to a subset of the multipole parameters, as in the database approach of Lecomte and coworkers, and transferable multipole parameters are required to be treated as a complete and inseparable set. The present structures investigated with the invariom approach give valuable insight into possible causes for the contradictory results obtained earlier and the invariom approach should significantly improve future protein refinement.

2. Database and refinement

2.1. Database generation

The invariom database contains Hansen and Coppens multipole parameters (Hansen & Coppens, 1978) and information about the local atomic coordinate system. Multipole parameters are P_v (valence), P_{lm} (deformation), κ (valence contraction/expansion) and κ' (deformation contraction/expansion), defined by

$$\rho_{\rm atom}(\mathbf{r}) = \rho_{\rm core}(r) + P_v\kappa^3\rho_v(\kappa\mathbf{r}) + \sum_{l=0}^{l_{\max}}\kappa'^3R_l(\kappa'\mathbf{r})\sum_{m=0}^{l}P_{lm\pm}d_{lm\pm}(\theta,\varphi). \qquad (1)$$

The multipole parameters and their local atomic site symmetry (Kurki-Suonio, 1977) depend on a local coordinate system based on neighbouring atoms (in other words, a particular chemical environment). In order to fulfil transferability, atoms require a similar chemical environment and a common choice of local coordinate systems. These are identified from experimental geometry by examining the bond-distinguishing parameter χ of bonded atoms as detailed below, which is compared with values stored in the database. As we use local atomic site symmetry extensively, dummy atoms have to be calculated for some invarioms and in such cases the database contains a dummy atom in the coordinate-system definition.

[1] P_v, P_{lm}, κ and κ' are multipole parameters and are explained in §2.1.

Table 1
Invarioms assigned to the terminal group and the backbone atoms of the naturally occurring amino acids.

The superscripts are counters for invarioms and model compounds.

Atom	Invariom assigned	Site symmetry	Model compound
C^{α}	C1n1c1c1h^{1}	*m*	2-Aminopropane^{1}
Terminal C'	C1.5o1.5o1c^{-2}	*m*	Acetic acid anion^{2}
Peptide bond C'	C1.5o1.5n[1c1h]1c^{3}	*m*	Methylacetamide^{3}
Terminal N^{+}	N1c1h1h1h^{+4}	3	Methylamine cation^{4}
Terminal N	N1c1h1h^{5}	*m*	Methylamine^{5}
Peptide bond N'	N1.5c[1.5o1c]1c1h^{6}	*m*	Methylacetamide
Carboxylate O	O1.5c[1.5o1c]^{7}	*m*	Acetic acid anion
Peptide bond O'	O1.5c[1.5n1c]^{8}	*mm*2	Ethylamide^{6}
H@C^{α}	H1c[1n1c1c]^{9}	6	2-Aminopropane
H@ terminal N^{+}	H1n[1c1h1h]$^{+10}$	6	Methylamine cation
H@ terminal N	H1n[1c1h]^{11}	6	Methylamine
'Normal' C^{β}	C1c1c1h1h^{12}	*m*	Propane^{7}
H@C^{β}	H1c[1c1c1h]^{13}	6	Propane
Neutral carboxyl O1	O2c^{14}	*mm*2	Formaldehyde^{8}
Neutral carboxyl O2	O1c1h^{15}	*m*	Methanol^{9}
H@ oxygen	H1o[1c]^{16}	6	Methanol
Carboxyl C	C2o1o1c^{17}	*m*	Ethanol^{10}

We mimic the chemical environment in theoretical calculations of model compounds. These include nearest neighbours for single-bonded systems and next-nearest neighbours for delocalized/mesomeric systems (for example, carbon in benzene and, to a lesser extent, the atoms involved in the peptide bond) and for H atoms. When the procedure was originally introduced (Dittrich *et al.*, 2004), H atoms were modelled by nearest neighbours only. Recently, it has been found that a better fit to experimental or theoretical data and, more importantly, a reduced deviation from electroneutrality can be achieved by taking next-nearest neighbours into account for H atoms (Kingsford-Adaboh *et al.*, 2006). The outermost 'shell' of a model compound is usually saturated with H atoms and occasionally with larger fragments, such as C—H in the model compound benzene for the invariom C1.5c[1.5c1h]1.5c[1.5c1h]1h. For extended delocalized systems the entire delocalized part of a structure has to be calculated (see examples in Table 1).

Invariom multipole populations were obtained from B3LYP geometry optimizations for such model compounds with the D95++(3df,3pd) basis as available in *GAUSSIAN* (Frisch *et al.*, 1998). For these molecules, theoretical structure factors (cubic cell with $a = b = c = 30$ Å, space group $P\overline{1}$) were calculated (Chandler & Spackman, 1978) and a multipole refinement with *XDLSM* from the *XD* suite (Koritsánszky *et al.*, 2003) was performed. κ' parameters were kept at a value of 1 in our current refinements. The derivation of invarioms by double Fourier transform follows a procedure developed earlier (Koritsánszky *et al.*, 2002).

In invariom structure refinement each atom of a crystal structure is assigned an invariom. After the transfer process, the monopole populations (charges) are scaled to give electroneutrality for the asymmetric unit. As published earlier (Dittrich *et al.*, 2004), the difference between the number of valence electrons and the sum of monopole populations is usually negligible. There are currently several different ways to achieve electroneutrality in *INVARIOMTOOL* (Hübschle & Dittrich, 2004). In the present work, addition of the mean deviation from electroneutrality was applied only for H atoms, as for protein molecules differences are assumed to be absorbed by H atoms acting like a sponge for charge. For the examples given in §3.2, the largest relative charge difference for all trial structures occurred for *N*-acetyl-L-glutamic acid (Dobson & Gerkin, 1997) with +0.82 e (1.1%) for a total of 74 valence electrons in the structure, so that −0.07 e were subtracted from each of the 11 H-atom monopole populations. Details on invariom notation and the construction of the database of invariom multipole parameters have been published previously (Dittrich *et al.*, 2005). A publication about the *INVARIOMTOOL* preprocessor program is in preparation (Hübschle *et al.*, 2006). A complete list of all invarioms assigned to the atoms contained in amino acids and their side chains is given in §3.1.

2.2. Density modelling and refinement of example structures

Example structures in cif format were downloaded from the *Acta Crystallographica Section C: Crystal Structure Commnuications* or *Acta Crystallographica Section E: Structure Reports Online* web pages. For structures published in these journals, structure factors have been made available online since approximately 1996 and only those structures where structure factors were publicly available were chosen. The cifs were converted into *SHELX* format with the program *PLATON* (Spek, 2003) and initial IAM refinements with *SHELXL* (Sheldrick, 1997) reproduced the literature results. IAM refinements were repeated with the full-matrix multipole least-squares program *XDLSM* from the *XD* program package (Koritsánszky *et al.*, 2003). IAM refinements in *SHELXL* and *XDLSM* differ by the weighting scheme and the spherical scattering factors used. The *INVARIOMTOOL* preprocessor program (Hübschle & Dittrich, 2004) was then employed for invariom assignment and to modify master and input files for aspherical-atom refinement with *XDLSM*. This automated process was sufficient for all examples, except where refinement indices were very large and the geometry inaccurate or when single/double and mesomeric bonds give almost similar values for χ; the program allows manual invariom assignment for cases such as these. If a dummy atom has to be calculated for the atomic coordinate system, the program does so automatically. Only positional and thermal but no multipole parameters were refined and the residual minimized was $\varepsilon = \sum_{\mathbf{H}} [|F_o(\mathbf{H})| - |F_c(\mathbf{H})|]^2$ with $w(\mathbf{H}) = 1/\sigma^2[F_o(\mathbf{H})]$. A σ cutoff (3σ) was used. Depending on the quality and resolution of the experimental data, H atoms were either kept at orientations calculated from *SHELXL* and set to neutron distances or alternatively freely refined. For all data sets used in this work, anomalous dispersion was considered by standard corrections.[2]

[2] In cases where Friedel pairs for Mo $K\alpha$ light-atom structures were merged, the correction has only been applied to the merged data set.

3. Results and discussion

One main aim of this work was to validate the invariom database. Using example structures, it is demonstrated that all amino acids are described by the 73 entries. The extent to which the figures of merit improved for a particular structure was of secondary importance, although interesting new points emerge, especially from data sets of minor quality or disordered structures. After listing the invarioms that were assigned to the 20 mainly naturally occurring amino acids in proteins, results are presented for invariom refinements of 42 different data sets.

3.1. Invariom assignment to amino acids for aspherical-atom modelling

The atoms in the 20 naturally occurring amino acids in proteins require 73 different invarioms, including different possible protonation states and mesomeric structures, and details of their assignment are given in Tables 1 and 2. Selenocysteine was not included in the database and neither was the proposed new amino acid L-pyrrolysine (Hao *et al.*, 2002) owing to their rare occurrence, although their inclusion should be straightforward. Possible ambiguities initially arose for the mesomeric or delocalized ring systems histidine and tryptophan, where the assignment was verified with geometry-optimized model compounds to ensure that the invariom name is unique. These optimizations have shown that empirical values of $\chi = 0.0838$ and 0.184 are recommended for distinguishing between single and mesomeric and between mesomeric and double bonds in

$$\chi = [r_c(\text{atom1}) + r_c(\text{atom2}) - 0.08 \cdot |\Delta(\text{EN})|] - d, \quad (2)$$

where d is the bond distance, Δ(EN) the difference in the Allred–Rochow electronegativity (Allred & Rochow, 1958) and r_c is the covalent radius. A similar empirical relation (Schomaker & Stevenson, 1941) is also used to identify covalent bonds, as discussed previously (Dittrich *et al.*, 2005).

In Table 1 the main-chain and terminal group invarioms are given. For each atom, the invariom assigned is listed with the local atomic site symmetry that was manually chosen in the refinement against the theoretical structure factors of the model compounds. The common chemical name of the respective model compound is also given.

Invariom nomenclature has been described in detail previously (Dittrich *et al.*, 2005). Formal charges are indicated by a plus or minus at the end of the invariom name and are only given where a full charge can be assigned to the invariom and when the molecular model compound is an ion, such as CH_3COO^-. For ions where a spherical scattering factor is available, no aspherical density was modelled. Using the chloride ion Cl^- as an example, the charge (sign) is written behind the element name.

The separation between a single/mesomeric/double bond is arbitrary and not very distinctive for some delocalized systems such as arginine and histidine, but experience with a large number of X-ray structures and theoretically optimized model compounds led to the establishment of reliable values for the bond-distinguishing parameter χ as defined in (2). These values allow a sensible distinction between where to include next-nearest neighbours and where it is unnecessary. As an example, two model compounds were compared to represent the side chain of arginine. For the central C atom in the optimized model compound methylguanidine, the invariom assigned was found to be C1.5n[1h]1n1n with a χ value of 0.183 for the shortest C—N bond, whereas it is C2n1n1n in guanidine with a χ value of 0.187. Hence, for some mesomeric systems the value for the bond-distinguishing parameter χ can be very close above or below the threshold values for χ given earlier, so that the choice of the correct model compound for the database required care. Although the differences in the aspherical electron density are small between invarioms derived from nearest-neighbour and next-nearest-neighbour model compounds, the consequence for the automatization of invariom modelling is that low-resolution or low-quality X-ray data require occasional manual intervention in the assignment of the invariom name when such mesomeric systems occur.

Another challenge for automatization is disorder, because for the assignment of invariom name distances between partially occupied atoms should not be taken into account. We are currently working on an algorithm for the preprocessor program *INVARIOMTOOL* that allows the automatic detection of disorder. The algorithm will enable the detection of disorder by checking the sum of the formal charges as part of the invariom name and results will be reported in a subsequent paper. Table 2 lists the invarioms that were assigned to the side chains of the 20 amino acids in alphabetic order, analogous to Table 1.

3.2. Database validation

The database of 73 invarioms was verified on an extensive number of structures, where experimental intensities were available from the websites of the IUCr journals *Acta Crystallographica Section C: Crystal Structure Commnuications* and *Acta Crystallographica Section E: Structure Reports Online*. Trial structures were ordered according to their crystallographic *R* factor [*R*(F)] of the *XDLSM* IAM refinement and are listed in Table 3. Structures marked with an asterisk (*) are partly disordered. We tried to avoid disordered structures and the extent of disorder in the structures studied here is usually limited. For all structures listed in Table 3 an invariom as well as an IAM refinement was performed under identical refinement conditions. The figures of merit we considered most significant for this study were resolution, crystallographic *R* factor [*R*(F)], Δ (the maximal residual density) and DPI (Cruickshank's diffraction-component precision index; Cruickshank, 1999). These values are listed in Table 3 together with compound names, CSD refcode and their literature citation.

The DPI gives an approximation of the uncertainties of atomic coordinates $\sigma(x, B_{\text{avg}})$ and was calculated according to

Table 2
Invarioms assigned to side-chain atoms of the naturally occurring amino acids.

	Code†	Atom	Invariom assigned	Site symmetry	Model compound
Alanine	Ala, A	C$^{\beta}$	C1c1h1h1h[18]	3	Ethane[11]
		H@C$^{\beta}$	H1c[1h1h1h][19]	6	Ethane
Arginine	Arg, R	C$^{\gamma}$	C1c1c1h1h	*mm*2	Propane
		C$^{\delta}$	C1n1c1h1h[20]	1	Aminoethane[12]
		H@C$^{\delta}$	H1c[1n1c1h][21]	6	Aminoethane
		N$^{\varepsilon}$	N1c1c1h[22]	1	*N*,*N*-dimethylamine[13]
		H@N$^{\varepsilon}$	H1n[1c1c][23]	6	*N*,*N*-dimethylamine
		C$^{\zeta}$	C1.5n[1h]1n1n[24]	2	Methylguanidine[14]
		N$^{\nu 1}$	N1.5c[1n1n]1h[25]	*m*	Methylguanidine
		N$^{\nu 2}$	N1c1h1h	*m*	Methylamine
		H@N$^{\nu 1}$	H1n[1.5c][26]	6	Methaneimine[15]
		H@N$^{\nu 2}$	H1n[1c1h]	6	Methylamine
Arginine^{+}	Arg, R	N$^{\varepsilon}$	N1.5c[1.5n1.5n]1c1h^{+}[27]	*m*	Methylguanidinium cation[16]
		H@N$^{\varepsilon}$	H1n[1.5c1c]$^{+}$[28]	6	Methylguanidinium cation
		C$^{\zeta}$	C1.5n[1c1h]1.5n[1h1h]1.5n[1h1h]$^{+}$[28]	*mm*2	Methylguanidinium cation
		N$^{\nu 1}$	N1.5c[1.5n1.5n]1h1h^{+}[30]	*mm*2	Guanidinium cation[17]
		N$^{\nu 2}$	N1.5c[1.5n1.5n]1h1h^{+}	*mm*2	Guanidinium cation
		H@N$^{\nu 1,2}$	H1n[1.5c1h]$^{+}$[31]	6	Guanidinium cation
Asparagine	Asn, N	C$^{\gamma}$	C1.5o1.5n[1h1h]1c[32]	*m*	Acetamide[18]
		O$^{\delta 1}$	O1.5c[1.5n1c]	*mm*2	Acetamide
		N$^{\delta 2}$	N1.5c[1.5o1c]1h1h[33]	2	Acetamide
		H@N$^{\delta}$	H1n[1.5c1h][34]	6	Acetamide
Aspartic acid	Asp, D	C$^{\gamma}$	C2o1o1c	*m*	Ethanol
		O$^{\delta 1}$	O2c	*m*	Formaldehyde
		O$^{\delta 2}$	O1c1h	*m*	Methanol
		H@O$^{\delta 2}$	H1o[1c]	6	Methanol
Aspartate^{-}	Asp, D	C$^{\gamma}$	C1.5o1.5o1c^{-}	*m*	Acetic acid anion
		O$^{\delta 1}$	O1.5c[1.5o1c]	*m*	Acetic acid anion
		O$^{\delta 2}$	O1.5c[1.5o1c]	*m*	Acetic acid anion
Cysteine	Cys, C	C$^{\beta}$	C1s1c1h1h[35]	*m*	Ethanethiol[19]
		S$^{\gamma}$	S1c1h[36]	*m*	Methanethiol[20]
		H	H1s[1c][37]	6	Methanethiol
Cystine	Cys, C	S$^{\gamma}$	S1s1c[38]	*m*	Methylhydrodisulfide[21]
Glutamine	Gln, Q	C$^{\gamma}$	C1c1c1h1h	*mm*2	Propane
		H@C$^{\gamma}$	H1c[1c1c1h]	6	Propane
		C$^{\delta}$	C1.5o1.5n[1h1h]1c	*m*	Acetamide
		O$^{\varepsilon 1}$	O1.5c[1.5n1c]	*mm*2	Acetamide
		N$^{\varepsilon 2}$	N1.5c[1.5o1c]1h1h	2	Acetamide
		H@N$^{\varepsilon}$	H1n[1.5c1h]	6	Acetamide
Glutamic acid	Glu, E	C$^{\gamma}$	C1c1c1h1h	*mm*2	Propane
		H@C$^{\gamma}$	H1c[1c1c1h]	6	Propane
		C$^{\delta}$	C2o1o1c	*m*	Ethanol
		O$^{\varepsilon 1}$	O2c	*m*	Formaldehyde
		O$^{\varepsilon 2}$	O1c1h	*m*	Methanol
		H@O$^{\varepsilon}$	H1o[1c]	6	Methanol
Glutamate^{-}	Glu, E	C$^{\gamma}$	C1.5o1.5o1c^{-}	*m*	Acetic acid anion
		O$^{\varepsilon 1}$	O1.5c[1.5o1c]	*m*	Acetic acid anion
		O$^{\varepsilon 2}$	O1.5c[1.5o1c]	*m*	Acetic acid anion
Glycine	Gly, G	C$^{\alpha}$	C1n1c1h1h	1	Aminoethane
		H@C$^{\alpha}$	H1c[1n1c1h]	6	Aminoethane
Histidine 1	His, H	C$^{\gamma}$	C1.5n[1.5c]1.5c[1.5n1h]1c[39]	*m*	1-Methylimidazole[22]
		N$^{\delta 1}$	N1.5c[1.5c1c]1.5c[1.5n1h][40]	*m*	1-Methylimidazole
		C$^{\delta 2}$	C1.5n[1.5c1h]1.5c[1.5n1c]1h[41]	*m*	1-Methylimidazole
		C$^{\varepsilon 1}$	C1.5n[1.5c1h]1.5n[1.5c]1h[42]	*m*	Imidazole[23]
		N$^{\varepsilon 2}$	N1.5c[1.5n1h]1.5c[1.5c1h]1h[43]	*m*	Imidazole
Histidine 2	His, H	C$^{\gamma}$	C1.5n[1.5c1h]1.5c[1.5n1h]1c[44]	*m*	2-Methylimidazole[24]
		N$^{\delta 1}$	N1.5c[1.5c1c]1.5c[1.5n1h]1h[45]	*m*	2-Methylimidazole
		C$^{\delta 2}$	C1.5n[1.5c]1.5c[1.5n1c]1h[46]	*m*	2-Methylimidazole
		C$^{\varepsilon 1}$	C1.5n[1.5c1h]1.5n[1.5c]1h	*m*	Imidazole
		N$^{\varepsilon 2}$	N1.5c[1.5n1h]1.5c[1.5c1h][47]	*m*	Imidazole
Histidine^{+}	His, H	C$^{\gamma}$	C1.5c[1n1h]1n1c[48]	*m*	Methylimidazolium cation[25]
		N$^{\delta 1}$	N1.5c[1.5n1h]1c1h^{+} [49]	*mm*2	Methylimidazolium cation
		C$^{\delta 2}$	C1.5c[1n1c]1n1h^{+}	*m*	Methylimidazolium cation
		C$^{\varepsilon 1}$	C1.5n[1c1h]1.5n[1c1h]1h[50]	*mm*2	Imidazolium cation[26]
		N$^{\varepsilon 2}$	N1.5c[1.5n1h]1c1h^{+}	*mm*2	Methylimidazolium cation
Isoleucine	Ile, I	C$^{\beta}$	C1c1c1c1h[51]	3*m*	Isobutane[27]
		H@C$^{\beta}$	H1c[1c1c1c][52]	6	Isobutane
		C$^{\gamma 1}$	C1c1h1h1h	3	Ethane
		H@C$^{\gamma 1}$	H1c[1c1h1h]	6	Ethane
		C$^{\gamma 2}$	C1c1c1h1h	*mm*2	Propane
		H@C$^{\gamma 2}$	H1c[1c1c1h]	6	Propane

Table 2 (continued)

	Code†	Atom	Invariom assigned	Site symmetry	Model compound
		C^{δ}	C1c1h1h1h	3	Ethane
		$H@C^{\delta}$	H1c[1c1h1h]	6	Ethane
Leucine	Leu, L	C^{γ}	C1c1c1c1h	3*m*	Isobutane
		$H@C^{\gamma}$	H1c[1c1c1c]	6	Isobutane
		$C^{\delta 1}$	C1c1h1h1h	3	Ethane
		$C^{\delta 2}$	C1c1h1h1h	3	Ethane
		$H@C^{\delta}$	H1c[1h1h1h]	6	Ethane
Lysine	Lys, K	C^{γ}	C1c1c1h1h	*mm*2	Propane
		$H@C^{\gamma}$	H1c[1c1c1h]	6	Propane
		C^{δ}	C1c1c1h1h	*mm*2	Propane
		$H@C^{\delta}$	H1c[1c1c1h]	6	Propane
		C^{ε}	C1n1c1h1h	1	Aminoethane
		$H@C^{\varepsilon}$	H1c[1n1c1h]	6	Aminoethane
		N^{ζ}	N1c1h1h	*m*	Methylamine
		$H@C^{\zeta}$	H1n[1c1h]	6	Methylamine
Lysine^{+}	Lys, K	N^{ζ}	N1c1h1h1h^{+}	3	Methylamine cation
		$H@C^{\zeta}$	H1n[1c1h1h]$^{+}$	6	Methylamine cation
Methionine	Met, M	C^{γ}	C1s1c1h1h	*m*	Ethanethiol
		$H@C^{\gamma}$	H1c[1s1c1h]	6	Ethanethiol
		S^{δ}	S1c1c[53]	*mm*2	Dimethylsulfide[28]
		C^{ε}	C1s1h1h1h[54]	3	Methanethiol
		$H@C^{\varepsilon}$	H1c[1s1h1h][55]	6	Methanethiol
Phenylalanine	Phe, F	C^{γ}	C1.5c[1.5c1h]1.5c[1.5c1h]1c[56]	*mm*2	Toluene[29]
		$C^{\delta 1}$	C1.5c[1.5c1c]1.5c[1.5c1h]1h[57]	*mm*2	Toluene
		$C^{\delta 2}$	C1.5c[1.5c1c]1.5c[1.5c1h]1h	*mm*2	Toluene
		$C^{\varepsilon 1}$	C1.5c[1.5c1h]1.5c[1.5c1h]1h[58]	*mm*2	Benzene[30]
		$C^{\varepsilon 2}$	C1.5c[1.5c1h]1.5c[1.5c1h]1h	*mm*2	Benzene
		C^{ζ}	C1.5c[1.5c1h]1.5c[1.5c1h]1h	*mm*2	Benzene
		$H@C^{\delta,\varepsilon,\zeta}$	H1c[1.5c1.5c][59]	6	Benzene
Proline	Pro, P	C^{γ}	C1c1c1h1h	*mm*2	Propane
		C^{δ}	C1c1c1h1h	*mm*2	Propane
		$H@C^{\gamma,\delta}$	H1c[1c1c1h]	6	Propane
		C^{ε}	C1n1c1h1h	1	Aminoethane
		$H@C^{\gamma}$	H1c[1n1c1h]	6	Aminoethane
		N'	N1.5c[1.5o1c]1c1c[60]	*m*	*N*,*N*-dimethylacetamide[31]
		N_{term}	N1c1c1h1h^{+}	*mm*2	Dimethylamine cation[32]
Serine	Ser, S	C^{β}	C1o1c1h1h[61]	*m*	Ethanol
		$H@C^{\beta}$	H1c[1o1c1h][62]	6	Ethanol
		O^{γ}	O1c1h	*m*	Methanol
		$H@O^{\gamma}$	H1o[1c]	6	Methanol
Threonine	Thr, T	C^{β}	C1o1c1c1h[63]	*m*	2-Propenol
		$H@C^{\beta}$	H1c[1o1c1c][64]	6	2-Propenol
		$O^{\gamma 1}$	O1c1h	*m*	Methanol
		$H@O^{\gamma}$	H1o[1c]	6	Methanol
		$C^{\gamma 2}$	C1c1h1h1h	3	Ethane
		$H@C^{\gamma}$	H1c[1c1h1h]	6	Ethane
Tryptophan	Trp, W	C^{γ}	C1.5c[1.5c1.5c]1.5c[1n1h]1c[65]	*m*	3-Methylindole[33]
		$C^{\delta 1}$	C1.5n[1.5c1h]1.5c[1.5c1c]1h[66]	*m*	3-Methylpyrrole[34]
		$C^{\delta 2}$	C1.5c[1.5n1.5c]1.5c[1.5c1c]1.5c[1.5c1h][67]	*m*	3-Methylindole
		$N^{\varepsilon 1}$	N1.5c[1.5c1.5c]1.5c[1.5c1h]1h[68]	*m*	Indole[35]
		$C^{\varepsilon 2}$	C1.5n[1.5c1h]1.5c[1.5c1.5c]1.5c[1.5c1h][69]	*m*	Indole
		$C^{\varepsilon 3}$	C1.5c[1.5c1.5c]1.5c[1.5c1h]1h[70]	*m*	Naphthalene[36]
		$C^{\zeta 2}$	C1.5c[1.5n1.5c]1.5c[1.5c1h]1h[71]	*m*	Indole
		$C^{\zeta 3}$	C1.5c[1.5c1h]1.5c[1.5c1h]1h	*mm*2	Benzene
		$C^{\nu 2}$	C1.5c[1.5c1h]1.5c[1.5c1h]1h	*mm*2	Benzene
Tyrosine	Tyr, Y	C^{γ}	C1.5c[1.5c1h]1.5c[1.5c1h]1c	*mm*2	Toluene
		$C^{\delta 1}$	C1.5c[1.5c1c]1.5c[1.5c1h]1h	*mm*2	Toluene
		$C^{\delta 2}$	C1.5c[1.5c1c]1.5c[1.5c1h]1h	*mm*2	Toluene
		$C^{\varepsilon 1}$	C1.5c[1.5c1o]1.5c[1.5c1h]1h[72]	*mm*2	Phenol[37]
		$C^{\varepsilon 2}$	C1.5c[1.5c1o]1.5c[1.5c1h]1h	*mm*2	Phenol
		$H@O^{\delta,\varepsilon}$	H1c[1.5c1.5c]	6	Benzene
		C^{ζ}	C1.5c[1.5c1h]1.5c[1.5c1h]1o[73]	*mm*2	Phenol
		O^{ν}	O1c1h	*m*	Methanol
		$H@O^{\nu}$	H1o[1c]	6	Methanol
Valine	Val, V	C^{β}	C1c1c1c1h	3*m*	Isobutane
		$H@C^{\beta}$	H1c[1c1c1c]	6	Ethane
		$C^{\gamma 1}$	C1c1h1h1h	3	Ethane
		$C^{\gamma 2}$	C1c1h1h1h	3	Ethane
		$H@C^{\gamma}$	H1c[1h1h1h]	6	Ethane

† IUPAC standard one-letter and three-letter code (IUPAC–IUB Commission on Biochemical Nomenclature, 1970).

Table 3
Figures of merit and experimental conditions for 42 invariom structure refinements and comparison to IAM refinements.

Structures marked with an asterisk (*) are partly disordered and in those marked with a hash (#) we suspect dynamic disorder to occur.

	Compound	AAs†	Radiation	CSD refcode‡	Reference	$\sin\theta/\lambda_{\max}$ (Å^{-1})	DPI_{IAM} (Å)	Temp. (K)	$R(F)_{IAM}$	$R(F)_{inv}$	Δ_{IAM}§	Δ_{inv}
1	*N*-Acetyl-L-tyrosine ethyl ester·H_2O	Y	Mo *Kα*	ATYREE03	Dahaoui *et al.* (1999)	1.09	0.068	123	2.25	1.09	0.37	0.13
2	*Cis*-L-(5-oxo-L-prolyl)-L-prolinamide·H_2O¶	P	Cu *Kα*	TUDPAJ	Wouters *et al.* (1997)	0.62	0.271	293	2.29	1.95	0.11	0.10
3	L-Alanyl-L-threonine	A, T	Mo *Kα*	EWOVAN	Netland *et al.* (2004)	0.81	0.162	105	2.76	1.53	0.32	0.17
4	L-Phenylalanyl-L-valine	F, V	Mo *Kα*	XEGNAY	Görbitz (2000*b*)	0.66	0.268	150	2.77	1.48	0.16	0.11
5	L-Threonyl-L-alanine	T, A	Mo *Kα*	MAPKOE	Görbitz (2005)	0.66	0.302	100	2.84	1.93	0.19	0.14
6	Bis(L-tyrosinium) sulfate·H_2O*¶	Y	Mo *Kα*	MIFZIK	Sridhar *et al.* (2002*a*)	0.59	0.427	293	2.87	2.27	0.32	0.29
7	L-Seryl-L-valine	S, V	Mo *Kα*	EYIVAY	Moen *et al.* (2004)	0.64	0.306	105	2.91	2.16	0.18	0.14
8	*N*-Methyl-DL-aspartic acid·H_2O¶	D	Synchrotron	WOCVOZ	Madsen & Pattison (2000)	0.67	0.315	122	2.97	2.27	0.37	0.25
9	L-Alanyl-L-tryptophan·H_2O¶	A, W	Cu *Kα*	FUJZUF	Emge *et al.* (2000)	0.61	0.345	295	3.01	2.28	0.25	0.15
10	L-Valyl-L-phenylalanine	V, F	Mo *Kα*	MOBYAD	Görbitz (2002)	0.65	0.285	150	3.05	2.12	0.19	0.15
11	L-Valyl-L-serine·$3H_2O$	V, S	Mo *Kα*	FOBLUE	Johansen *et al.* (2005)	0.86	0.156	105	3.05	2.25	0.30	0.25
12	L-Tryptophan formic acid	W	Synchrotron	MUGKAA01	Scheins *et al.* (2004)	1.38	0.051	100	3.10	2.48	0.53	0.37
13	L-Cysteine	C	Mo *Kα*	LCYSTN04	Görbitz & Dalhus (1996)	0.86	0.191	120	3.11	2.91	0.44	0.43
14	L-Isoleucyl-L-isoleucine	I	Mo *Kα*	YAGZOW	Görbitz (2004*a*)	0.81	0.167	105	3.15	2.00	0.28	0.16
15	L-Asparaginyl-L-valine.$1\frac{1}{3}H_2O$	D, V	Mo *Kα*	FOBXAW	Bonge *et al.* (2005)	0.85	0.175	100	3.18	2.09	0.31	0.23
16	Bis(L-glutamic acid) sulfate.$\frac{1}{2}H_2O$	E	Mo *Kα*	FACXIR	Sridhar *et al.* (2002*b*)	0.59	0.391	293	3.27	3.02	0.34	0.34
17	Cyclo-(D,L-proline)$_2$-(L-alanine)$_4$·H_2O	A, P	Synchrotron	CAMVES01	Dittrich *et al.* (2002)	1.32	0.058	100	3.38	2.69	0.51	0.37
18	L-Asparaginium nitrate*¶	N	Mo *Kα*	MAPFIT	Aarthy *et al.* (2005)	0.59	0.681	293	3.40	3.33	0.21	0.20
19	DL-Arginine·H_2O	R+	Mo *Kα*	FUGXID	Kingsford-Adaboh *et al.* (2000)	0.62	0.285	100	3.47	2.06	0.31	0.27
20	DL-Alanyl-methionine	A, M	Mo *Kα*	ALAMET01	Guillot, Muzet *et al.* (2001)	1.00	0.106	100	3.49	2.45	0.74	0.33
21	Glycyl-L-aspartic acid·$2H_2O$	G, D	Mo *Kα*	BEVXEF01	Pichon-Pesme *et al.* (2000)	1.20	0.108	123	3.52	2.79	0.43	0.31
22	L-Isoleucyl-L-leucine·H_2O#	I, L	Mo *Kα*	ETITUW	Görbitz (2004*b*)	0.62	0.405	105	3.55	3.01	0.33	0.33
23	L-Seryl-L-phenylalanine	S, F	Mo *Kα*	PAJFIQ	Helle *et al.* (2004)	0.63	0.396	105	3.59	3.25	0.25	0.21
24	L-Leucyl-L-alanine.$4H_2O$	L, A	Mo *Kα*	RAVMOQ	Görbitz (1997)	0.87	0.160	150	3.62	2.84	0.36	0.25
25	Glycyl-L-threonine·$2H_2O$#	G, T	Mo *Kα*	GLYTRE03	Benabicha *et al.* (2000)	1.15	0.109	100	3.73	2.96	0.50	0.44
26	L-Alanyl-L-methionine.$\frac{1}{2}H_2O$	A, M	Mo *Kα*	EMIPAR	Görbitz (2003)	0.64	0.470	105	3.74	3.38	0.30	0.25
27	L-Argininium chloride*	R+	Mo *Kα*	LARGIN02	Sridhar *et al.* (2002*c*)	0.82	0.194	293	3.85	3.56	0.40	0.42
28	Glycyl-L-tryptophan·$2H_2O$¶	G, W	Cu *Kα*	GLTRH01	Emge *et al.* (2000)	0.61	0.510	295	4.04	3.39	0.23	0.20
29	L-Phenylalanyl-L-alanine·$2H_2O$	F, A	Mo *Kα*	QIMBUJ	Görbitz (2001)	0.81	0.227	150	4.05	3.81	0.44	0.38
30	L-Valyl-L-glutamine	V, Q	Mo *Kα*	TIPTOB	Görbitz & Backe (1996)	0.70	0.445	120	4.43	3.98	0.32	0.26
31	Glycyl-DL-leucine¶	G, L	Mo *Kα*	XEGHOG	Bombicz *et al.* (2000)	0.58	0.439	120	4.62	4.13	0.21	0.15
32	L-Tryptophan formic acid	W	Mo *Kα*	MUGKAA	Hübschle *et al.* (2002)	0.77	0.345	183	4.71	3.83	0.40	0.30
33	Bis(L-proline) nitrate¶	P	Mo *Kα*	LUDFOF	Pandiarajan *et al.* (2002)	0.61	0.644	293	4.74	4.29	0.30	0.29
34	L-Glutaminyl-L-valine	Q, V	Mo *Kα*	TIPTVH	Görbitz & Backe (1996)	0.70	0.465	120	4.82	4.45	0.34	0.30
35	L-Tyrosyl-glycyl-glycine·H_2O	Y, G	Mo *Kα*	LTYRGG01	Pichon-Pesme *et al.* (2000)	1.15	0.164	123	4.83	3.71	0.80	0.58
36	L-Histidinium L-histidine glutarate*¶	H, H+	Cu *Kα*	ADAVUW	Saraswathi & Vijayan (2001)	0.63	0.470	293	5.49	5.50	0.35	0.41
37	*N*-Acetyl-L-glutamic acid¶	E	Mo *Kα*	TERRUD	Dobson & Gerkin (1997)	0.65	0.579	296	5.51	5.33	0.32	0.31
38	L-Leucyl-L-phenylalanine.2-propanol#	L, F	Mo *Kα*	COCGOQ	Görbitz (1999)	0.86	0.227	150	5.54	5.07	0.48	0.42
39	Hippuryl-L-histidinyl-L-leucine.$5H_2O$	H+, L	Mo *Kα*	FACCIV10	Vrielink *et al.* (1996)	0.54	1.049	293	5.55	5.42	0.24	0.26
40	L-Seryl-L-alanine	S, A	Mo *Kα*	KIYHOP	Görbitz (2000*a*)	0.81	0.369	153	5.92	4.93	0.49	0.41
41	L-Aspartic acid monohydrate	D	Mo *Kα*	IJEQET	Umadevi *et al.* (2003)	0.59	0.840	293	6.14	5.73	0.36	0.32
42	L-Lysine L-lysinium dichloride nitrate#	K, K+	Mo *Kα*	BOQWOT	Srinivasan *et al.* (2001)	0.59	1.136	293	6.15	6.02	0.48	0.50

† Amino-acid residues. ‡ CSD version 5.27. § Δ is the positive residual electron density. ¶ Extinction reported. For L-Asp·H_2O (No. 41), an extinction parameter can be refined fulfilling the criteria mentioned in the text, although it has not been reported in the literature.

$$\mathrm{DPI} = \sigma(x, B_{\mathrm{avg}}) = \left(\frac{N_i}{n_{\mathrm{obs}} - n_{\mathrm{par}}}\right)^{1/2} \cdot C^{1/3} \cdot R(F) \cdot d_{\min}, \quad (3)$$

where N_i is the number of atoms of type i possessing a scattering power $\bar{s}$ similar to the j atoms in the asymmetric unit ($\sum_j f_j^2 = N_i f_i^2$), n_{obs} and n_{par} are the number of observations and parameters, C is the percentage completeness and $d_{\min}$ is the resolution of the experiment [3].

Fig. 1(*a*) shows the crystallographic *R* factor for IAM and invariom models for all trial structures in the same order as in

[3] N_i was calculated by dividing the sum of the squares of the atomic number (number of electrons) of the atoms i in the asymmetric unit ($\sum_i^N Z_i^2$) by Z_C^2 (36). This assumes the average scattering contribution to be that of a C atom and N_i to be the equivalent number of C atoms in the asymmetric unit as in Allen *et al.* (1995). Further information on the structural data used to calculate the DPI can be found in the supplementary data, which have been desposited in the IUCr electronic archive (Reference: VR5059). Services for accessing these data are described at the back of the journal.

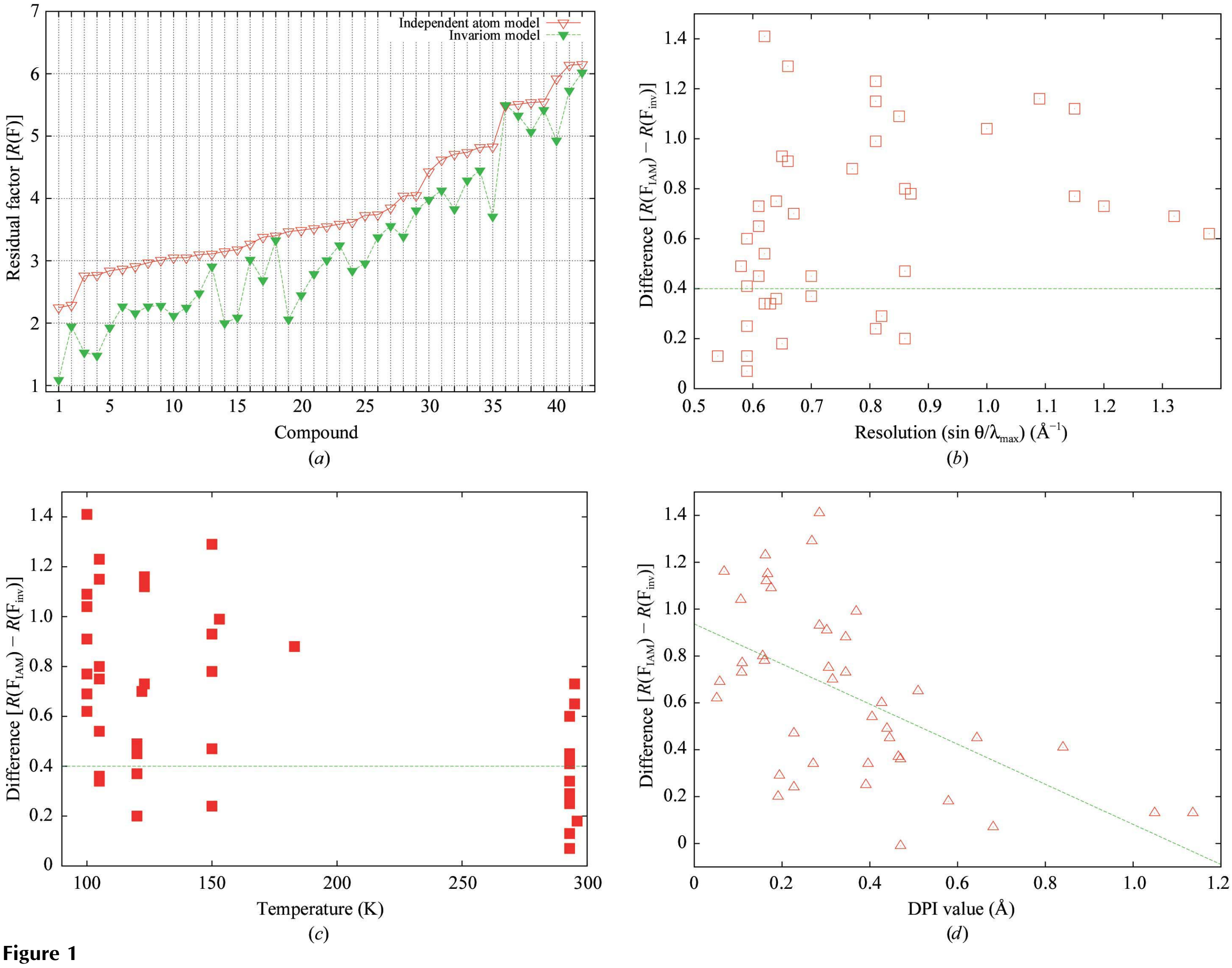

Figure 1
(*a*) Comparison of the crystallographic *R* factor between IAM and invariom model. (*b*) Difference of the *R* factor between IAM and invariom model plotted *versus* resolution and (*c*) *versus* temperature of the structures given above. (*d*) Difference $R(F_{IAM}) - R(F_{inv})$ plotted *versus* DPI as calculated from (3). A correlation can be seen and an expected improvement of the *R* factor can be predicted.

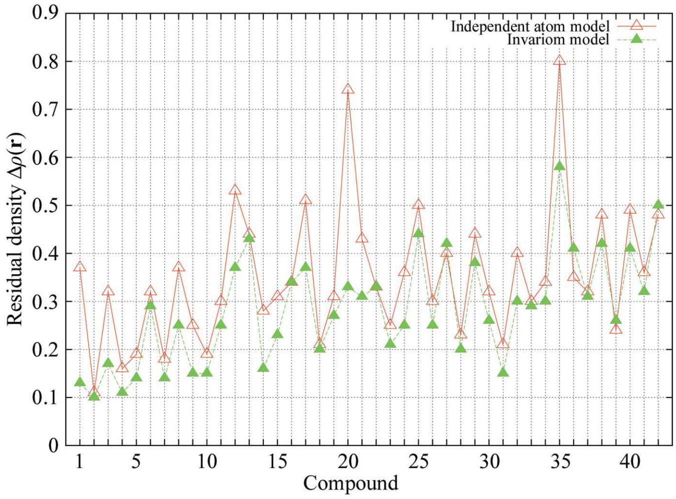

Figure 2
Comparison of the residual density $\Delta\rho(\mathbf{r})$ between IAM and invariom model.

Table 3. It is obvious that *R*(F) is always equal to or better for the invariom model when compared with the IAM. Cases where the difference $R(F_{IAM}) - R(F_{inv})$ are small (<0.4%) will be discussed: low-resolution data sets Nos. 2 and 39 (Wouters *et al.*, 1997; Vrielink *et al.*, 1996) collected at room temperature both show a small difference of $R(F_{IAM}) - R(F_{inv})$. As the agreement factors are already comparably low for these two structures, this is probably a consequence of the measurement conditions. For the next group of examples where the difference $R(F_{IAM}) - R(F_{inv})$ is small, static disorder occurs. These structures, where disorder has been already detected and modelled in the literature are No. 6, bis(L-tyrosinium) sulfate·H_2O (Sridhar *et al.*, 2002*a*), No. 22, L-isoleucyl-L-leucine (Görbitz, 2004*b*), No. 38, L-leucyl-L-phenylalanine (Görbitz, 1999), and No. 42, L-lysine L-lysinium (Srinivasan *et al.*, 2001). In structure No. 13, monoclinic L-cysteine-II, the authors discuss possible disorder of the H atom of the thiol group. Disorder has also been reported in the orthorhombic phase I of L-cysteine at ambient conditions (Moggach *et al.*,

2005). The last group of examples are structure No. 16, *bis*(L-glutamic acid) sulfate (Sridhar *et al.*, 2002*b*), No. 18, L-asparaginium nitrate (Aarthy *et al.*, 2005), No. 27, L-argininium chloride (Sridhar *et al.*, 2002*c*), and No. 36, L-histidinium L-histidine glutarate, where it is possible that dynamic hydrogen disorder occurs so that the structural model is incomplete. All have hydrogen-bond patterns where mobile protons could compete with counter-ions to temporarily compensate charges by forming $NH_3^+ \cdots COO^-$ links, in the process neutralizing two otherwise separated charges of the amino and carboxylate groups. From the earlier examples and these 'suspect' structures (as marked with a # in Table 3), we can conclude that when disorder is present in a structure, the modelling of disorder largely determines the agreement between F_{obs} and F_{calc} and the more sophisticated scattering model does not improve the situation significantly in such cases. This has important implications for modelling of protein data, which is usually strongly affected by disorder, and suggests that improved treatment of disorder is a desirable feature for future software developments for the refinement of protein data. Recent substantial advances have been summarized (Guillot, Viry *et al.*, 2001; Jelsch *et al.*, 2005).

In Fig. 1(*b*) the difference $R(\mathrm{F_{IAM}}) - R(\mathrm{F_{inv}})$ is plotted against the resolution of the respective experiments. A threshold of $R(\mathrm{F_{IAM}}) - R(\mathrm{F_{inv}}) = 0.4$ separates the examples discussed before from most of the other structures. Adding to our earlier findings, the improvement from using aspherical scattering factors does not depend much on the resolution of the experiment if it is higher than $\sin\theta/\lambda = 0.6$ Å^{-1} ($d = 0.83$ Å; Dittrich *et al.*, 2005). On the other hand, an interesting observation directly related to resolution is that the improvement arising from the invariom valence density with high-resolution data is less good than for some low-order cases, probably to a small degree owing to the crystal field effect, but more importantly owing to the receding contribution of valence in relation to total scattering. The unmodelled contribution of anharmonic thermal motion of the N atom could also limit a possible improvement of figures of merit for at least some of the compounds studied. We are currently investigating this effect on a suitable data set.

Although model bias (Bränden & Jones, 1990) is not an issue with the small-molecule structures studied here, calculation of R_{free} (Brünger, 1992) will become useful or even necessary for the application of invarioms to larger protein structures [4].

In Fig. 1(*c*) we have investigated the temperature dependence of the difference $R(\mathrm{F_{IAM}}) - R(\mathrm{F_{inv}})$. The results from the structures investigated in this paper also support earlier findings; the higher the temperature of an experiment, the smaller the possible improvement, which suggests that most of today's experiments should aim for the lowest possible temperature.

Fig. 1(*d*) shows the correlation between the difference $R(\mathrm{F_{IAM}}) - R(\mathrm{F_{inv}})$ and the DPI. The linear fit [$m = -0.9$ (2), $b = 0.94$ (8)] to the DPI enables a rough estimate of the improvement of the *R* factor that can be expected from invariom modelling from the DPI of a structure. It is interesting to note that the approximate coordinate error as predicted by the DPI is reduced as much as the *R* factor is by invarioms. Only the value for the IAM refinement $\mathrm{DPI_{IAM}}$ is given in Table 3, as the DPI using the invariom *R* factor can be obtained by scaling $\mathrm{DPI_{IAM}}$ with $R(\mathrm{F_{inv}})/R(\mathrm{F_{IAM}})$. Concerning the resolution of an experiment, Blow found that a rearrangement of the DPI formula (3) reveals a dependence on $(\mathrm{resolution})^{5/2}$ of the DPI (Blow, 2002). From (3) it can be concluded that the amount of improvement of the conventional *R* factor is an indicator of data quality in terms of information content, *i.e.* the fine details of the electron-density distribution. In light of the discussed effects of disorder, resolution and temperature, the DPI nicely summarizes the quality of a structure. Hence, the best possible resolution and lowest possible temperature should be aimed for in an experiment.

Fig. 2 compares the positive residual electron density for IAM and invariom models. Here, the most interesting feature can be observed for the high-resolution structure of DL-alanyl-methionine (No. 20) containing sulfur (Guillot, Muzet *et al.*, 2001), where the residual density for the invariom model is considerably reduced and becomes comparable to the other structures that do not contain sulfur. In structures with disorder an analogous behaviour is seen for $\Delta\rho(\mathbf{r})$ and the reduction of the *R* factor. For these cases, the residual density is similar or even increased for the invariom model.

Only three of the example structures crystallize in centrosymmetric space groups. In non-centrosymmetric space groups phases can be not well defined and there are roughly only half as many reflections available per refined parameter. However, we did not observe an influence of the space group on the figures of merit of the structures studied.

One can conclude that the remaining unmodelled density causes the disappointingly small improvement in the *R* factor for disordered structures. In other words, invariom modelling enhances the signal in disordered regions and should have an impact on structure validation. In cases where the figures of merit do not improve, the structural model should be revised, since it is likely that unresolved disorder occurred. Enhancement of the Fourier signal with the invariom model should allow an improvement of the structural model beyond the promolecule signal, especially in heavily disordered structures, and although an improvement of the trial structures was not within the scope of this work, we will pursue this topic in future research. Nevertheless, we marked structures where hydrogen or other kind of disorder is likely to have occurred with an asterisk in Table 3. Improving the structural model will be a necessity when modelling protein data, where the resolution of an experiment is the limiting factor. It is noteworthy that disorder seems to be a lot more common than marked in the Cambridge Structural Database.

[4] To obtain an indication of data quality, we attempted to relate the $I/\sigma(I)$ ratio of the outermost resolution shell to the difference $R(\mathrm{F_{IAM}}) - R(\mathrm{F_{inv}})$, but no correlation was found.

When extinction was reported in the literature (as marked in Table 3), we attempted to refine isotropic extinction assuming a Gaussian distribution in promolecule and invariom model. In most of these refinements we found that extinction was not really present, as either the extinction parameter became zero, the *R* factor remained constant and the residual density increased for both models.[5] It therefore seems that a correction for extinction is unnecessary in a considerable part of the cases where an extinction parameter can be successfully refined in the promolecule model. Two special cases are discussed. In the structure of glycyl-DL-leucine (Bombicz *et al.*, 2000), both invariom and IAM *R* factors as well as the invariom residual densities are reduced and extinction effects appear to be real. For *N*-methyl-DL-aspartic acid measured with synchrotron radiation (Madsen & Pattison, 2000) both *R* factors and residual densities increase when the significant extinction parameter is refined: possibly low-order reflections are affected by systematic errors. To summarize, extinction appears to be less frequent than reported for the organic molecules studied here. Testing invariom and promolecule *R* factors together with invariom residual densities gives more information on whether or not extinction effects are real.

4. Conclusion

An invariom database of intermolecular transferable pseudoatoms (equivalent to individual aspherical form factors) for amino-acid, oligopeptides and protein molecules has been generated from *ab initio* calculations of small model compounds *via* theoretical structure factors. For evaluation purposes, the database has been applied to 42 example structures where experimental structure factors were available from IUCr journals covering the naturally occurring amino acids, some of their derivatives or their protonated/unprotonated states and also most common solvents. In order to apply invariom modelling for standard small-molecule structures, no further calculations nor extra experimental procedures are necessary, making it a rapid, easily accessible and useful tool for standard crystallographic work. Invariom modelling usually reduces the *R* factor and other figures of merit and, since the electron density is imposed, allows better deconvolution of electronic and thermal effects. This holds also for H atoms, resulting in an improved description of their geometry.

The following main points emerged from the study of the example structures. (i) Extinction appears to be a lot less common than reported in the literature. (ii) An important conclusion relevant for protein refinement can be drawn from the example structures that were disordered: it is the modelling of disorder and the completeness of the structural model rather than the aspherical electron-density contribution that limit the fit of calculated to experimental structure factors and therefore the quality of the results. (iii) The DPI value nicely summarizes the information content a structure can be expected to provide. It is proportional to the improvement of the *R* factor of the invariom refinement.

We intend to investigate ultrahigh-resolution protein data in the near future in order to refine these data with the invariom aspherical scattering model. Information gathered in this study provides criteria to assess the suitability of protein structural data to invariom refinement. An improved molecular geometry, the location of more hydrogen positions in Fourier maps and physically more meaningful thermal parameters are anticipated. For invariom modelling of protein data, a resolution of $d \leq 0.9$ Å or $\sin\theta/\lambda \geq 0.55$ Å^{-1} and a low DPI are recommended. Pending the development of improved algorithms for modelling of dynamic disorder, such results can realistically be expected.

This work was supported by the Australian Synchrotron Research Program, which is funded by the Commonwealth of Australia under the Major National Research Facilities Program. BD thanks the ASRP for a postdoctoral fellowship. We thank Dr C. R. Pulham and Dr F. P. A. Fabbiani (CSEC, Edinburgh) for support, S. Siemsen for help in preparing the manuscript and one of the referees for suggesting studying the DPI. We also acknowledge support from the Deutsche Forschungsgemeinschaft within the SPP 1178 project.

References

Aarthy, A., Anitha, K., Athimoolam, S., Bahadurb, S. A. & Rajaram, R. (2005). *Acta Cryst.* E**61**, o2042–o2044.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995). *Acta Cryst.* A**51**, 95–111.

Allred, A. L. & Rochow, E. G. (1958). *J. Inorg. Nucl. Chem.* **5**, 264–268.

Benabicha, F., Pichon-Pesme, V., Jelsch, C., Lecomte, C. & Khmou, A. (2000). *Acta Cryst.* B**56**, 155–165.

Blow, D. M. (2002). *Acta Cryst.* D**58**, 792–797.

Bombicz, P., Dittrich, B., Strümpel, M., Nabein, H.-P. & Luger, P. (2000). *Acta Cryst.* C**56**, 1447–1449.

Bonge, H. T., Rosenberg, M. L., Riktor, M. & Görbitz, C. H. (2005). *Acta Cryst.* E**61**, o524–o527.

Bränden, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 678–689.

Brock, C. P., Dunitz, J. D. & Hirshfeld, F. L. (1991). *Acta Cryst.* B**47**, 789–797.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Chandler, G. S. & Spackman, M. A. (1978). *Acta Cryst.* A**34**, 341–343.

Coppens, P. (1997). *X-Ray Charge Densities and Chemical Bonding*, 1st ed. Oxford University Press.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Dahaoui, S., Jelsch, C., Howard, J. A. K. & Lecomte, C. (1999). *Acta Cryst.* B**55**, 226–230.

Dittrich, B., Hübschle, C. B., Messerschmidt, M., Kalinowski, R., Girnt, D. & Luger, P. (2005). *Acta Cryst.* A**61**, 314–320.

Dittrich, B., Koritsánszky, T., Grosche, M., Scherer, W., Flaig, R., Wagner, A., Krane, H.-G., Kessler, H., Riemer, C., Schreurs, A. M. M. & Luger, P. (2002). *Acta Cryst.* B**58**, 721–727.

Dittrich, B., Koritsánszky, T. & Luger, P. (2004). *Angew. Chem. Int. Ed.* **43**, 2718–2721.

Dobson, A. J. & Gerkin, R. E. (1997). *Acta Cryst.* C**53**, 73–76.

Emge, T. J., Agrawal, A., Dalessio, J. P., Dukovic, G., Inghrim, J. A., Janjua, K., Macaluso, M., Robertson, L. L., Stiglic, T. J., Volovika, Y. & Georgiadis, M. M. (2000). *Acta Cryst.* C**56**, e469–e471.

[5] The IAM residual density is increased by introducing an extinction parameter for most of the examples studied including glycyl-DL-leucine; using a unit weighting scheme, we could not find an indication for extinction effects for most cases where is has been reported.

Frisch, M. J. *et al.* (1998). *GAUSSIAN* 98, revision A.7. Gaussian Inc., Pittsburgh, PA, USA.

Görbitz, C. H. (1997). *Acta Cryst.* C**53**, 736–739.

Görbitz, C. H. (1999). *Acta Cryst.* C**55**, 2171–2177.

Görbitz, C. H. (2000*a*). *Acta Cryst.* C**56**, 500–502.

Görbitz, C. H. (2000*b*). *Acta Cryst.* C**56**, 1496–1498.

Görbitz, C. H. (2001). *Acta Cryst.* C**57**, 575–576.

Görbitz, C. H. (2002). *Acta Cryst.* B**58**, 512–518.

Görbitz, C. H. (2003). *Acta Cryst.* C**59**, o730–o732.

Görbitz, C. H. (2004*a*). *Acta Cryst.* B**60**, 569–577.

Görbitz, C. H. (2004*b*). *Acta Cryst.* E**60**, o626–o628.

Görbitz, C. H. (2005). *Acta Cryst.* E**61**, o2012–o2014.

Görbitz, C. H. & Backe, P. H. (1996). *Acta Cryst.* B**52**, 999–1006.

Görbitz, C. H. & Dalhus, B. (1996). *Acta Cryst.* C**52**, 1756–1759.

Guillot, R., Muzet, N., Dahaoui, S., Lecomte, C. & Jelsch, C. (2001). *Acta Cryst.* B**57**, 567–578.

Guillot, B., Viry, L., Guillot, R., Lecomte, C. & Jelsch, C. (2001). *J. Appl. Cryst.* **34**, 214–223.

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

Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A. & Chan, M. K. (2002). *Science*, **296**, 1462–1466.

Helle, I. H., Løkken, C. V., Görbitz, C. H. & Dalhus, B. (2004). *Acta Cryst.* C**60**, o771–o772.

Housset, D., Benabicha, F., Pichon-Pesme, V., Jelsch, C., Maierhofer, A., David, S., Fontecilla-Camps, J. C. & Lecomte, C. (2000). *Acta Cryst.* D**56**, 151–160.

Hübschle, C. B. & Dittrich, B. (2004). *INVARIOMTOOL, a Preprocessor Program for Aspherical Atom Modelling with XD Using Invarioms.* Freie Universität Berlin, Berlin, Germany.

Hübschle, C. B., Dittrich, B. & Luger, P. (2002). *Acta Cryst.* C**58**, o540–o542.

Hübschle, C. B., Dittrich, B. & Luger, P. (2006). In the press.

IUPAC–IUB Commission on Biochemical Nomenclature (1970). *Biochemistry*, **9**, 3471–3479.

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

Jelsch, C., Pichon-Pesme, V., Lecomte, C. & Aubry, A. (1998). *Acta Cryst.* D**54**, 1306–1318.

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 3171–3176.

Johansen, A., Midtkandal, R., Roggen, H. & Görbitz, C. H. (2005). *Acta Cryst.* C**61**, o198–o200.

Kingsford-Adaboh, R., Dittrich, B., Hübschle, C. B., Gbewonyo, W. S. K., Okamoto, H., Kimura, M. & Ishida, H. (2006). In the press.

Kingsford-Adaboh, R., Grosche, M., Dittrich, B. & Luger, P. (2000). *Acta Cryst.* C**56**, 1274–1276.

Koritsánszky, T., Richter, T., Macchi, P., Volkov, A., Gatti, C., Howard, S., Mallinson, P. R., Farrugia, L., Su, Z. W. & Hansen, N. K. (2003). *XD: A Computer Program Package for Multipole Refinement and Topological Analysis of Electron Densities from Diffraction Data.* Freie Universität Berlin, Berlin, Germany.

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

Kurki-Suonio, K. (1977). *Isr. J. Chem.* **16**, 115–123.

Madsen, D. & Pattison, P. (2000). *Acta Cryst.* C**56**, 1157–1158.

Moen, A., Frøseth, M., Görbitz, C. H. & Dalhus, B. (2004). *Acta Cryst.* C**60**, o564–o565.

Moggach, S. A., Clark, S. J. & Parsons, S. (2005). *Acta Cryst.* E**61**, o2739–o2742.

Muzet, N., Guillot, B., Jelsch, C., Howard, E. & Lecomte, C. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 8742–8747.

Netland, K. A., Andresen, K., Görbitz, C. H. & Dalhus, B. (2004). *Acta Cryst.* E**60**, o951–o953.

Pandiarajan, S., Sridhar, B. & Rajaram, R. K. (2002). *Acta Cryst.* E**58**, o862–o864.

Petrova, T. & Podjarny, A. (2004). *Rep. Prog. Phys.* **67**, 1565–1605.

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

Pichon-Pesme, V., Lachekar, H., Souhassou, M. & Lecomte, C. (2000). *Acta Cryst.* B**56**, 728–737.

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

Saraswathi, N. T. & Vijayan, M. (2001). *Acta Cryst.* B**57**, 842–849.

Scheins, S., Dittrich, B., Messerschmidt, M., Paulmann, C. & Luger, P. (2004). *Acta Cryst.* B**60**, 184–190.

Schomaker, V. & Stevenson, D. P. (1941). *J. Am. Chem. Soc.* **63**, 37–40.

Sheldrick, G. M. (1997). *SHELXL-97. A Program for Refinement of Crystal Structures.* University of Göttingen, Germany.

Spek, A. L. (2003). *J. Appl. Cryst.* **36**, 7–13.

Sridhar, B., Srinivasan, N. & Rajaram, R. K. (2002*a*). *Acta Cryst.* E**58**, o211–o214.

Sridhar, B., Srinivasan, N. & Rajaram, R. K. (2002*b*). *Acta Cryst.* E**58**, o272–o276.

Sridhar, B., Srinivasan, N. & Rajaram, R. K. (2002*c*). *Acta Cryst.* E**58**, o1372–o1374.

Srinivasan, N., Sridhar, B. & Rajaram, R. K. (2001). *Acta Cryst.* E**57**, o888–o890.

Stewart, R. F. (1969). *J. Chem. Phys.* **51**, 4569–4577.

Umadevi, K., Anitha, K., Sridhar, B., Srinivasan, N. & Rajaram, R. K. (2003). *Acta Cryst.* E**59**, o1073–o1075.

Volkov, A., Koritsánszky, T., Li, X. & Coppens, P. (2004). *Acta Cryst.* A**60**, 638–639.

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

Vrielink, A., Obel-Jorgensen, A. & Codding, P. W. (1996). *Acta Cryst.* C**52**, 1300–1302.

Wouters, J., Norberg, B. & Evrard, G. (1997). *Acta Cryst.* C**53**, 477–480.