International Tables for Crystallography (2006). Vol. F, Section 18.3.2, pp. 382–392.

18.3. Structure quality and target parameters

BY R. A. ENGH AND R. HUBER

18.3.1. Purpose of restraints

> Believe statistics! Shun bias! – these, we see, are two materially different laws. (Adapted from William James.) A wise man, therefore, proportions his restraints to the evidence. (Adapted from David Hume.)

If we could adequately measure the complex diffuse scattering function of an X-ray beam from a single macromolecular structure to arbitrary resolution, we could, given an accurate model for X-ray scattering, calculate a measurement-time-averaged electron-density distribution for that structure independent of most model assumptions. Alternatively, if we knew the potential energy for the system accurately as a function of all relevant parameters, classical and non-classical, and had the computational power to analyse the function appropriately, we could model the same distribution. If both experiment and theory were accessible at these extremes, the available information would be redundant and could be used together in arbitrary ways. Of course, both experiment and theory are limited far below these extremes, and the researcher must consider what information is available in order to address the pertinent protein-structure question in the most appropriate way.

The most general theoretical assumptions used to interpret experimentally determined X-ray reflections typically include the spherical symmetry of scattering functions centred at nuclei, harmonicity (and isotropy) of deviations about average positions, infinity of a crystal lattice of identical unit cells and so on. While the majority of structures determined to date share a standard set of assumptions, significant alternatives have been proposed, such as, for example, the modelling of motion with normal-mode analyses instead of harmonic temperature factors (Kidera *et al.*, 1994). Any significant change in the set of assumptions used in structure determination must be expected to leave its mark on the statistical properties of the structure.

Besides the necessity of general model assumptions for structure determination, there is the simple necessity to supplement the measured reflection data with additional sources of information to improve the ratio of experimental observables to model parameters. For example, a crystal of a protein with some 3000 atoms that diffracts to a moderate resolution of 2.5 Å might produce some 14 000 measurable unique reflections. In this case, fitting a model that includes three Cartesian coordinates and a temperature factor for each atom (12 000 parameters) is effectively underdetermined when considering experimental error. The solution is either to reduce the number of parameters by introducing constraints to the model, or to increase the number of effective observations by using additional restraints in the structure determination.

18.3.1.1. *Utility of restraints: protein/special geometries*

One set of restraints arises from the expectation that the macromolecular structure should be chemically reasonable, that is, it should reflect the geometries required by chemical physics. This restricts especially values of bond lengths, bond angles, planarity and improper dihedral angles. Additional restriction comes from the expectation that these values should most closely conform to structures determined by the same experimental method and corresponding model assumptions. There are now two sources of structural information of sufficient quality for use as restraints in structure refinement: chemical fragments from the Cambridge Structural Database (CSD) (Chapter 23.4) and the very high resolution protein structures that can be refined without recourse to restraining parameters (Longhi *et al.*, 1998; Dauter *et al.*, 1997). Restraints derived from the CSD have come into widespread use for protein (Engh & Huber, 1991; Brünger, 1993; Priestle, 1994) and nucleic acid (Parkinson *et al.*, 1996) refinement; since their derivation, the database has grown from some 80 000 structures to around 200 000. The number of very high resolution protein structures solved has also increased to include dozens of structures (or thousands of chemical fragments). Chemical-fragment geometries can be studied using either or both sets of structures.

The optimal choice of restraints to use depends on the character and expected use of the refined structure. For example, the quality of comparative studies (Kleywegt & Jones, 1998) is enhanced by standardized refinement conditions, removing at least one source of systematic variation (*e.g.* Laskowski, Moss & Thornton, 1993), although effects from individual crystals and data collection and processing remain. On the other hand, a structure with an unusual chemical environment might require a tailored parameterization scheme to study the effects of this environment. Such cases might occur, for example, at enzyme active sites, where bound ligands may be distorted toward intermediate geometries (Bürgi & Dubler-Steudle, 1988).

18.3.1.2. *Risk of restraints: bias, lack of cross validation*

As an alternative to implementation as restraints, statistical information from known structures provides an independent check of structural integrity. However, the more information which is used for restraints, the less is available for such cross validation. For example, the use of a force field in protein refinement that strictly enforces a physical distribution of the protein backbone φ–ψ angles of the Ramachandran plot would transfer error-induced strain into other degrees of freedom while eliminating a Ramachandran analysis as a tool for judging protein quality (Hooft *et al.*, 1997). There is an additional risk associated with the intentional introduction of bias into protein refinement: Restraining a model to conform to known structures is in error if the structure is unique in some way and therefore should *not* conform to these structures, or if there is erroneous bias in the set of structures from which the restraints were derived. Furthermore, the bias may be exaggerated, especially if the biased feature is so probable that deviations are considered suspect. This can be seen, for example, in the increasing frequency of *cis*-prolines as protein-structure resolution increases (Stewart *et al.*, 1990): when there is uncertainty in lower-resolution structures, the more probable *trans*-prolines are preferentially chosen in model building, exaggerating their frequency.

18.3.2. Formulation of refinement restraints

A priori information regarding protein structure can be used in two fundamental ways: as constraints by fixing parameters to target values, or as restraints by allowing limited deviation from target values. The two methods differ fundamentally when counting the total number of degrees of freedom (important, for example, when thermodynamic quantities of simulations are considered), but both improve the observation-to-parameter ratio (constraints reduce parameters, restraints increase observations). In this chapter, we focus on the use of restraints.

A common form of restraint is an energy function parameterized to represent a conformational energy of a protein, driving the refinement toward a low-energy conformation but allowing reasonable deviations from it. Strictly, however, protein-structure refinement involves fitting model parameters to data measured from the ensemble of structures in the crystal; the model parameters

should reflect ensemble properties rather than conformational energies of a single protein. This distinction may be subtle but it is statistically significant, particularly since the models themselves are simplified. For example, if the protein structure to be refined is modelled as a static structure with harmonic vibrations, then systematic errors could be expected if bond vibrations include significant anharmonicity. The systematically shortened C—H bonds of high-resolution X-ray crystal structures relative to those from neutron diffraction studies also show the systematic error involving the strictly erroneous approximation that scattering functions are spherically symmetric and centred at nuclei. These examples illustrate parameter problems that can in principle be minimized by correcting the refinement model, such as with the application of force-field restraints to an ensemble of structures, the addition of a statistically derived shift to proton positions, anisotropic scattering functions *etc.*

The 'stiffness' of the restraints should reflect the degree of confidence in the restraints. Here it is possible to distinguish different kinds of confidence, depending on the type of restraint used. The use of a conformational energy function as a refinement target function reflects the expectation that the magnitude of true deviations from the energy minima are related to the slopes of the physical energy potential surface; 'confidence' for bond lengths can thus be estimated from their vibration constants. An alternative approach is to derive target values from a database of independently determined structures. Distributions of parameter values derived from such databases have, in principle, arbitrary functional forms; a series expansion delivers the common descriptors of mean, variance, skew and so forth. The relatedness of the database to the system to be restrained determines how closely the distribution in the database should reflect the distribution in the restrained system. For proteins, it has seemed reasonable to expect that bond and angle parameters should share similar mean and sample variance distributions.

18.3.2.1. *Choice of properties for restraint*

Standard refinement procedures usually include the use of harmonic restraints of bond lengths, bond angles, planarity and 'improper' dihedrals, which, along with nuclear repulsions, comprise the 'hardest' restraints. Other geometric properties, such as dihedral angles, electrostatic interactions, Ramachandran energies *etc.*, can contribute further information to the refinement; these softer restraints can either distort the structure if weighted too strongly, or not contribute significantly to the refinement if weighted too weakly. Further, the softer restraints may be more useful as statistical parameters for judging the quality of a structure than as refinement parameters themselves. Particular care should be taken when refining with significant electrostatic energies: the forces have significant effects over long ranges, are usually based on simplified polarizability models and usually arise from assumptions of standard intrinsic pK_a values.

18.3.2.2. *Simple derivation of force constants from parameter distributions*

For a given distribution of parameter values, we take the average as the target value. If the choice of fragment geometries gives a distribution that corresponds to the varieties in the refinement problems, restraints should be applied to allow a similar range of freedom. In general, Gaussian distributions are assumed (with some obvious checks to avoid bimodal distributions) and are not corrected for non-orthogonality when parameters are introduced. If the refinement program uses a least-squares minimization method, the statistical mean and variance values can be used as tabulated. If an energy-function target is used, the variance values may be converted to force constants k defined by $E = k(\mathrm{d}x)^2$ by equating the Boltzmann probability, $\exp(-\mathrm{d}E/bT) = \exp[-k(\mathrm{d}x)^2/bT]$, with the Gaussian distribution function, $\exp[-(\mathrm{d}x)^2/(2\sigma^2)]$, such that $k = bT/2\sigma^2$, where b is the Boltzmann constant. (Note that this force constant k is one-half the magnitude of the force constant k' defined by Newton's law, $F = \mathrm{d}E/\mathrm{d}x = -k'x$.) This treatment resembles the generation of a mean field potential from the structure (*e.g.* Sippl, 1995), but is simplified for bond and angle parameters in several respects. Firstly, a single temperature is assumed; temperature dependence of equilibrium constants would require consideration of the temperatures of individual crystal structure determinations. Secondly, a normal-mode analysis would be required to eliminate redundancies arising from coupling between the parameters. Finally, the values of the variances are in many cases uncertain, due to poor statistical representation, non-Gaussian effects, or other causes (see below). Note that this simplified treatment of parameter uncertainty can be implemented more rigorously in maximum-entropy refinement methods (Pannu & Read, 1996; Bricogne, 1993).

18.3.2.2.1. *Clustering*

With ample computational power available for most refinement applications, the clustering of parameters into atom types might seem unnecessary. However, clustering is inevitable and occurs at least in the choice of fragments for deriving statistics. Values derived for individual residues, for example, effectively cluster three-dimensional structures together with a concomitant loss of information. For at least some residues with limited representation in the CSD, peptide geometries are best analysed as a general class, requiring clustering fragments into side-chain and main-chain classifications. Some side chains, such as arginine, also require smaller fragment definitions due to relatively small numbers in the CSD. The statistics listed here can be further clustered into fewer atom or bond-type classes. Inappropriate clustering, that is, the simultaneous analysis of fragments that are better represented by two or more fragment types with correspondingly various average values and sample variances, will exaggerate non-Gaussian distribution characteristics. In extreme cases, skew, kurtosis and especially multimodal distributions then provide evidence for a requirement to subdivide the fragment classes (see for example *cis*/*trans*-proline below). The effective clustering of the data presented here into largely conformation-independent fragment definitions could be replaced by more specific information, especially as the database grows.

18.3.2.2.2. *Treatment of outliers*

A truly Gaussian distribution should include outliers at high σ values (about 0.01% for 4σ). We should expect, however, that the width of the distribution is affected not only by inherent variation in the variables to be parameterized, but also by variability in the experimental conditions (*e.g.* resolution) and by erroneous structures. This weakens a strategy of automatic rejection of outliers beyond a specific cutoff value. The possibility of visualizing the distributions with CSD software allows refinement of this rejection strategy, with, however, the introduction of considerable subjectivity in the criteria. For this work, a 4σ cutoff was generally considered a flag for erroneous outliers. However, broad and flat tails in the distribution were relatively frequent and often asymmetric. These deviations from Gaussian behaviour 'artificially' increased σ values. In these cases, the 4σ cutoff rule was not applied automatically, but was applied after examination and rejection of conspicuous outliers. From an algorithmic viewpoint, this was the additional use of skew and kurtosis (third and fourth moments of the distribution) for rejection criteria. In

most cases, uncertainty in rejection criteria affected the average values little, but could significantly alter standard deviations.

18.3.2.3. *Bonds and angles*

18.3.2.3.1. *Peptide parameters: proline, glycine, alanine and CB substitution*

Fragments representing five-atom lengths of the backbone currently provide adequate statistics for peptide compositions of varieties including glycine, proline and side chains branched at CB. Peptide cyclicity was generally allowed on the assumption that this does not introduce distortions greater than typical protein secondary-structure interactions. The results are presented in Table 18.3.2.3. With one exception, none of the values deviates from those of 1991 by more than one sample standard deviation. However, the very large σ values for the proline C—N—CA and C—N—CD angles (Table 18.3.2.1) are conspicuous. Using high-resolution protein structures, Lamzin *et al.* (1995) identified geometries of proline that were inconsistent with high-resolution protein structures and also noted inconsistencies in C—CA—CB angle parameters (see also the sections on individual amino acids below). In the case of proline, a bimodal distribution of these parameters could be resolved with the discrimination between *cis* and *trans* forms (Fig. 18.3.2.1). A scatter plot of the angles against ω torsion angle resolves the averages (and σ's) of 122.6 (50) and 125.4 (44)° for C—N—CA and C—N—CD, respectively, into *cis*- and *trans*-dependent values with much smaller sample deviations (see Table 18.3.2.2). The large σ value for CB—CG remains, however, particularly for *trans*-proline. Its origin is unknown, but proline pucker may play a role.

Glycine, with its unique CH_2 as CA, required new atom-type definitions for Engh & Huber (EH) (1991) parameterization to account for parameter-average differences of about one-half of a sample standard deviation. These also included C—N—CA, for which the average angles were 120.6° for glycine and 121.7° for the rest. The new statistics with 83 C—CO—NH—CH_2—C fragments estimate a larger value of 122.3° for the glycine C—N—CA angle.

'Extended atom'-type parameterizations, which cluster carbon atoms according to the number of bound hydrogen atoms, naturally separate parameters involving CB into values representing alanine and branched and unbranched side chains. Separate analyses of the bonds and angles for fragments depending on the number of hydrogen atoms at CB (1, 2 or 3) revealed significant variation for the C—CA—CB and N—CA—CB angles. The fragments chosen for peptide parameterization did not cover all possibilities for the peptide chain. In particular, effects of charges at the termini were not analysed. Also, specific residue sequences likely to have statistical effects, such as Pro-Pro (Bansal & Ananthanarayanan, 1988), were not analysed here. With 50–60 relevant fragments from the predominantly α-helical ROP protein, Vlassi *et al.* (1998) were able to compile statistics for main-chain bonds and angles and compare them with protein refinement parameters. Differences from EH were particularly significant for CO and CA—C bonds (1.237 and 1.508 Å, respectively) and for the O—C—N angle (121.35°). Excepting the proline O—C—N angle, for which the new CSD statistics predict an average value lowered to 121.1°, these values remained relatively unchanged. A likely source of the difference might be the predominantly helical structure of the ROP protein; the helical hydrogen bonding directly involves the C—O group in a systematic way.

18.3.2.3.2. *Aromatic residues: tryptophan, phenylalanine, tyrosine, histidine*

With the exception of generally lower σ values, tryptophan parameters remain essentially unchanged. Phenylalanine, also with generally lower σ values, is also essentially unchanged with the assumption of Gaussian distributions. However, a scatter plot of the CB—CG—CD1 *versus* CB—CG—CD2 angles shows an inverse correlation between these two angles, corresponding to ring rotations about an axis perpendicular to the ring face. Non-Gaussian distributions were most evident for tyrosine. In addition to the phenomenon described for phenylalanine, a clearly multimodal distribution was observed for the CE(1,2)—CZ—OH angles, with maxima at 118 and 122° (Fig. 18.3.2.2). The scatter plot of CE1—CZ—OH *versus* CE2—CZ—OH demonstrates that this distribution typifies individual fragments and does not arise from differing classes of fragments. This justifies an asymmetric parameterization for these angles; symmetric parameterization would require correspondingly soft force constants. The major difference between the histidine parameters listed here compared to those of EH arise from the appearance of HISD (uncharged; unprotonated at NE2) fragments in the CSD. The EH parameterization assumed values from other fragments. The total of 12 fragments is not large, but does predict some alterations in parameters involving the ring nitrogens. The fragment selection reported here did not investigate effects of noncovalent binding. For the aromatic residues, these include hydrogen-bonding effects (especially for histidine) and π-cloud interactions. Appropriate fragments exist in the database, so such dependencies are, in principle, accessible to investigation.

18.3.2.3.3. *Aliphatic residues: leucine, isoleucine, valine*

Compared to EH parameterization, the only notable features of the aliphatic residues were the leucine bonds and the C—CA—CB angles of isoleucine and valine. The leucine CD—CG(1,2) bonds retained relatively large σ values, which rather increased compared to the previous values. The C—CA—CB angle values, clustered as bare carbon/tetrahedral CH extended atom/tetrahedral CH_2 extended atom in EH, are sensitive to the degree of substitution at the CB carbon (Table 18.3.2.3, see the discussion of peptide fragments above). The statistics here show that the EH (1991) parameters were too small by about 2°.

18.3.2.3.4. *Neutral polar residues: serine, threonine, glutamine, asparagine*

These residues share neutral polarity, but are all geometrically distinct. Like leucine, valine and isoleucine described above, threonine is branched at CB, and the parameterization for C—CA—CB should be chosen accordingly. Additionally for threonine, the CA—CB—CG2 angle, clustered with valine as CH1E—CH1E—CH3E in EH (1991), should be altered from 110.5 to 112.4° according to the statistics reported here. The tabulated glutamine and asparagine parameters are taken from identical amide-group statistics, and parameters for the aliphatic atoms of glutamine are taken from arginine. This choice of fragments arose from a desire to maximize the number of fragments for the amide group; however, the individual residues might be expected to exhibit residue-specific amide structures.

18.3.2.3.5. *Acidic residues: glutamate, aspartate*

The fragment definitions were chosen to select both symmetrically and asymmetrically encoded carboxylate structures; that is, the statistics include carboxylate groups with delocalized charges as well as carboxylate groups encoded with a single charged oxygen atom. This distribution presumably reflects the variations in proteins as well. For both glutamic and aspartic acids, statistical variation in the asymmetry of delocalization was evident. One measure of parameter variation as a function of varying charge delocalization is the anticorrelation of C—O bond lengths and CH_2—C—O bond angles. For example, while the standard deviation

Table 18.3.2.1. *Bond lengths of standard amino-acid side chains*

EH denotes the values of Engh & Huber (1991), which were clustered according to atom type. The EH99 values are taken from recent Cambridge Structural Database releases with clustering of parameters only in the choice of fragments, based on amino acids. Parameters marked with an asterisk involving CA—CB bonds were taken from peptide fragment geometries. Two asterisks mark long-chain aliphatic parameters taken from arginine statistics. The number of fragments and the number of structures containing these fragments are noted after the amino-acid name. The fragments used for generating the statistics are described after the amino-acid name: incomplete valences indicate unspecified substituents with, however, specified orbital hybridization.

Alanine, 163/268, CO—NH—CH(CH_3)—CO—NH

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.521	0.033	1.520	0.021

Arginine, 71/98, CH—$(CH_2)_3$—NH—$C(NH_2)_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521	0.027
CG—CD	1.520	0.030	1.515	0.025
CD—NE	1.460	0.018	1.460	0.017
NE—CZ	1.329	0.014	1.326	0.013
CZ—NH(1,2)	1.326	0.018	1.326	0.013

Asparagine, 145/247, —C—CH_2—CO—NH_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.527	0.026
CB—CG	1.516	0.025	1.506	0.023
CG—OD1	1.231	0.020	1.235	0.022
CG—ND2	1.328	0.021	1.324	0.025

Aspartate, 265/404, C—CH_2—CO_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.516	0.025	1.513	0.021
CG—OD(1,2)	1.249	0.019	1.249	0.023

Cysteine, 10/17, N—CH(CO)—CH_2—SH

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.526	0.013
CB—SG	1.808	0.033	1.812	0.016

Disulfides, 53/68, C—CH_2—S—S—CH_2—C

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—SG	1.808	0.033	1.818	0.017
SG—SG	2.030	0.008	2.033	0.016

Glutamate, 74/88, C—CH_2—CH_2—CO_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.517	0.019
CG—CD	1.516	0.025	1.515	0.015
CD—OE(1,2)	1.249	0.019	1.252	0.011

Glutamine, 145/247, —C—CH_2—CO—NH_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521**	0.027**
CG—CD	1.516	0.025	1.506	0.023
CD—OE1	1.231	0.020	1.235	0.022
CD—NE2	1.328	0.021	1.324	0.025

Glycine: see peptide parameters, Table 18.3.2.3

Histidine (HISE), 35/37, C—CH_2—imidazole; NE protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.496	0.018
CG—ND1	1.371	0.017	1.383	0.022
CG—CD2	1.356	0.011	1.353	0.014
ND1—CE1	1.319	0.013	1.323	0.015
CD2—NE2	1.374	0.021	1.375	0.022
CE1—NE2	1.345	0.020	1.333	0.019

Histidine (HISD), 10/12, C—CH_2—imidazole; ND protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.492	0.016
CG—ND1	1.378	0.011	1.369	0.015
CG—CD2	1.356	0.011	1.353	0.017
ND1—CE1	1.345	0.020	1.343	0.025
CD2—NE2	1.382	0.030	1.415	0.021
CE1—NE2	1.319	0.013	1.322	0.023

Histidine (HISH), 50/54, C—CH_2—imidazole; NE, ND protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.492	0.010
CG—ND1	1.378	0.011	1.380	0.010
CG—CD2	1.354	0.011	1.354	0.009
ND1—CE1	1.321	0.010	1.326	0.010
CD2—NE2	1.374	0.011	1.373	0.011
CE1—NE2	1.321	0.010	1.317	0.011

Isoleucine, 54/80, NH—CH(CO)—CH(CH_3)—CH_2—CH_3

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.544	0.023
CB—CG1	1.530	0.020	1.536	0.028
CB—CG2	1.521	0.033	1.524	0.031
CG1—CD1	1.513	0.039	1.500	0.069*

Table 18.3.2.1. *Bond lengths of standard amino-acid side chains* (*cont.*)

Leucine, 178/288, NH—CH(CO)—CH_2—$CH(CH_3)_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.533	0.023
CB—CG	1.530	0.020	1.521	0.029
CG—CD(1,2)	1.521	0.033	1.514	0.037

Lysine, 232/380, —$(CH_2)_3$—NH_3

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521**	0.027**
CG—CD	1.520	0.030	1.520	0.034
CD—CE	1.520	0.030	1.508	0.025
CE—NZ	1.489	0.030	1.486	0.025

Methionine, 37/49, C—$(CH_2)_2$—S—CH_3

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.509	0.032
CG—SD	1.803	0.034	1.807	0.026
SD—CE	1.791	0.059	1.774	0.056*

Phenylalanine, 1076/1616, C—CH_2—phenyl

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.502	0.023	1.509	0.017
CG—CD(1,2)	1.384	0.021	1.383	0.015
CD(1,2)—CE(1,2)	1.382	0.030	1.388	0.020
CE(1,2)—CZ	1.382	0.030	1.369	0.019

Proline, 262/255, *trans*, C—CO—pyrrolidine—CO—N

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.531	0.020
CB—CG	1.492	0.050	1.495	0.050
CG—CD	1.503	0.034	1.502	0.033
CD—N	1.473	0.014	1.474	0.014

Proline, 262/158, *cis*, C—CO—pyrrolidine—CO—N

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.533	0.018
CB—CG	1.492	0.050	1.506	0.039
CG—CD	1.503	0.034	1.512	0.027
CD—N	1.473	0.014	1.474	0.014

Serine, 33/39, NH—CH(CO)—CH_2—OH

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.525	0.015
CB—OG	1.417	0.020	1.418	0.013

Threonine, 20/25, NH—CH(CO)—CH(OH)—CH_3

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.529	0.026
CB—OG1	1.433	0.016	1.428	0.020
CB—CG2	1.521	0.033	1.519	0.033

Tryptophan, 123/135, CH_2—indole

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.498	0.031	1.498	0.018
CG—CD1	1.365	0.025	1.363	0.014
CG—CD2	1.433	0.018	1.432	0.017
CD1—NE1	1.374	0.021	1.375	0.017
NE1—CE2	1.370	0.011	1.371	0.013
CD2—CE2	1.409	0.017	1.409	0.012
CD2—CE3	1.398	0.016	1.399	0.015
CE2—CZ2	1.394	0.021	1.393	0.017
CE3—CZ3	1.382	0.030	1.380	0.017
CZ2—CH2	1.368	0.019	1.369	0.019
CZ3—CH2	1.400	0.025	1.396	0.016

Tyrosine, 124/161, *para*-(—C—CH_2)—phenol

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.512	0.022	1.512	0.015
CG—CD(1,2)	1.389	0.021	1.387	0.013
CD(1,2)—CE(1,2)	1.382	0.030	1.389	0.015
CE(1,2)—CZ	1.378	0.024	1.381	0.013
CZ—OH	1.376	0.021	1.374	0.017

Valine, 198/313, N—CH(CO)—CH—$(CH_3)_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.543	0.021
CB—CG(1,2)	1.521	0.033	1.524	0.021

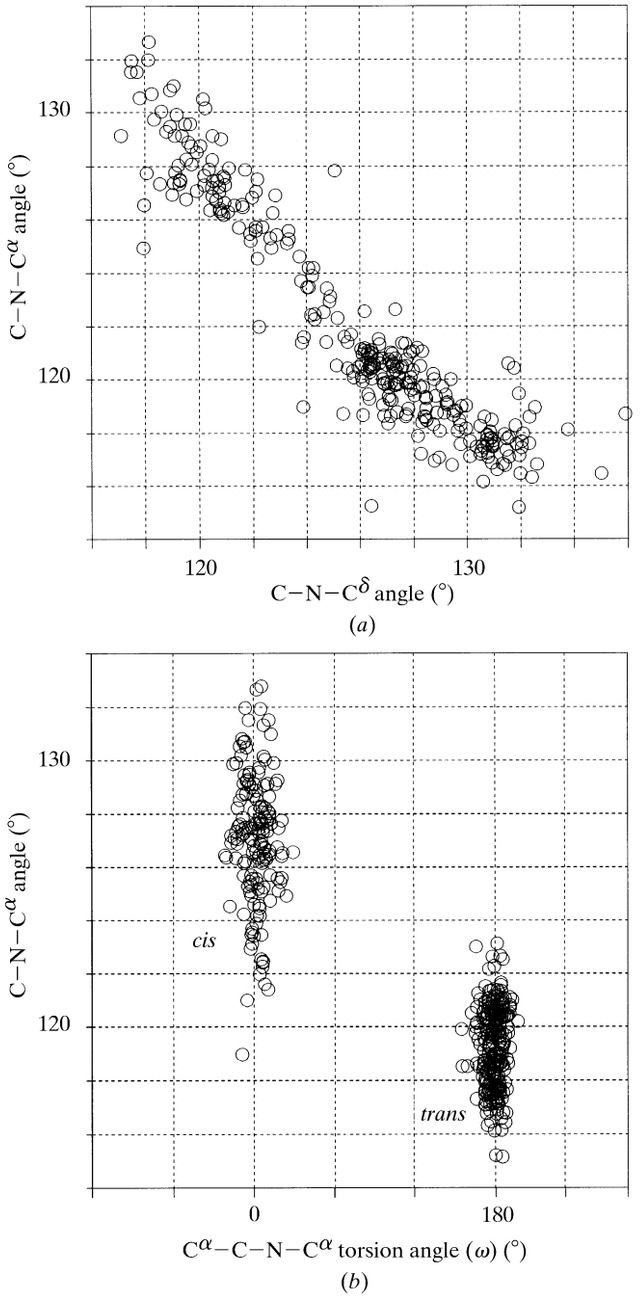

Fig. 18.3.2.1. Torsion dependence of proline angle geometry. A one-dimensional frequency plot of either C—N—C^{α} or C—N—C^{δ} angles shows a broad and bimodal distribution. (*a*) A scatter plot of the two angles shows a very strong anticorrelation and suggests two minima. (*b*) Plotting either angle against the ω torsion angle resolves the broad distribution into two separate peaks.

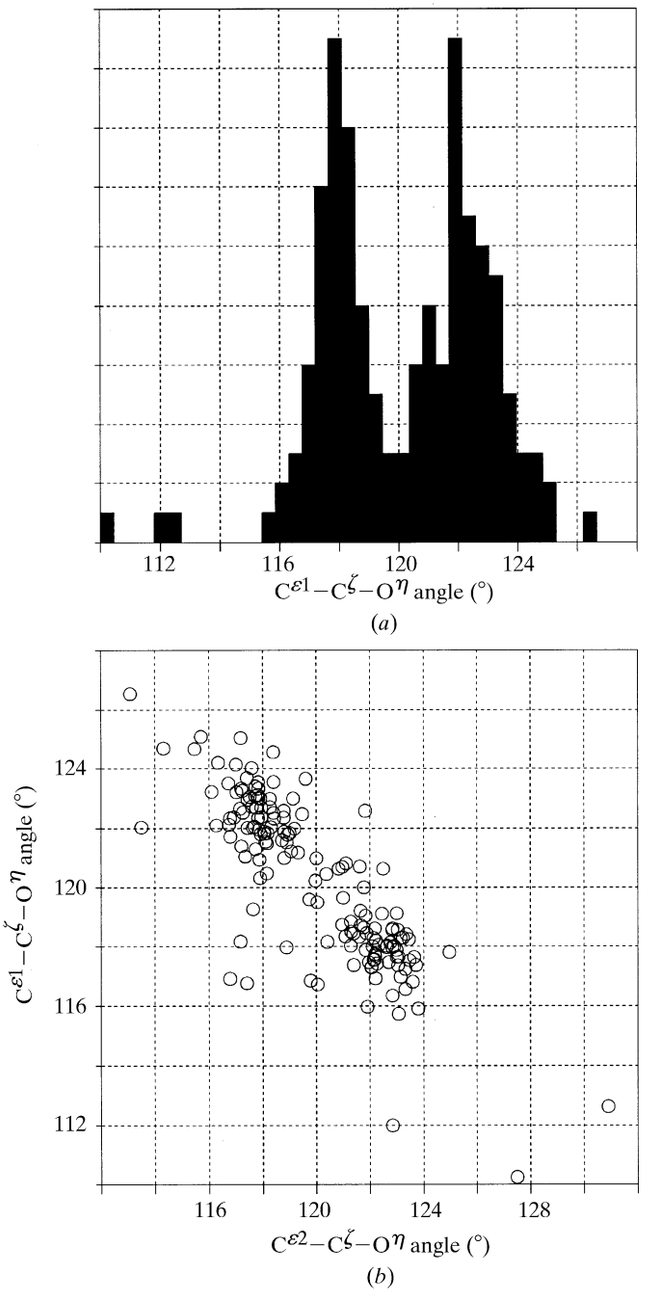

Fig. 18.3.2.2. Bimodal distributions for tyrosine. (*a*) The C^{ε}—C^{ζ}—O^{η} angle distributions involving the tyrosine alcohol have maxima at 120 (2)°. (*b*) A scatter plot of the $C^{\varepsilon 2}$—C^{ζ}—O^{η} angle against the $C^{\varepsilon 1}$—C^{ζ}—O^{η} angle confirms that the C^{ζ}—O^{η} bond projects asymmetrically away from the aromatic ring.

of the corresponding aspartate bond lengths individually is 0.024 Å, the standard deviation of their pairwise average is 0.012 Å. Similarly, the standard deviation of the glutamate CH_2—C—O bond angles individually is 2.1° but the standard deviation of the pairwise average is 0.6°. This coupling of parameters is an example of additional information potentially available for structure refinement, but which would require new formulations of restraints.

18.3.2.3.6. *Basic residues: arginine, lysine*

The 98 arginine fragments in the database did not show alterations from the EH values, except generally tighter restraints at the guanidinium group. Lysine CD—CE bond lengths are somewhat shorter in the new statistics, while the two angles derived from the fragments remained similar.

18.3.2.3.7. *Sulfur-containing residues: methionine, cysteine, disulfides*

One of the most conspicuous features of the EH parameters is the soft force constant for the methionine SD—CE bond length. The 49 fragments now in the CSD also show a sample deviation for the 1.774 Å average bond length of 0.056 Å, and after one 4σ outlier rejection, the tabulated value of 1.779 Å still has a large sample

Table 18.3.2.2. *Bond angles of standard amino-acid side chains*

For details see Table 18.3.2.1.

Alanine, 163/268, CO—NH—$CH(CH_3)$—CO—NH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.4	1.5	110.1	1.4
CB—CA—C	110.5	1.5	110.1	1.5

Arginine, 71/98, CH—$(CH_2)_3$—NH—$C(NH_2)_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4	2.2
CB—CG—CD	111.3	2.3	111.6	2.6
CG—CD—NE	112.0	2.2	111.8	2.1
CD—NE—CZ	124.2	1.5	123.6	1.4
NE—CZ—NH(1,2)	120.0	1.9	120.3	0.5
NH1—CZ—NH2	119.7	1.8	119.4	1.1

Asparagine, 145/247, —C—CH_2—CO—NH_2

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	112.6	1.0	113.4**	2.2**
CB—CG—ND2	116.4	1.5	116.7	2.4
CB—CG—OD1	120.8	2.0	121.6	2.0
ND2—CG—OD1	122.6	1.0	121.9	2.3

Aspartate, 265/404, C—CO_2

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	112.6	1.0	113.4*	2.2**
CB—CG—OD(1,2)	118.4	2.3	118.3	0.9
OD1—CG—OD2	122.9	2.4	123.3	1.9

Cysteine, 10/17, N—CH(CO)—CH_2—SH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.8	1.5
CB—CA—C	110.1	1.9	111.5	1.2
CA—CB—SG	114.4	2.3	114.2	1.1

Disulfides, 53/68, C—CH_2—S—S—CH_2—C

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—SG	114.4	2.3	114.0	1.8
CB—SG—SG	103.8	1.8	104.3	2.3

Glutamate, 74/88, C—CH_2—CH_2—CO_2

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4**	2.2**
CB—CG—CD	112.6	1.7	114.2	2.7
CG—CD—OE(1,2)	118.4	2.3	118.3	2.0
OE1—CD—OE2	122.9	2.4	123.3	1.2

Glutamine, 145/247, —C—CH_2—CO—NH_2

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4**	2.2**
CB—CG—CD	112.6	1.7	111.6**	2.6**
CG—CD—OE1	120.8	2.0	121.6	2.0
CG—CD—NE2	116.4	1.5	116.7	2.4
OE1—CD—NE2	122.6	1.0	121.9	2.3

Glycine: see Table 18.3.2.3

Histidine (HISE), 35/37, C—CH_2— imidazole; NE protonated

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.6	1.7
CB—CG—ND1	121.6	1.5	121.4	1.3
CB—CG—CD2	129.1	1.3	129.7	1.6
CG—ND1—CE1	105.6	1.0	105.7	1.3
ND1—CE1—NE2	111.7	1.3	111.5	1.3
CE1—NE2—CD2	106.9	1.3	107.1	1.1
NE2—CD2—CG	106.5	1.0	106.7	1.2
CD2—CG—ND1	109.2	0.7	108.8	1.4

Histidine (HISD), 10/12, C—CH_2— imidazole; ND protonated

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.6	1.7
CB—CG—ND1	122.7	1.5	123.2	2.5
CB—CG—CD2	129.1	1.3	130.8	3.1
CG—ND1—CE1	109.0	1.7	108.2	1.4
ND1—CE1—NE2	111.7	1.3	109.9	2.2
CE1—NE2—CD2	107.0	3.0	106.6	2.5
NE2—CD2—CG	109.5	2.3	109.2	1.9
CD2—CG—ND1	105.2	1.0	106.0	1.4

Table 18.3.2.2. *Bond angles of standard amino-acid side chains* (*cont.*)

Histidine (HISH), 50/54, C—CH_2—imidazole; NE, ND protonated

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.6	1.6
CB—CG—ND1	122.7	1.5	122.5	1.3
CB—CG—CD2	131.2	1.3	131.4	1.2
CG—ND1—CE1	109.3	1.7	109.0	1.0
ND1—CE1—NE2	108.4	1.0	108.5	1.1
CE1—NE2—CD2	109.0	1.0	109.0	0.7
NE2—CD2—CG	107.2	1.0	107.3	0.7
CD2—CG—ND1	106.1	1.0	106.1	0.8

Isoleucine, 54/80, NH—CH(CO)—$CH(CH_3)$—CH_2—CH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	111.5	1.7	110.8	2.3
CB—CA—C	109.1	2.2	111.6	2.0
CA—CB—CG1	110.4	1.7	111.0	1.9
CB—CG1—CD1	113.8	2.1	113.9	2.8
CA—CB—CG2	110.5	1.7	110.9	2.0
CG1—CB—CG2	110.7	3.0	111.4	2.2

Leucine, 178/288, NH—CH(CO)—CH_2—$CH(CH_3)_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.4	2.0
CB—CA—C	110.1	1.9	110.2	1.9
CA—CB—CG	116.3	3.5	115.3	2.3
CB—CG—CD(1,2)	110.7	3.0	111.0	1.7
CD1—CG—CD2	110.8	2.2	110.5	3.0

Lysine, 232/380, —$(CH_2)_3$—NH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4*	2.2**
CB—CG—CD	111.3	2.3	111.6**	2.6**
CG—CD—CE	111.3	2.3	111.9	3.0
CD—CE—NZ	111.9	3.2	111.7	2.3

Methionine, 37/49, C—$(CH_2)_2$—S—CH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.3	1.7
CB—CG—SD	112.7	3.0	112.4	3.0
CG—SD—CE	100.9	2.2	100.2	1.6

Phenylalanine, 1076/1616, C—CH_2—phenyl

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.9	2.4
CB—CG—CD(1,2)	120.7	1.7	120.8	0.7
CD(1,2)—CG—CD(2,1)	118.6	1.5	118.3	1.3
CG—CD(1,2)—CE(1,2)	120.7	1.7	120.8	1.1
CD(1,2)—CE(1,2)—CZ	120.0	1.8	120.1	1.2
CE(1,2)—CZ—CE(2,1)	120.0	1.8	120.0	1.8

Proline, 262/255, *trans*, C—CO—pyrrolidine—CO—N

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	103.0	1.1	103.3	1.2
CB—CA—C	110.1	1.9	111.7	2.1
CA—CB—CG	104.5	1.9	104.8	1.9
CB—CG—CD	106.1	3.2	106.5	3.9
CG—CD—N	103.2	1.5	103.2	1.5
CA—N—CD	112.0	1.4	111.7	1.4
C—N—CA	122.6	5.0	119.3	1.5
C—N—CD	125.0	4.1	128.4	2.1

Proline, 262/158, *cis*, C—CO—pyrrolidine—CO—N

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	103.0	1.1	102.6	1.1
CB—CA—C	110.1	1.9	112.0	2.5
CA—CB—CG	104.5	1.9	104.0	1.9
CB—CG—CD	106.1	3.2	105.4	2.3
CG—CD—N	103.2	1.5	103.8	1.2
CA—N—CD	112.0	1.4	111.5	1.4
C—N—CA	122.6	5.0	127.0	2.4
C—N—CD	125.0	4.1	120.6	2.2

Serine, 33/39, NH—CH(CO)—CH_2—OH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.5	1.5
CB—CA—C	110.1	1.9	110.1	1.9
CA—CB—OG	111.1	2.0	111.2	2.7

Threonine, 20/25, NH—CH(CO)—CH(OH)—CH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	111.5	1.7	110.3	1.9
CB—CA—C	109.1	2.2	111.6	2.7
CA—CB—OG1	109.6	1.5	109.0	2.1
CA—CB—CG2	110.5	1.7	112.4	1.4
OG1—CB—CG2	109.3	2.0	110.0	2.3

Table 18.3.2.2. *Bond angles of standard amino-acid side chains* (*cont.*)

Tryptophan, 123/135, CH_2—indole

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.6	1.9	113.7†	1.9†
CB—CG—CD1	126.9	1.5	127.0	1.3
CB—CG—CD2	126.8	1.4	126.6	1.3
CD1—CG—CD2	106.3	1.6	106.3	0.8
CG—CD1—NE1	110.2	1.3	110.1	1.0
CD1—NE1—CE2	108.9	1.8	109.0	0.9
NE1—CE2—CD2	107.4	1.3	107.3	1.0
CE2—CD2—CG	107.2	1.2	107.3	0.8
CG—CD2—CE3	133.9	1.0	133.9	0.9
NE1—CE2—CZ2	130.1	1.5	130.4	1.1
CE3—CD2—CE2	118.8	1.0	118.7	1.2
CD2—CE2—CZ2	122.4	1.0	122.3	1.2
CE2—CZ2—CH2	117.5	1.3	117.4	1.0
CZ2—CH2—CZ3	121.5	1.3	121.6	1.2
CH2—CZ3—CE3	121.1	1.3	121.2	1.1
CZ3—CE3—CD2	118.6	1.3	118.8	1.3

Tyrosine 124/161, *para*—C—CH_2—phenyl—OH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.9	1.8	113.4	1.9
CB—CG—CD(1,2)	120.8	1.5	121.0	0.6
CD(1,2)—CG—CD(2,1)	118.1	1.5	117.9	1.1
CG—CD(1,2)—CE(1,2)	121.2	1.5	121.3	0.8
CD(1,2)—CE(1,2)—CZ	119.6	1.8	119.8	0.9
CE(1,2)—CZ—CE(2,1)	120.3	2.0	119.8	1.6
CE(1,2)—CZ—OH	119.9	3.0	120.1	2.7‡

Valine, 198/313, N—CH(CO)—CH—$(CH_3)_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	111.5	1.7	111.5	2.2
CB—CA—C	109.1	2.2	111.4	1.9
CA—CB—CG(1,2)	110.5	1.7	110.9	1.5
CG1—CB—CG2	110.8	2.2	110.9	1.6

† Alternate fragment definition including CA.
‡ Bimodal distribution (see text).

deviation of 0.041 Å. In practice, the use of soft restraints during refinement often leads to warnings of relatively large deviations from the target value. Inspection of the CSD structures did not reveal an artificial source of this greater variability. Cysteines and disulfides here show reduced sample σ values for generally similar average target values.

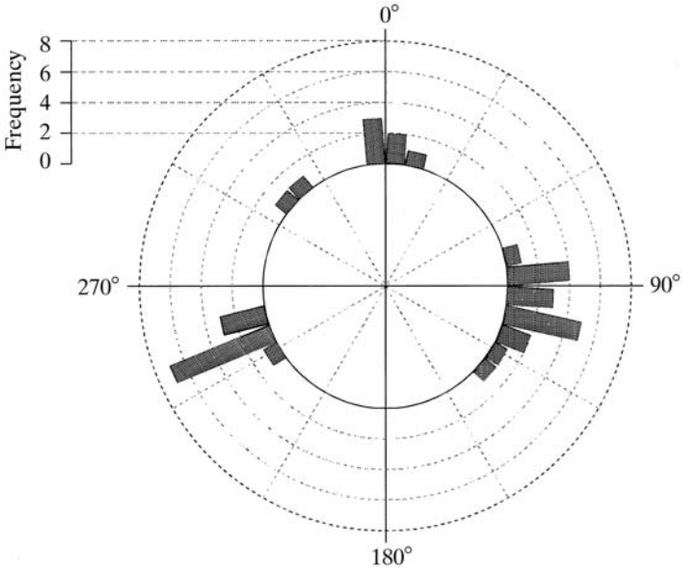

Fig. 18.3.2.3. Tryptophan χ_2 dihedral angle distribution. The 36 tryptophan fragments in the CSD show several apparent minima, including an eclipsed C^{α}—C^{β}—C^{γ}—$C^{\delta 1}$ dihedral conformation. This is apparent in χ_1, χ_2 tryptophan distribution plots from protein structures as well (Laskowski, MacArthur *et al.*, 1993).

18.3.2.4. *Planarity restraints*

Planarity and improper dihedral restraints, being 'hard' restraints, are amenable to the same kind of parameterization described above. Physically realistic deviations should be allowed. A survey of several planar atoms, such as CG of aromatic residues, the inter-ring carbons (CD2, CE2) of tryptophan and CZ of tyrosine, showed standard deviations about strict planarity of 1–2°. Statistically significant deviations of average values from perfect planarity might also be expected, particularly as a function of the protein fold environment. For example, an average nonzero planarity of the ω angle of the peptide bond has been noted (Marquart *et al.*, 1983) and attributed at least in part to secondary structure; such effects may be misrepresented in statistics from small molecules.

18.3.2.5. *Torsion angles*

Since torsion angles are generally more adequately determined by protein structures of typical resolutions, there is less need to derive parameters from the CSD for refinement purposes. Further, distributions derived from small molecules may not be representative of torsion angles among proteins, since typical fragments lack 'typical' protein secondary structure. This kind of environmental dependence will affect softer parameters such as dihedral angles more than bonds or angles. However, since the torsion-angle distribution is a function not only of potential interactions of the peripheral groups but of the electronic character of the bond itself, some questions may be best examined by the study of chemical fragments. One such feature might be the $\chi_1\chi_2$ distributions for aromatic residues. Statistical distributions of side-chain orientations show the apparent stability of an eclipsed χ_2 conformation, particularly for $\chi_1 = 300°$ (g^- conformer, Fig. 18.3.2.3). If the relative dearth of secondary structure leads to atypical torsion-angle distributions in proteins, it must be conversely expected that torsion-angle statistics derived from the set of all proteins will be

Table 18.3.2.3. *Bond lengths (Å) and angles (°) of peptide backbone fragments*

EH denotes parameters from Engh & Huber (1991). Bold values mark important updates for angles involving proline (with *cis* and *trans* distinction) and branched CB atoms (isoleucine, valine, threonine). The number of fragments used for the statistics is given. The standard deviation of each value is given in parentheses following the value and is on the scale of the least significant digit of the value.

(*a*) Bonds

	Peptide (1165)	Proline (297)	Glycine (83)	EH	EH proline	EH glycine
N—CA	1.459 (20)	1.468 (17)	1.456 (15)	1.458 (19)	1.466 (15)	1.451 (16)
CA—C	1.525 (26)	1.524 (20)	1.514 (16)	1.525 (21)	EH	1.516 (18)
C—O	1.229 (19)	1.228 (20)	1.232 (16)	1.231 (20)	EH	EH
CA—CB (all)	1.532 (31)	1.531 (20)	—	1.530 (20)	EH	—
CA—CB (CH_3)	1.521 (33)	—	—	1.521 (33)	—	—
CA—CB (CH_2)	1.535 (22)	1.531 (20)	—	1.530 (20)	EH	—
CA—CB (CH)	1.542 (23)	—	—	1.540 (27)	—	—
C—N	1.336 (23)	1.338 (19)	1.326 (18)	1.329 (14)	1.341 (16)	EH

(*b*) Angles

	Peptide (1165)	Proline (297)	Glycine (83)	EH	EH proline	EH glycine
N—CA—C	111.0 (27)	112.1 (26)	113.1 (25)	111.2 (28)	111.8 (25)	112.5 (29)
N—CA—CB (all)	110.6 (21)	103.1 (12)	—	110.5 (17)	103.0 (11)	—
N—CA—CB (CH_3)	110.4 (15)	—	—	110.4 (15)	EH	—
N—CA—CB (CH_2)	110.6 (18)	103.1 (12)	—	110.5 (17)	103.0 (11)	—
N—CA—CB (CH)	111.1 (23)	—	—	111.5 (17)	EH	—
CA—C—N	117.2 (22)	117.1 (28)	116.2 (20)	116.2 (20)	116.9 (15)	116.4 (21)
CA—C—O	120.1 (21)	120.2 (24)	120.6 (18)	120.8 (17)	EH	120.8 (21)
O—C—N	122.7 (16)	121.1 (19)	123.2 (17)	123.0 (16)	122.0 (14)	EH
C—CA—CB	110.6 (23)	111.8 (20)	—	110.1 (19)	EH	—
C—CA—CB (CH_3)	110.5 (15)	—	—	110.5 (15)	—	—
C—CA—CB (CH_2)	110.4 (20)	111.8 (20)	—	110.1 (19)	EH	—
C—CA—CB (CH)	**111.3 (20)**	—	—	109.1 (22)	—	—
C—N—CA	121.7 (25)	122.0 (42) all **119.3 (15)** ***trans*** **127.0 (24)** ***cis***	**122.3 (21)**	121.7 (18)	122.6 (50)	120.6 (17)

less applicable, for example, to helical proteins than statistics derived from mostly helical proteins. This illustrates the dilemma of selecting the ideal statistical database for protein refinement.

18.3.2.6. *Non-bonded interactions*

Like the potential parameterization of torsion-angle statistics with the CSD, parameterization of non-bonded interactions, typically into terms representing packing (an empirical mix of London dispersion forces and solvent effects), electrostatics and hydrogen bonding, is probably more strongly influenced by protein environment than are bond and angle terms. Specific chemical questions are likely to be best addressed with fragment structure databases, however, for improved parameterization or structure-quality evaluation. Possible examples include hydrogen bonds, salt bridges (see, *e.g.*, the discussion on charge delocalization of Glu and Asp above) *etc.* These are particularly relevant for structures generally atypical among proteins, such as at enzyme reactive sites.

18.3.2.7. *Effects of hydrogen atoms in parameterization*

While many CSD fragments include hydrogen-atom positions, their accuracy is necessarily the most limited. Evaluation of CSD statistics without hydrogens and the subsequent addition of parameters to refine hydrogens adds an additional artifactual coupling between parameters involving non-hydrogen atoms as well. This artifactual coupling might theoretically be of some concern, but is a second-order effect and presumably introduces effects smaller than the artifactual coupling between non-hydrogen parameters, for example.

18.3.2.8. *Special geometries: cofactors, ligands, metals etc.*

Most crystallographers will experience neither the need nor desire to derive their own parameterization for general protein structure refinement; many, however, need new parameters for ligands or other entities that are not amino-acid residues. The accuracy required for their application will determine the appropriate effort. If the purpose of the structure is to determine the general orientation of an inhibitor for molecular modelling studies with a still lower effective accuracy, it may be that no refinement (and no parameterization) is necessary at all. As the accuracy requirements increase, so does the need for good parameterization and a way of estimating when the density is incompatible with known structures. Such incompatibility may be decisive in identifying the stereochemistry of ligands selected from racemic mixtures, or the occurrence of a chemical reaction, or even falsely characterized substances. For such applications, small-molecule structural databases will remain the only choice for parameter derivation, which can be done exactly as for amino-acid fragments.

18.3.2.9. *Addition of tailored information sources*

When specific structural effects are observed which are not otherwise parameterized, new parameters might be desirable to encode this information. In the case of new statistical minima or variances of parameters already encoded by the refinement program, it is for most programs a simple matter to introduce new atom types, residues, or fragment names to the topology and parameter libraries. Examples include new parameterization for charged states or for *cis*- and *trans*-proline. If the reason for the new statistics is structural, the structure must be appropriately monitored during refinement to ensure that the conditions continue to hold, particularly in the case of simulated-annealing refinement steps.

18.3.3. Strategy of application during building/refinement

Refinement parameters necessarily and intentionally introduce bias into the refinement which may not disappear with later alterations of parameters. The importance of this fact is reflected by the observation that the parameters of refined structures can be recognized by statistical studies of the structures (Laskowski, Moss & Thornton, 1993). It is therefore important that the parameters initially reflect what can be confidently predicted about the structure. If unknown geometries may be expected, at metal or catalytic sites, for example, or if isomerization states need to be recognized from the refined structure, all relevant parameters must be initially eliminated from the refinement. Depending on the resolution of the structure and the detail required, the unbiased final refined structure may sufficiently demonstrate the unknown structural quantities. On the other hand, insufficient restraint may allow unreasonable geometries that do not allow recognition of the desired quantity. In this case, it may be necessary to test all possible restraint conditions and compare the results of the refinements.

18.3.3.1. *Confidence in restraints versus information from diffraction*

Primarily in cases of new structures, such as small-ligand- or metal-binding proteins, the refinement may indicate that the expected geometries and applied restraints seem incompatible with evidence from the electron density. Several sources for such discrepancies must be considered for an evaluation of the true geometries or the confidence level of such an evaluation. The quality of the experimental information, such as data resolution and reduction parameters, must be considered. Physical phenomena possibly ignored by the refinement model might include anisotropies of motion and/or electron distribution, or disorder in the crystal. These might lead to systematic deviations in the refined structure that mimic alternate parameterizations. Finally, newly derived parameters should be examined to decide whether the fragments and chemical environments were inappropriate for the refinement problem, or whether errors in fragment structures artifactually distorted the parameterization.

18.3.4. Future perspectives

It seems obvious to seek the best (most accurate) possible parameterization and establish it as a standard (to enable statistical structure comparisons). This does not seem to be a realistic goal for several reasons. Firstly, the parameterization is less a determinant of accuracy than the quality of the data and the method of refinement. Secondly, the quality of existing parameterization and the potential for new environment-dependent parameters improves as more structures are solved and databases grow. Such new parameters can be derived from conformation-dependent statistics (*cis*- and *trans*-proline is an example described above), hydrogen-bonding geometries *etc.* Finally, protein structures are generally solved not to build a statistically optimized protein database, but to discover biophysical functional mechanisms.

The growth of structural databases will improve our understanding of structural properties (Wilson *et al.*, 1998); the highest-resolution protein structures will contribute most to the database, while low-resolution structures will profit most from improved predictive power. Structures that require restrained refinement both draw on the database for refinement parameters and integrity checks, and also contribute to it; a kind of boot-strapping procedure to re-refine deposited structures with iteratively improved parameters is conceivable (if convergent). The consequent removal of parameterization 'signatures' in, *e.g.*, bond and angle parameters seems unlikely to have practical consequences beyond identification of, *e.g.*, catalytically relevant outliers, but qualitative improvements in structure comparison might be revealing in unexpected ways. Such an effort will require adequate computational resources and the deposition of structure factors or, even better, diffraction images.

references