Celotna doktorska disertacija, vključno s prilogami, naj ne bi obsegala več kot 200 strani.

Vsako glavno poglavje se mora pričeti na lihi strani.

Poglavja si morajo slediti v zaporedju, predpisanem s predlogo.

V kolikor so v delo vključeni algoritmi in/ali slike in/ali tabele, so kazala zanje obvezna.

Pri navajanju virov in literature je potrebno upoštevati pravila, opisana v dokumentu Navajanje virov.doc.

Complete doctoral dissertation including apendices should not exceed 200 pages.

Each main chapter should begin on an odd page.

The chapters should follow as prescribed in the template.

When the work contains algorithms and/or figures and/or tables their indexes are obligatory.

The references should consider the rules as described in the document Citation style.doc.

KRATKA NAVODILA
SHORT INSTRUCTION

NE BRIŠI
DON'T DELETE

Miha Andrejašič

PURY: A TOPOLOGY AND GEOMETRY PARAMETER LIBRARY FOR SMALL MOLECULES

Doctoral Dissertation

PURY: KNJIŽNICA TOPOLOŠKIH IN GEOMETRIJSKIH PARAMETROV ZA MALE MOLEKULE

Doktorska disertacija

Supervisor: prof. dr. Dušan Turk

April 2009

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenia

Index

Abstract

The number and variety of macromolecular structures in complexes with "hetero" ligands is growing. The need for fast delivery of correct geometric parameters for their refinement, which is often crucial for understanding the biological relevance of the structure, is growing correspondingly. The current standard for describing protein structures is the Engh-Huber parameter set. It is an expert data set resulting from selection and analysis of crystal structures gathered in the Cambridge Structure Database (CSD). Clearly such a manual approach cannot be applied to the vast and ever growing number of chemical compounds. Therefore, a database, termed PURY, of geometric parameters of chemical compounds, together with a server accessing it, has been developed. PURY is a compilation of the whole CSD. It contains lists of atom classes and bonds connecting them, as well as angles, chirality, planarity and conformation parameters. The current compilation is based on CSD 5.28 and contains 1978 atom classes, 32702 bonding, 237068 angle, 201860 dihedral and 64193 improper geometric restraints. Analysis has confirmed that the restraints from the PURY database are suitable for use in macromolecular crystal structure refinement and should be of value for the crystallographic community. The database can be accessed through the web server "http://pury.ijs.si/", which from deposited co-ordinates creates topology and parameter files in forms for refinement programs MAIN, CNS and RefMac. The server will, in the near future, move to the CCDC web site.

Povzetek

Število in raznolikost makromolekulskih struktur v kompleksih s hetero ligandi raste, s tem pa tudi potreba po hitri pripravi pravilnih geometrijskih parametrov za izboljševanje (refinement) ujemanja le-teh, saj primerjava ujemanja pokaže razlike, pogosto odločilne za razumevanje biološke pomembnosti struktur. Sedanji standard za opis geometrije proteinskih struktur je set Engh-Huberjevih parametrov. Set je pripravljen s skrbno izbiro in analizo kristalnih struktur majhnih molekul, izbranih iz baze Cambridge Structural Database (CSD). Vendar je ročna izbira in analiza struktur nepraktična in je zato ne uporabljamo za pripravo seta, ki bi zagotavljal pravilne geometrijske parametre za naraščajoče število različnih struktur malih molekul. Baza PURY je bila razvita za avtomatizirano analizo, generiranje geometrijskih parametrov in distribucijo le-teh preko spletne strani. Pripravljena je z analizo celotne baze CSD. Vsebuje listo vseh razredov atomov, vezi, ki jih povezujejo, prav tako kotov, ravninskih (dihedralnih) kotov ter planarnih in konformacijskih izrazov. Trenutna baza parametrov temelji na verziji 5.28 baze CSD in vsebuje 1978 razredov atomov, 32702 vezi, 237068 kotov, 201860 ravninskih (dihedralnih) kotov ter 64193 izrazov za opis planarnosti oziroma kiralnosti. Analize so potrdile, da so parametri, ki jih vsebuje in ponuja baza PURY, primerne za izboljševanje struktur malih molekul in proteinov in da predstavljajo pomemben dodatek kristalografski skupnosti. Do baze lahko uporabniki dostopajo preko spletne strani “http://pury.ijs.si“, ki po vnosu koordinat molekule uporabniku ponudi datoteke z opisom topologije molekule in pripadajočimi parametri, pripravljene za izboljševanje s programi MAIN, CNS in RefMac. Spletni strežnik bo v bližnji prihodnosti prestavljen na strani CCDC.

Abbreviations

ASCII	=	American Standard Code for Information Interchange
Å	=	Ångström [10^{-10} m]
CCDC	=	Cambridge Crystallographic Data Centre
CGI	=	Common Gateway Interface
CNS	=	Crystallography & NMR System
CSD	=	Cambridge Structural Database
EH	=	Engh-Huber protein parameter set
HTML	=	HyperText Markup Language
MAD	=	Multiple wavelength anomalous dispersion
MIR	=	Multiple wavelength isomorphous scattering
NMR	=	Nuclear Magnetic Resonance
PDB	=	Protein Data Bank
RMSD	=	Root mean square deviation

1 Introduction

1.1 Historical insight

Macromolecules are the principal non-aqueous components of living cells. To understand cellular processes, knowledge of the three-dimensional structure of these macromolecules is vital. The word *protein* comes from the Greek word *prota* meaning "of primary importance". This name was introduced by Jons Jakob Berzelius in 1838 for large organic compounds with almost the same empirical formulas. He used the name because the organic compounds he was studying were primitive, but seemed to be very important for animal nutrition. Today it is known that all functions of living organisms are related to a large group of macromolecules called proteins. Each protein or a group of proteins has a specific function. That is why in bacterial cells, proteins represent about one half of the cell dry weight (Loewenstein and Cohen, 1964).

Proteins play crucial roles in nearly all biological processes - in catalysis, transport, coordinated motion and growth and differentiation control. This remarkable range of functions arises from their folding into many distinctive three-dimensional structures that bind highly diverse molecules. One of the major goals in biochemistry is to determine how amino acid sequence specifies the conformation of proteins and how proteins bind specific substrates and other molecules, mediate catalysis and transduce energy and information. It became obvious that the three-dimensional structure is the key determinant to understanding the function and mechanism of a protein. The understanding of protein structure and function has been greatly enriched by the x-ray crystallography, also known as macromolecular crystallography, a technique that can reveal the precise three-dimensional positions on most of the atoms in a protein molecule. (Stryer, 1996).

Two techniques are widely used for structural determination of macromolecules at atomic resolution: X-ray diffraction of crystals and nuclear magnetic resonance (NMR). While the mass limit for NMR technique is around 40,000 Da, the X-ray technique is favourable to many proteins and is hence today's method of choice. One of the first crucial steps in studying proteins was made by James B. Sumner in 1926 who showed enzymes could be isolated and crystallised. In 1955, Sir Frederick Sanger succeeded in determining the complete amino acid sequence of the protein insulin (Sanger and Tuppy, 1951). This was the first proof that all proteins have specific structures. In the mean time, protein crystallography was evolving. X-ray crystallography was initiated in 1912 by Max von Laue and Peter Ewald. Lawrence Bragg and his father William Bragg transformed the Laue equation into a physically intuitive form known as "Bragg's Law" (Bragg, 1913):

$$n \times \lambda = 2 \times d \times \sin \theta \qquad (1)$$

This equation led to structure determination of various inorganic salts and metals. In the 1930s, Bragg applied the Fourier methods into crystallography and thus introduced the central crystallographic problem: phasing. Simple organic compounds started to be examined in the 1920s (Rossmann, 2001).

The first breakthrough and begin of organic and protein crystallography was in 1956 by Dorothy Hodgkin and her colleagues with solved structures of penicillin and vitamin B-12 (Hodgkin et al., 1956). The structure of vitamin B-12 was solved using multiple isomorphous replacement and heavy atom refinement. Cobalt atoms were used for the purpose. This method is still used today to solve the phase problem. X-ray crystallography of biological molecules took off with Dorothy Crowfoot Hodgkin as she developed methods for indexing and processing X-ray intensities. The new methods were first successfully used by Max Perutz and Sir John Cowdery Kendrew in 1958 at solving the three-dimensional structures of hemoglobin and myoglobin (Kendrew et al., 1958, Muirhead and Perutz, 1963) for which they were awarded the Nobel Prize in Chemistry in 1962. In 1964, Hodgkin was awarded the Nobel Prize in Chemistry for solving the vitamin B-12 structure. In 1969, she succeeded in solving the insulin structure (Blundell et al., 1971), which took over thirty years to accomplish (Crowfoot Hodgkin, 1935).

Other techniques were invented for the solution of the phase problem including molecular replacement

and anomalous dispersion. Blow and Rossmann introduced the single isomorphous replacement method in 1961 (Blow and Rossmann, 1961) where only one heavy atom derivative is enough to solve the phasing problem. The method of multiwavelength anomalous dispersion (MAD) introduced later became the most popular with the use of synchrotrons (Hendrickson, 1991). MAD techniques greatly expanded by using proteins in which methionine residues are replaced by selenomethionine (Hendrickson et al. 1990, Yang et al. 1990) where selenium atoms are used as anomalous scatterers (Rossmann, 2001). Another advance was introduction of the molecular replacement technique (Hoppe 1957, Rossmann and Blow, 1961, Rossmann 1972). The idea behind it is that many proteins can crystallise in different forms but have common fold and common structures have common Patterson maps. It was only in the 1970s when more structures become available and when was possible to use the technique to solve homologous structures with suitable search models. Furthermore, computer graphics greatly facilitated electron density and atomic models representations (Jones 1978, O'Donnel and Olson 1981, Diamond 1982, Barry and McAlister 1982, Wright, 1982, Conolly and Olson 1985, Hubbard 1986, Ferrin et al. 1988). With introduction of computers into protein crystallography, refinement was introduced. Macromolecular models based on X-ray diffraction data, initially derived from electron density maps, were not very accurate because the phases used to calculate the electron densities were not very accurate. With the process of refinement, model parameters (coordinates, temperature factors) are adjusted in order to improve the agreement between intensities of the observed reflections and the values calculated from the model parameters. The accuracy of models is greatly improved; hence more information can be extracted from the model (Lyle, 1985, Hoppe and Gassmann 1968, Hughes 1941, Booth 1946a,b, Busing et al. 1962, Diamond 1971, Watenpaugh et al. 1972, Watenpaugh et al. 1973, Konnert 1976, Hendrickson 1985).

With the increasing number of macromolecular structures it became evident that a publicly available depository should be organized. In 1971 Protein Data Bank (PDB) was established in at Brookhaven National Laboratory, Upton, Long Island, New York (Bernstein et al., 1977). In 1972, two structures were deposited. By the First PDB Newsletter (1974), atomic coordinates were available for 12 proteins including carboxypeptidase A, alpha-chymotrypsin, cytochrome b5, lactate dehydrogenase, pancreatic trypsin inhibitor, subtilisin, myoglobin, rubredoxin, papain, and three hemoglobins. The increasing power of computers, computer graphics, synchrotron X-ray sources and recombinant techniques of molecular biology transformed protein crystallography into a vital area of science and a powerful tool for the pharmaceutical industry. Today PDB contains over 50 000 entries, many of them are complexes with ligands which are of natural origin or obtained by of chemical synthesis.

1.2 Principles of X-ray crystallography

Diffraction from a single molecule is too weak to be measurable. An ordered three-dimensional array of molecules, a crystal, is used instead to intensify the signal (**Figure 1**). Even a small protein crystal might contain 10^{15} molecules (Salemme et al. 1988, Jullien et al. 1994).

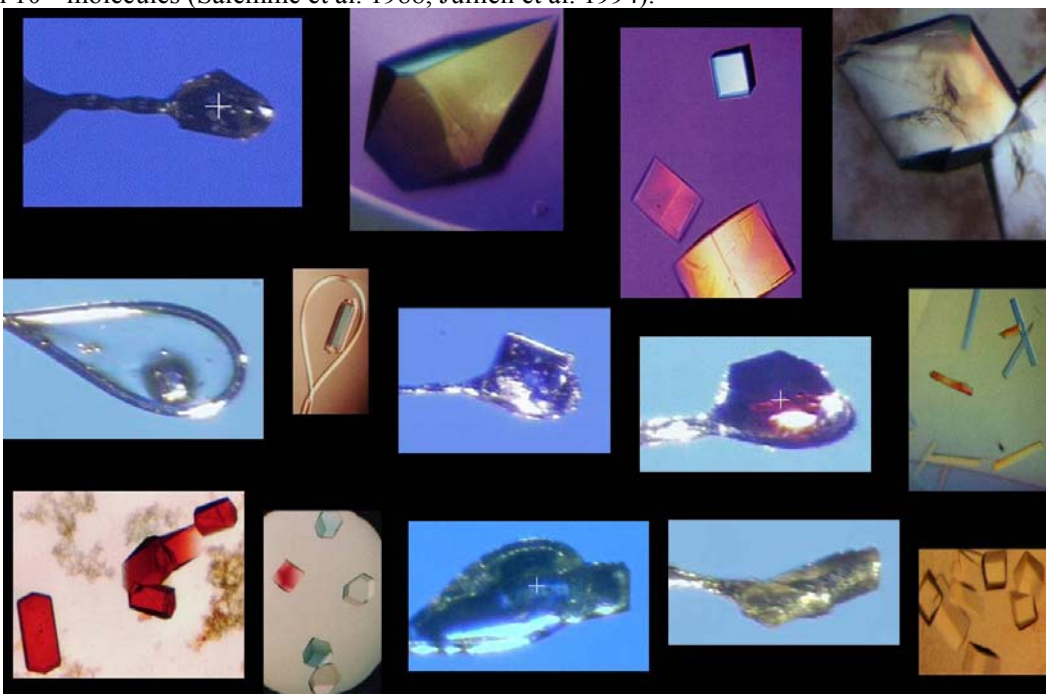


Figure 1: *Protein Crystals*

A crystal behaves like a three-dimensional diffraction grating which gives rise to both constructive and destructive interference effects in the diffraction pattern. When the internal order of the crystal is poor, X-rays will not be diffracted to high angles and the data will not result in a detailed high resolution structure. When the crystal is well ordered, diffraction is measurable at high angles and a detailed structure high resolution is obtained. On a diffraction image, the pattern appears as a series of discrete spots, known as reflections (**Figure 2**). The diffraction pattern is recorded using a detector: once an X-ray sensitive film, nowadays usually an image plate (Amemiya et al. 1988, Eikenberry et al. 1992) or a charge-coupled device (CCD) (Tate et al., 1995, Eikenberry et al, 1991).

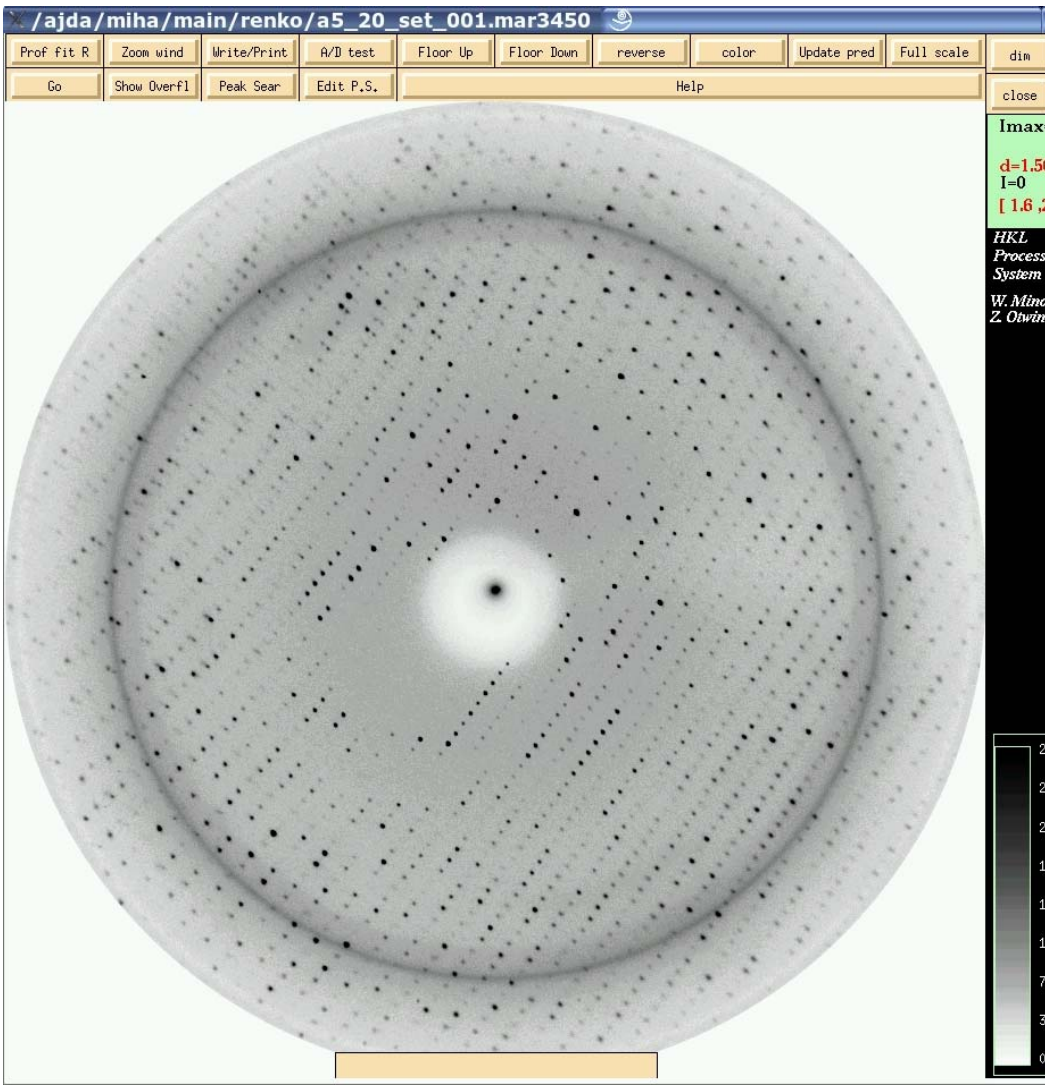


Figure 2: A diffraction pattern. The pattern after exposing a protein crystal to X-rays rotating the crystal for 1º viewed in HKL2000 (Otwinowski and Minor, 1997)

Unfortunately, it is not possible physically to focus an X-ray diffraction pattern to obtain the image of the diffracting object. This has to be done mathematically with the help of computers. The lens simulation with the computer is called Fourier transform, where structural factors are deascribed as

$$F(h,k,l) = \int_x \int_y \int_z f(x,y,z) e^{-2\pi i(hx+ky+lz)} dxdydz \qquad (2)$$

X-rays are diffracted by the electrons in the protein crystal. During the experiment only the intensities and not he phases of the structure factors are measured. The magnitude of the structural factor $|F_{hkl}|$ is proportional to the amplitude of the recorded reflection $(I_{hkl})^{1/2}$. Each reflection intensity contains contribution of all atoms in the structure and conversely, each atom contributes to the intensity of each reflection. Each structure factor has amplitude and phase. The lack of phase information is known as the phase problem. Consequently, the result of an X-ray experiment after the solution of the phase problem is a three-dimensional map showing the distribution of electrons in the structure called an electron density map.

The representations that connects electron density ρ(x,y,z) to the diffraction pattern is Equation 3.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} F_{hkl} e^{-2\pi i(hx+ky+lz)} \quad (3)$$

Where F_{hkl} are structural factors that describe the reflection and h,k and l are indices of the reflection.

1.3 Preparing the Protein

First, it is necessary to obtain a pure sample of the target protein. It can be done by either isolating it from its source, or by cloning its gene into a high expression system (Hughes and Stock, 2001). The sample needs to be assessed for suitability for crystallization by inspecting its activity, stability, solubility and other biophysical properties. If the sample fails one or more of the above criteria, it may be worthwhile returning to the expression and purification protocols and trying different approaches, such as the addition of ligands known to interact with the protein, or adding extra purification steps. It may be worthwhile switching to a different expression system altogether or working with a mutated or truncated construct.

1.4 Protein Crystallization

Obtaining suitable single crystals is the least understood step in macromolecular crystallography. The methods predominantly used are the sitting and the hanging drop vapour diffusion. It is a trial-and-error diffusion method in which the protein is slowly precipitated from its solution (Hampel et al. 1968, Giege et al. 1977, Richard et al. 1995, Jerusalmi and Steitz, 1997). Normally crystallisation experiments are set up on crystallization plates with up to 384 different conditions per tray using pipetting robots and as much as 2000 different precipitating conditions. Pipetting robot systems are mainly designed for the sitting drop technique. (**Figure 3**)

Figure 3: *Phoenix pipetting robot.* (http://www.artrobbinsinstruments.com/phoenix.html)

When sufficient amounts of protein are available for crystallization, it is usually used to perform one or more sparse matrix initial screens. Under optimal circumstances it is possible to get one or more "hits" in the initial screens, upon which further experiments are designed. Knowing the composition of all crystallisation components of the initial hit(s), varying the concentrations of all components, slight pH changes, adding additives, switching to similar buffers or precipitants, or even using different crystallization methods such as seeding, streaking, micro batch crystallization under oil and dialysis crystallization, are used to obtain diffraction quality single crystals. Sometimes good crystals will form overnight, but usually it takes from several days to several weeks for the crystals to grow (Bergfors, 1999).

1.5 X-ray Data Collection

Once the crystals are obtained, they are exposed to X-rays. A single crystal is mounted either in a capillary to be exposed at room temperature, or caught in a loop and flash-cooled to 100 K using liquid nitrogen cryo system (Petsko 1975, Hope et al. 1989, Hope 1990, Gotz et al. 1991, Vali 1995, Garman and Mitchell 1996). Nowadays, most data collection is done using the latter method. The next step is attaching the prepared crystal to a goniometer head in order for the crystal to be positioned accurately in the X-ray beam by means of a number of adjustment screws. For the cryogenic data collection, a cold nitrogen gas stream keeps the crystal at 100 K throughout the experiment. Today's X-ray sources are mostly synchrotron beam lines although in-house Cu rotating anode X-ray generators (Yoshimatsu and Kozaki, 1977, Phillips 1985) or similar may still be used (**Figure 4**). Focused X-rays emerge from a narrow tube called the collimator and strike the crystal to produce a diffraction pattern. This is recorded on the X-ray detector being an image-plate or a CCD. In a successful experiment, clean sharp spots should be observed: a uniform lattice of spots indicating a single crystal, and there should be no evidence of salt or ice crystals present which give rise to very strong spots or rings. Water rings overlap and hide protein diffraction spots, making data less interpretable.

Figure 4: *An X-ray source.* A Cu-rotating anode X-ray source RU-H2R (Rigaku Corporation, Japan).

From such an image, the crystal symmetry, the unit cell parameters (dimensions, angles), the crystal orientation and the resolution limits are determined. The data collection strategy should be designed in a way to maximize both the resolution and completeness of the data set. After choosing the appropriate strategy, a whole data set is collected: the crystal is rotated through a small angle, typically 1 degree, and images (X-ray diffraction patterns) are recorded. The total rotation angle depends on the crystal symmetry. The lower the symmetry, the more data (bigger total angle) are required.

1.6 Solving the Phase Problem

In order to visualise the structure of the target protein, it is necessary to solve the phase problem. In protein crystallography, it can be done in several ways.

When the 3-dimensional coordinates of a similar protein to the one of interest exist, the structure might be solved by a molecular replacement. The method involves rotating and translating the existing model in order to match it with the data of the protein of interest. Hence the name: molecular replacement. The unknown protein is replaced by a known model in order to solve the phase problem. If the method was successful, the amplitudes and phases for the electron density map of the new model can be calculated.

When such a model is not available, the isomorphous replacement method in one of its variants should be used. One or more heavy atoms are introduced into the crystal lattice without perturbing it. Preparing a heavy-atom derivative can be a trial-and-error experiment. Heavy atoms have higher electron density compared to the lighter element atoms that compose biomolecules. Thus, heavy atoms give measurable differences in the spots intensities in the diffraction pattern. By measuring these differences for each reflection, it is possible to derive the missing phase data. In practice, usually more than one heavy atom derivative is required to get good enough phases - hence the name multiple isomorphous replacement (MIR).

In some cases, crystallographers can make use of the anomalous scattering of certain atoms in the lattice at or near their X-ray absorption edges to gain the phase information. Many of the heavy atoms used in isomorphous replacement can also be used for this and the additional information can enhance the structure solution. This method, called multi wavelength anomalous dispersion (MAD), relies entirely on the anomalous differences measurement produced by one or more anomalously scattering atoms in the crystal. MAD potent phasing has been exploited and widely used in last few decades. In practice, three or more consecutive data sets are recorded of the same crystal at different wavelengths around the X-ray absorption edge of the anomalous scatterer. As this method requires a tuneable X-ray source, it can only be performed at a synchrotron. The resulting phase information often produce very high-quality electron density maps. Selenium is a particularly useful anomalous scatterer as it can be incorporated into proteins by over-expressing them in auxotrophic *E. coli* strains that are grown on minimal media supplemented with selenomethionine instead of methionine (Hendrickson et al., 1990).

1.7 Model Building, Refinement and Validation

After the density map has been completed, the structural model of the molecule has to be built. **(Figure 5).**

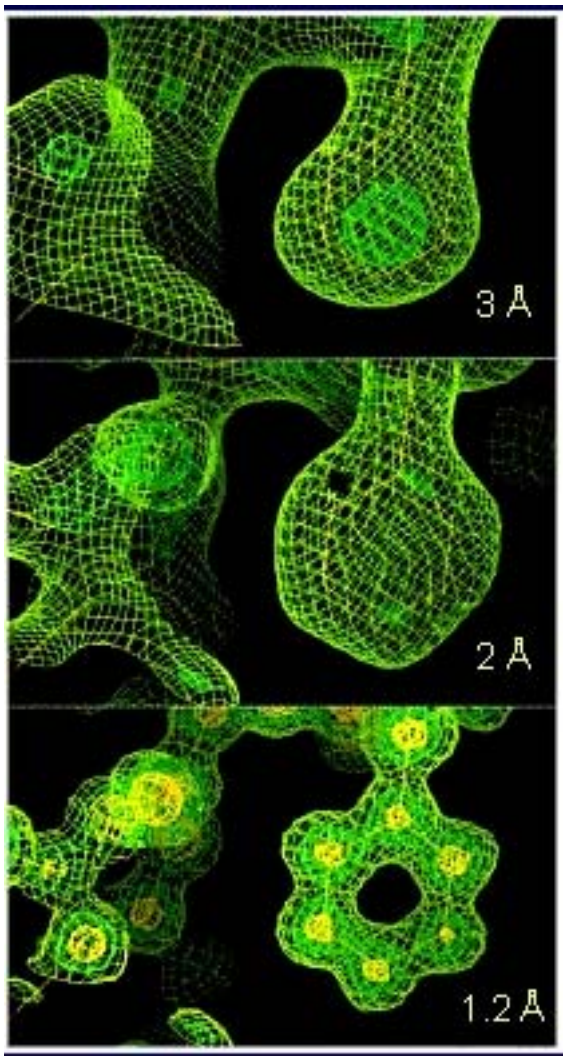


Figure 5: P*rotein electron density.* The electron density map resolution has a big impact on the protein model interpretation.

In the case of molecular replacement, the starting coordinate set is obtained from the model of the homologous protein structure. In the case of MIR and MAD, only the electron density map is obtained. In either case each amino acid residue must be fitted into the electron density map. The usual procedure is to fit the protein backbone first and then fit the amino acid residues. The details that can be seen in the map depend on the resolution and the quality of the phases. Often regions of high flexibility are not visible at all.

When crystal structures of proteins or small molecules are used to address questions of scientific relevance, the accuracy and precision of atomic coordinates are crucial. Therefore, once the preliminary model is built is solved, the atomic model is generally improved by refining it to increase the agreement with the observed diffraction data. In this way, the phases are improved and a more accurate electron density map and a better model are derived. Refinement and model building are an iterative process. It takes several cycles until convergence.

Methods used to find an optimal structure are either one of the gradient-descent methods or use of protocols that include simulated-annealing optimization and/or molecular-dynamics optimization (Tronrud, 2004, Eyck and Watenpaugh, 2001).

The most common macromolecular crystallographic refinement involves restrained optimization of the agreement between diffraction amplitudes calculated from an atomic model and those derived from the experimental data (Jensen, 1985). Diamond (1971) introduced the use of a constrained chemical model in the fitting of a calculated electron density map in a 'real-space' refinement procedure. Diamond (1971) and Watenpaugh et al. (1972) showed that phases derived from previous cycle of real-space fitting could be used to calculate the next electron-density map. The next step was application of target function in reciprocal (Fourier space). Konnert and Hendrickson (1980) for the first time applied chemical restrains a refinement program.

Different programs have different schemes for incorporating the stereochemical information, but for ease of conception the energy model is used to explain the process. The different terms can be thought of as energy terms in an equation combining all the information. In crystal structure refinement involving the use of energy restraints, an energy function based on the chemical information is combined with an X-ray target function to be minimized (Jack & Levitt, 1978),

$$E_{total} = E_{molecule} + E_{X\text{-}ray} \tag{4}$$

where E_{total} is the function to be minimized, $E_{molecule}$ is the geometric energy function (explained later – equation 8). The restrains are also weighted depending on the resolution of the data. When using high resolution data stereochemical restrains have low weight and when data is collected at low resolution restrains have high weight (Tronrud, 2004, Eyck and Watenpaugh, 2001). $E_{X\text{-}ray}$ is the X-ray target function. The straightforward choice for the X-ray target function is the least-squares residual.

Methods of improving and assessing the accuracy of the position of atoms in crystals rely on the agreement between the observed and calculated data. Goal of refinement is to derive structures, which correspond best to the observed measurements. In crystallography the measure of agreement between between the observed and calculated amplitudes is the R-factor (equation 5) defined as

$$R = \frac{\sum_{hkl} \left\| F_o(hkl) \right| - k \left| F_c(hkl) \right\|}{\sum_{hkl} \left| F_o(hkl) \right|} \tag{5}$$

where |Fo| is the experimentally observed structure factor and |Fc| is that calculated from the model.

The lower the value the better the model. As the agreement increases the value of R-factor becomes smaller until it reaches 0.0. For everyday cases the R-factor value of the final model should be around 0.20. An improvement from traditional refinement practices, which use all the observed reaction data to refine structures, is a procedure that includes cross-validation, as suggested by Brunger (1992). In this procedure, about 5±10% of the reflections are set aside for the purpose of validating the model and excluded from the working set of X-ray diffraction data used in minimization. The R value computed for this reserved set of reflections is called the free R value and is an unbiased indicator of how well the structural model fits the experimental observations. Only then detailed conclusions regarding chemistry, structure and function of the protein can be drawn.

$$E_{X-ray} = E^{LSQ} = \sum_{hkl} \left[\left| F_o(hkl) \right| - k \left| F_c(hkl) \right| \right]^2 \tag{6}$$

Fo and Fc in equation 6 are the experimental and scaled calculated structure factors for reflection hkl.

However, as pointed out by Read and others (Brunger & Adams, 2002; Read, 1986, Read, 1990; Pannu & Read, 1996; Adams et al., 1997, Murshudov et al. 1997), use of the residual as the target function is only justified for models that are very close to the true structure, which is often not the case in macromolecular refinement. An improved target function can be derived using the maximum- likelihood formalism (MLF),

$$E_{X-ray} = E^{MLF} = \sum_{hkl}\left(\frac{1}{\sigma^2_{ML}}\right)\left[\left|F_o(hkl)\right| - k\left\langle\left|F_c(hkl)\right|\right\rangle\right]^2 \quad (7)$$

which is more suitable for the general case of incomplete models and models that contain initial bias. The difference between least-squares and maximum-likelihood refinement is in the criteria which are used to measure the agreement between the observed and calculated data. Least-squares method evaluates refinement problem by minimizing the variance between the observed and model structure factors, whereas in the maximum-likelihood approach maximizes the probability of the model to describe the observed data.

The basic idea of maximum likelihood is quite simple: the best model is most consistent with the observations. Consistency is measured statistically, by the probability that the observations should have been made. If the model is changed to make the observations more probable, the likelihood goes up, indicating that the model is better. (You could also say that the model agrees better with the data, but bringing in the idea of probability defines "agreement" more precisely.) (Bricogne and Irwin, 1996, Murshudov et al. 1997, Pannu and Read, 1996)

The probabilities have to include the effects of all sources of error, including not just measurement errors but also errors in the model itself. But as the model gets better, its errors clearly get smaller, which means the probabilities become sharper. The sharpening of probabilities also increases the likelihood, as long as they are no sharper than appropriate. Mathematically, the likelihood is defined as the probability of making the set of measurements. One way to think about likelihood is that we imagine we haven't measured the data yet. We have a model, with various parameters to adjust (coordinates and B-factors in the case of crystallography), and some idea of sources of error and how they would propagate. This allows us to calculate the probability of any possible set of measurements. Finally, we bring in the actual measurements and see how well they agree with the model (Pannu and Read, 1996).

Practically during the process of refinement of macromolecule several stages are evident. In the initial phase, the search model, positioned by the molecular replacement or experimental phase, is used to calculate the electron density map. In the intermediate phase, the models are partially refined and more or less complete. They still contain regions with errors and lack flexible loops, ligands and solvents molecules. The positions of the residues still need to be examined and adjusted to best fit the electron density maps. In the final phase the last remaining, weak density and dubious map features require interpretation. They are commonly occupied by ligands attached to the macromolecular structure or flexible, likely surface located regions and residue side chains, and exhibit larger degree of disorder, when compared to the core of the structure.

Finalizing atomic **r**esolution models is generally straightforward. Unless one is working at very high resolutions, which are still very rare (Wlodlauer et al. 1984, Jelsh et a. 2000), macromolecular X-ray crystallography is a notoriously poor method for determining the structure of small molecules that are bound to macromolecules and it has been pointed out by a number of people that the stereochemical quality of more than a few small-molecule structures encountered in the worldwide Protein Data Bank (wwPDB; Berman et al., 2003) is less than overwhelming. Part of the explanation of this phenomenon lies in the general limitations of macromolecular crystallography, where due to limited resolution (and information content) and weak data (leading to a low signal-to-noise ratio), the data-to-parameter ratio of observed (Fo) to refinable parameters (x, y, z, B) is below 1.0 and the problem is undetermined. This means that in typical cases the data-to-parameter ratio is of the order of 0.5–5, where one would prefer to have values in excess of 10. The lack of data can to some extent be compensated for by the use of prior knowledge in the model refinement process. The data-to-parameter ratio can be improved by reducing the number of model parameters (by applying constraints) or by increasing the number of observations (in the form of restraints). A restraint expresses empirical knowledge (or expectations) regarding the chemistry or physics of a system in the form of a condition on one or more parameters (often in the form of a target value for a single parameter, with some indication of the allowed deviations from that value) (Kleywegt, 2007). Sometimes even at high resolutions, the error in the data, due to the relatively weak scattering of protein crystals, especially at higher resolutions, make restrained refinement a good idea at all resolutions (Jack & Levitt, 1978; Hendrickson, 1985; Tronrud et al., 1987, Dauter et al., 1997). The purpose of the former is to reduce the number of adjustable parameters, whereas the latter essentially increases the amount of observations by

supplementing the X-ray data with stereochemical information. Generally, the lower the resolution of the X-ray data, the more heavily the restraints are weighted in the refinement. The typical restraints are very similar to the energy terms in molecular-mechanics force fields, which are energy terms representing bond lengths, bond angles, chirality, planarity and nonbonded repulsion (Hendrickson, 1985; Brunger & Adams, 2002) and are weighted in such a way that the deviations match those found in databases of high-resolution structures. Refinement of the crystal structures of biological macromolecules on general therefore cannot be performed using the measured data alone when owing to limited resolution of the data observed, the data to parameter ratio is insufficient.

1.8 Geometry Parameters

The determination and refinement of macromolecular crystal structures, specially at resolutions below 1Å (Dauter et al., 1997), requires the knowledge of the geometry of the residue components of amino or nucleic acids, including bond lengths, bond angles, planar and chiral improper angles and dihedral angles as a supplemental information to the experimental X-ray data. This increases the ratio of the observations (reflections, geometric information) to model parameters (co-ordinates, temperature factors). The geometric restraints arise from the macromolecular structure chemistry: it should reflect the geometries governed by chemical physics. This restrains especially bond lengths, bond angles, planarity and improper dihedral angles. Additional restrictions come from the fact that these values should closely conform to the structures determined by the same experimental method. After introducing of the amino acid parameter sets by Engh and Huber and nucleic acid parameter sets by Parkinson (Engh and Huber, 1991; Parkinson et al. 1996, Engh and Huber 2001), derived from the Cambridge Structural Database (CSD) (Allen et al., 1979), refinement of macromolecular crystal structures has improved.

However, for the so called "hetero" compounds (The small or hetero molecules are physiological ligands, cofactors, lead compounds, substrate analogues, etc.) the situation is not so favourable. When the system under investigation contains a ligand, inaccurately determined parameters for the ligand can lead to significant structural errors in the final model. As a result, the structures of small molecules found in complexes with biomacromolecules are often less reliable than those of the surrounding amino or nucleic acids. The reason for this is most likely the infinite diversity of small molecules compared to proteins and nucleic acids that are conveniently made up of small set of building blocks (Kleywegt et al., 2003). The accuracy of their structures may be of crucial importance in interpreting their (potential) biological roles. Therefore, it is desirable to have restraints that would be able to cover entire chemical space and still be able to accuartely represent the physical interactions. The proposed thesis represents a way to overcome this gap.

When determining the crystal structure of the macromolecule-small molecule complex, the null hypothesis is usually: ‘The crystal contains the compound that had been soaked in or had been co-crystallised with and it has the ideal geometry’. The assumption that the geometry is ideal is usually true, although it has to be kept in mind that deviations may occur owing to steric strains, unexpected pH effects or ionic strength, so only a very convincing density in high resolution maps should be allowed to tempt one to change the assumption. Usually, the major problem is to define the restraints that influence the ideal geometry, as well as to find the appropriate (‘ideal’) target values for the restraints (Kleywegt, 2007). The created hetero molecules geometry parameters library - PURY (Andrejasic et al., 2008) poses a solution to both questions.

Several attempts have been made to fill the gap between the geometric parameters for macromolecules and small molecules. There are a few software packages capable of generating topological descriptions together with the corresponding geometric restraints for the hetero molecules refinement, such as PRODRG (Schüttelkopf et al., 2004), smiles2dict (Greaves et al., 1999)/libcheck (Vagin et al., 2004), HicUp/XPLO2D (Kleywegt and Jones, 1998), CORINA by Molecular Networks GmbH, eLBOW by PHENIX (Adams et al., 2002) and AFITT (Wlodek et al., 2006). The common features are that all of them use parameters with a predefined set of atom classes originating from various force fields and that all of them use only a selection of published values of bond distances and angle values. The “Hess2FF” software is an attempt to construct restraints on a purely theoretical basis (Nilsson et al., 2003).

Quantum-mechanical methods have been used to study the electronic structures, equilibrium geometries and thermo- dynamic properties of molecules both in the gas phase and in the condensed phase, usually with a continuum description of the solvation environment. Currently, the available computational power has limited the applications of most ab initio and density functional theory (DFT) methods to small molecules of up to a couple of hundred atoms. Nevertheless, there has been continuous development in electronic structure methods whose computational costs scale linearly with system sizes (Yu et al. 2005, Yang, 1991; Yang & Lee, 1995; Lee et al., 1996; Dixon & Merz, 1996, 1997). Recently, the application of quantum-mechanical calculations to validate protein models in the crystalline state has attracted much attention. Ryde and coworkers first proposed a method, ComQum-X, that combines quantum-mechanical/molecular-mechanical (QM/MM) calculations with crystallographic raw data to refine the structures of substrates in the presence of protein environments (Ryde et al., 2002; Ryde & Nilsson, 2003a,b,c). In this approach, the ligands were treated by quantum-mechanical calculations using DFT and the proteins and solvent were modeled by the force-field parameters in CNS and AMBER (Ryde & Nilsson, 2003b).

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

2 Aims and Hypothesis

2.1 Aims

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

The first aim of the thesis is to improve the accuracy of the parameters of geometric restraints describing the existing small molecules as well as those not yet synthesised. The plan was to develop an algorithm for the atom classes recognition based on their chemical environment. On the basis of atom classification, each energy term in which they appear was elaborated in order to obtain the target values and the force constants which assume Gaussian distribution of the target values. The geometrical restraints obtained in this way are expected to improve the accuracy of the crystal structures of macromolecules in complexes with ligands of small molecular weight.

The second aim of this work is to create a geometric restraints database applicable to the widest chemical space based on the large number of high resolution experimental structures of small molecules, which is expected to serve as public portal for automatic creation of highly accurate topological and geometrical libraries. The automatic and centralised design is expected to eliminate the vast number of errors likely to happen during manual construction of topological libraries and should – owing the accuracy - improve the geometries based on the parameters in use today.

2.2 Hypothesis

Our hypothesis is that it is possible to create a database of geometric restraints of hetero molecules, which will improve geometric restraints of small molecules used in macromolecular refinement.

3 Materials and Methods

3.1 Methods

3.1.1 Chemical geometric knowledge

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) and the very high resolution protein structures that can be refined without recourse to restraining parameters (Dauter 1997). The CSD is the appropriate source of geometrical information as the structures in CSD are determined by X-ray crystallography and are small enough to be fully determined by diffraction data. Parameters derived from these structures are ideally suited for the refinement of protein structures determined by X-ray crystallography since they are derived from X-ray structures, directly reflect average centres of electron density and are accurate to within the deviations from the target values suggested for X-ray structure refinement. The geometric restraints consist of average or equilibrium values and corresponding energy constants which directly reflect the variability or uncertainty of the average values. Both the equilibrium geometry and the energy constants can be determined from the statistical mean values and the sample standard deviations of a dependable set of high-resolution crystal structures. The geometric restraints in macromolecular crystallography are usually applied such that the bond lengths are distributed about their ideal values with a standard deviation of less than 0.02Å and bond angles about their ideal values with a standard deviation of about 2º. Restraints for geometric parameters which vary more widely should then have weaker corresponding force constants but their sole definition is based on the target refinement software.

Two main approaches have been used in past to organise prior chemical geometric knowledge. The first approach is based on using chemical atom classes and the second approach is based on larger monomer fragments. In atom class approach chemical elements are divided in certain classes depending on the kind of fragment and chemical environment they belonged to. For example, C atoms characterised by different sp-hybridisation or aromatic C atoms constitute different atom classes. Clearly each assigned atom class is a compromise between structural similarity and diversity of its surroundings. When simplifications are too crude (the fragments gets too small) the chemical diversity gets lost, whereas too detailed differentiation results in limited number of analyzed geometries compromising the statistical validity of the derived terms. The most popular atom classes are those used by the AMBER (Allinger, 1977; Pearlman et al, 1995) and CHARMM (Brooks et al, 1983) programs. All possible bond distances between two atoms, all possible angle terms between three atoms and all possible torsion angles defined by four atoms are generated. The drawback of this approach is the number of all possible atom classes and their derived bond and angle values. This is used by XPLOR, CNS and MAIN crystallographic refinement programs. The second approach used encodes prior chemical knowledge into library constructed of monomers. This approach can be mainly used for biological macromolecules (proteins, DNA/RNA, polysaccharides) where number of building blocks is small and limited. (Vagin, et al, 2004) The monomeric approach can handle changes in monomers by simply adding or modifying single atoms while preserving the main structure of monomer. The widely used amino acids and nucleic acids refinement libraries constructed by Engh and Huber and Parkinson et al, use the monomeric approach despite the fact that all monomers are constructed out of atom classes. For hetero compounds library the method using atom classes was selected since assuming the vast diversity of potential chemical space, the preparation of fragment library is unimaginable.

3.1.2 Molecular mechanics and Potential function

Molecular mechanics approach to handle large macromolecular systems on computers arose from a practical fact that most ab-initio and other molecular modelling techniques are too demanding in terms of time and computer processor power. Molecular mechanics is a mathematical formalism which attempts to reproduce molecular geometries, energies and other features by adjusting bond lengths, bond angles and torsion angles to equilibrium values that depend on the hybridization of an atom and its bonding scheme (this atom description is referred to as the atom class). Rather than utilising quantum physics, the method relies on the laws of classical Newtonian physics and experimentally derived parameters to calculate geometry as a function of potential energy. In molecular mechanics calculations, molecules are described using Newtonian mechanics, as a set of "balls" (atoms) held together by "springs" (bonds).

Each bond is therefore represented as a tight spring of a fixed length. Varying the length of a bond from its equilibrium value incurs a very high energy penalty. In the simplest representation, this extra energy is proportional to the square of the difference between the bond's actual length and its equilibrium bond length $\boldsymbol{l_0}$. The equation is also known as Hooke's Law. Similarly, bond angles have default equilibrium, or low energy values. For example, any of the equivalent H-C-H angles in methane will be (very close to) 109^o and the angle between any three carbons around a benzene ring will be 120^o. Bending an angle away from its equilibrium value also incurs an energy penalty. On the other hand it is relatively easy to twist around some types of bond, such as single C-C bonds. However, some positions are more energetically favourable than others. If you twist the central bond of ethane through 360^o you will pass through several preferred energy minima. Dihedral angles are terms which orient molecules into favourable twisting positions. Another geometric restrain must be applied to systems containing planar rings such as those found in tyrosine, phenylalanine and thryptophan. The forces keeping the rings planar are called 'improper dihedrals' and their associated forces involve a central and three bound atoms.

Refinement programs incorporate restraints into the target function (i.e. the function that is minimised, which can be a least-squares, maximum-likelihood or energy-based function) by adding empirical restraint functions that take different functional forms depending on the nature of the restraints. For restrained refinement of a model three things are needed: a set of definitions (atom types, bonds, angles, planar groups etc.), a set of target (‘ideal’) values for the restraints and appropriate weights for the individual restraints and for the restraint functions (to determine the relative importance of the experimental data and the restraints). In the following, this collection of items will be called a restraint set, but it goes by many other names: (stereochemical) dictionary, library, force field or topology and parameter definitions (Kleywegt, 2007).

The molecular mechanics force fields utilise potential energy functions to describe interactions between bonded and non-bonded atoms, using bonds, bond angles, dihedral and improper angles and electrostatic and van der Waals interactions. The general form of the force field equation is shown in **Equation 8,** where $E_{molecule}$ is total molecule potential energy. It is composed of various parts describing bonding and nonbonding interactions.

$$E_{molecule} = \sum_{bonds} K_b\left(b - b_0\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_0\right)^2 + \sum_{dihedral} K_\chi\left(1 + \cos\left(n\chi - \delta\right)\right) + \sum_{improper} K_\varphi\left(\varphi - \varphi_0\right)^2 + \sum_{nonbonded} \varepsilon\left[\left(\frac{R_{\min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{\min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon_0 r_{ij}} \quad (8)$$

K_b, K_θ, K_χ, and K_φ are the bond, angle, dihedral and improper angle force constants. b, θ, χ and φ are the bond, bond angle, dihedral angle and improper angle, with the subscript zero representingthe equilibrium values. Lennard-Jones attractive and repulsive 6-12 nonbonded and Coulomb electrostatics interactions terms follow. ε is Lennard-Jones well depth and R_{min} is the distance at the minimum. q_i is the partial atomic charge, ε_0 is the effective dielectric constant, and r_{ij} is the distance between atoms i and j. (MacKerell et al. 1998)

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, MM energies have no meaning as absolute quantities. Only differences in energy between two or more conformations have the meaning. Therefore the goal of the experiment is to obtain a model with the lowest potential energy of all the possible conformations. The search for minimum potential energy of example structure through artificial energy landscape is shown in **Figure 6**.

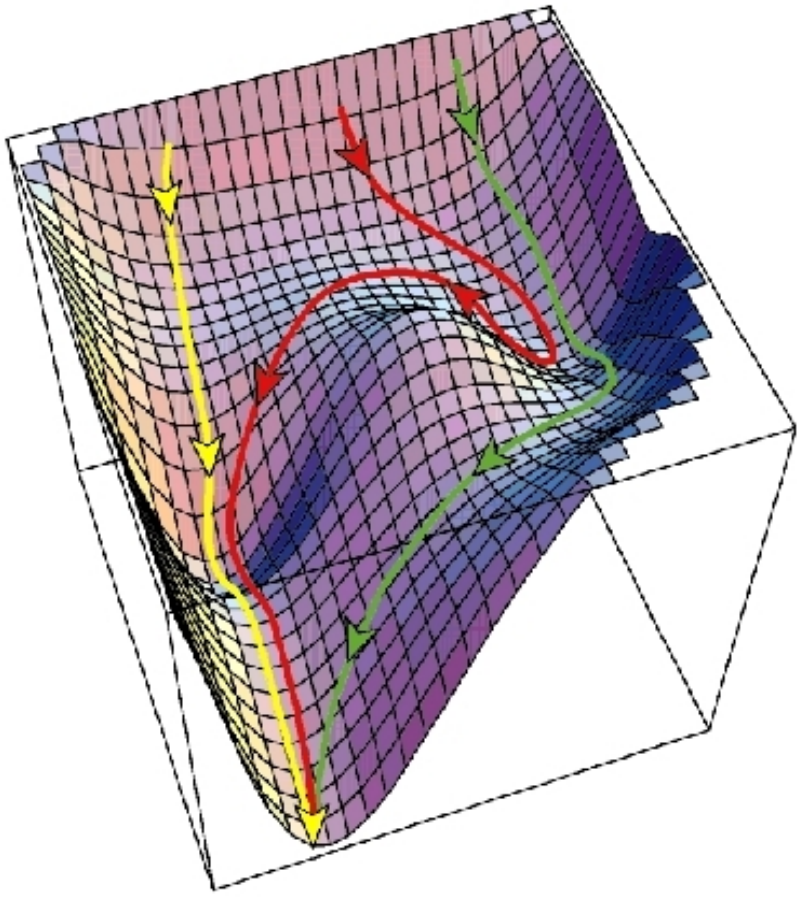


Figure 6: *Virtual potential energy landscape.* Various paths leading towards local conformation minima of protein structure in virtual potential energy landscape.

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

3.1.3 Energy terms

The terms of geometric restraints comprise bonding terms (bond distances and angles), and improper and dihedral angles. The bonding angle, improper and dihedral terms are compatible with most of the refinement programs, including XPLOR (Brunger et al., 1987), MAIN (Turk, 1992), CNS (Brünger et al., 1998), PHENIX (Adams et al., 2002) and Refmac (Murshudov et al., 1997). Bond angles are also analysed alternatively as angle distances for compatibility with the SHELX refinement programs (Sheldrick and Schneider, 1997). There is slight difference between software packages on whether they use sigma value or force constant as a restraint. SHELX, Refmac and PHENIX use estimated standard deviations or sigma of a bond length and bond angles in their formulation of restraints, rather than the force constant as is used in MAIN, CNS and XPLOR. The energy terms describing the non-bonding interactions have not been considered. Their values have been assigned in accordance with the XPLOR TOP_19 parameter set (Brunger et al., 1987).

Their distribution is in general assumed to be Gaussian and their variance is used (Figure 7) and it was accepted according to the literature. The Gaussian distribution is symmetric. The assumption that bonds stretch and compress symmetrically is therefore accepted. We are aware of the possibility that bonds may stretch more than they can be compressed due to repulsion between atoms. For small deviations this model is not a bad assumption and it can be accepted and justified. How valid is this assumption for larger deviations and how do bonds stretch in reality and is symmetrical deviation correct or should we investigate possibility that asymmetrical distribution model should be applied to the data? This two questions will be addressed in the future work.

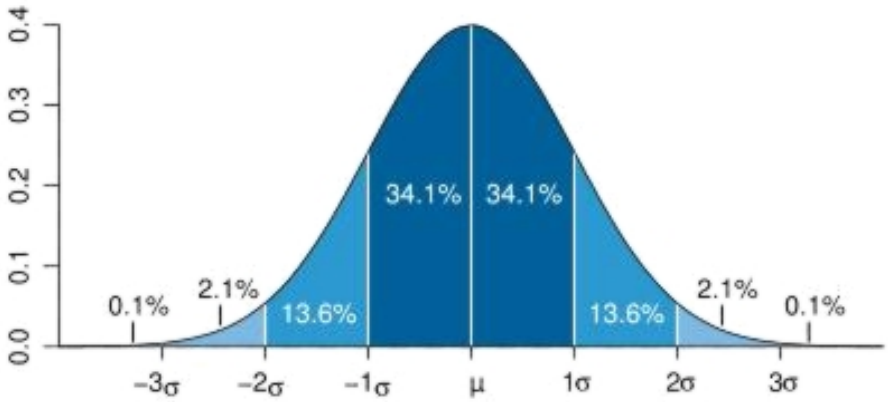


Figure 7: *Gaussian chart.* Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set (dark blue) while two standard deviations from the mean (medium and dark blue) account for about 95% and three standard deviations (light, medium, and dark blue) account for about 99.7%.

All terms, apart from the dihedral angles describing rotations about the single bonds (freely rotatable bonds), use the quadratic form of restraint function (**Equation 9)**:

$$E = k(g - g_t)^2 \tag{9}$$

Where E is the energy of the term, k is its force constant, g the geometric value of the term and g_t its target or ideal value. When preparing restraining parameters their average values are usually taken as target values. If the refinement program uses energy-function target the variance values have to be converted to force constants *k* defined by equating Boltzmann probability with the Gaussian distribution function such that $k = RT_{298}/\sigma^2$ where RT equals 0.592 kcal/mol. The force constant is specific for each specific restraint. It is derived from the sigma value of the geometric distribution, using the distribution law in the form used by Engh and Huber [Engh and Huber, 2001] (**Equation 10**):

$$k = \frac{0.592}{\sigma^2} \tag{10}$$

$$E = E_\chi * (1 + \cos(n\chi - \delta)) \tag{11}$$

For the dihedral angles the quite commonly used periodic function, the cosine term, is used (**Equation 11)**, where E is the energy of the term, E_0 is the energy barrier, χ the geometric value and δ the target value, and n represents the periodicity of the dihedral term, usually 2 for cis or trans configuration and 3 for freely rotatable bonds. In the proximity of the equilibrium the cosine function quite well mimics the quadratic form (as used for the bonding term). The force constant and energy barrier in the case of dihedrals, and the geometrical target value, are specific for each particular combination of atom classes involved.

3.1.4 Creation of connectivity

Generation of the atom classes starts from the list of bonding connections. It is followed by assigning atom states after which PURY algorithm assigns atom names, from which the data base of geometric restraints is generated. The connectivity list is read from the structure file. In the case of its absence, the list is calculated from the overlap of covalent radii (http://www.ccdc.cam.ac.uk/products/csd/radii) (Allen et al., 1979) (**Equation 12**).

$$d_{bond} = R_1 + R_2 + 0.45\text{Å} \quad (12)$$

3.1.5 Assignment of atomic states

The hybridization state of atoms is derived from the bonding parameters, taking into account neighbours and their geometrical arrangement, including bond length and angles. For example, four neighbours of a carbon atom define sp3 hybridisation. Three neighbours of a carbon atom in tetrahedral arrangement define the sp3 hybridization, whereas a planar arrangement defines the sp2 hybridization. Such sp3 atoms are checked for chirality. For carbon atoms with two neighbours, the non-linear arrangement specifies sp3 hybridization if the angle falls below the sp2 threshold, otherwise sp2 hybridisation is assumed. A linear arrangement suggests that the carbon atom is involved in either a triple bond or two double bonds, thus specifying sp1 hybridisation. It should be noted that, for atoms with only one covalent bond, the neighbouring atom and the bond length are the only source of information about its chemical environment.

The next state of atoms is their involvement in rings. Only ring structures with up to 12 members are considered. Ring structures are divided into planar and non-planar systems, in accordance with their aromaticity. When the maximal distance of a ring member from the LSQ plane of the rings stays below the cut-off value (0.1Å), the ring is considered planar, otherwise it is considered non-planar. Exceptionally, a ring is considered planar (aromatic) when all ring members are sp2 hybrids. For a chain of non-ring aromatic systems, planarity is not considered explicitly. Atoms in such chains are evaluated as sp2 hybrids.

3.1.6 PURY atom class assignment

The atom class code is meant to be human readable. Each atom class is composed of 4 ASCII characters to ensure compatibility with the existing software. (The meaning of characters is occasionally position and context dependent to compensate for the limitation of length of the class code.) The resulting atom class algorithm is rather branched, using numerous conditions. Below only the basic rules are presented.

The generation of classes for organic elements (C, H, N, O, P, and S) is different from that for others, in order to accommodate the larger diversity of the compounds. For organic elements, only the first position is reserved for the element name, whereas for the others the first two positions are used. In the cases of the two letter chemical elements, the second position is always written in small caps, whereas in cases of the non-organic elements with a single letter code, the second position is "_".

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

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

The fourth character describes the geometric arrangement of the neighbours of an atom. The non-ring atoms are described with the characters “X”, “Y”, “I”, and “<”. "X" denotes four valences, including the sp3 hybrids such as ammonium (NH3) which has one free electron pair. "Y" denotes a planar sp2 hybrid such as is present in the amide group. "I" and "<" indicate arrangements around an atom with two neighbours. "I" describes a linear arrangement of a carbon atom involved in one triple or two double bonds, whereas "<" describes an arrangement at an angle (a sp3 hybrid) such as that for the sulphur atom in methionine "S__<".

For the aromatic ring systems the fourth character describes the size of the ring, whereas in the case of non-aromatic rings the size is written on the second place. The code for the ring accepts up to twelve members, extending the code beyond "9" by the use of “0”, “A”, and “B”. Rings with more than twelve members are however not considered as rings. For example, atom class "CC_6" describes a phenyl ring

carbon attached to a carbon outside the ring. The bridge atoms between two rings are marked with the asterisk character "*" on either the second or the fourth place. The fourth character can also describe certain exceptions such as the guanidinium group for which the character "G" is used.

In the first row of **Figure 8,** three quite common fragments are used to illustrate the coding algorithm on a non-ring system, whereas the second line illustrates three atom classes from ring systems. **Figure 9** demonstrates the procedure of atom class assignment in the case of 4-chlorophenyl acetic acid.

In the case of metal atoms involved in coordinate bonds, the metals with bonding partners of the same kind are marked as "Ca6_", where "6" indicates that the calcium ion has six coordinated atoms. The periodic elements of the 4th and 5th row can form metallo acids. They are denoted with "O" on position three followed by the number of oxygen atoms, for example "MoO5". Co-ordinate bonds are not considered as part of the atom class assignment procedure of organic compounds.

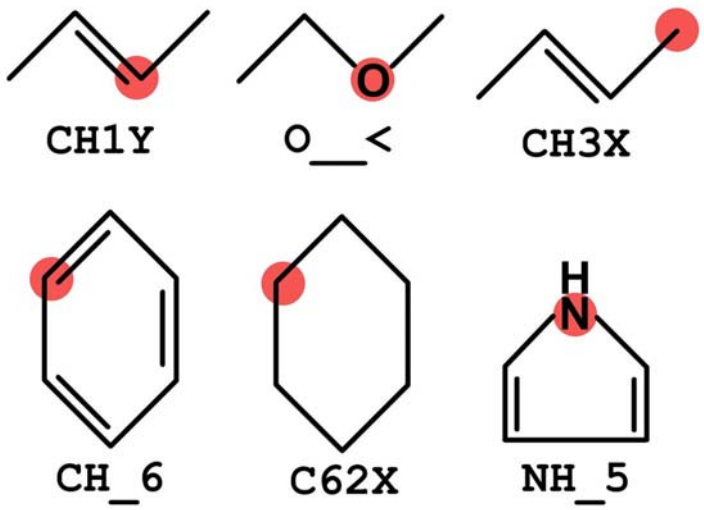


Figure 8: *Demonstration of PURY atom classes for six common fragments.* The atom with assigned class is marked with the red dot. From left to right and top to bottom are: CH1Y - sp2-hybridised carbon atom forming one double bond with bonded hydrogen atom, O__< - ether oxygen atom, CH3X - sp3-hybridised carbon atom with 3 bonded hydrogen atoms (methyl group), CH_6 - benzene carbon atom with bonded hydrogen atom, C62X - cyclohexane carbon atom with 2 bonded hydrogen atoms, NH_5 - pyrrole nitrogen atom with a bonded hydrogen atom

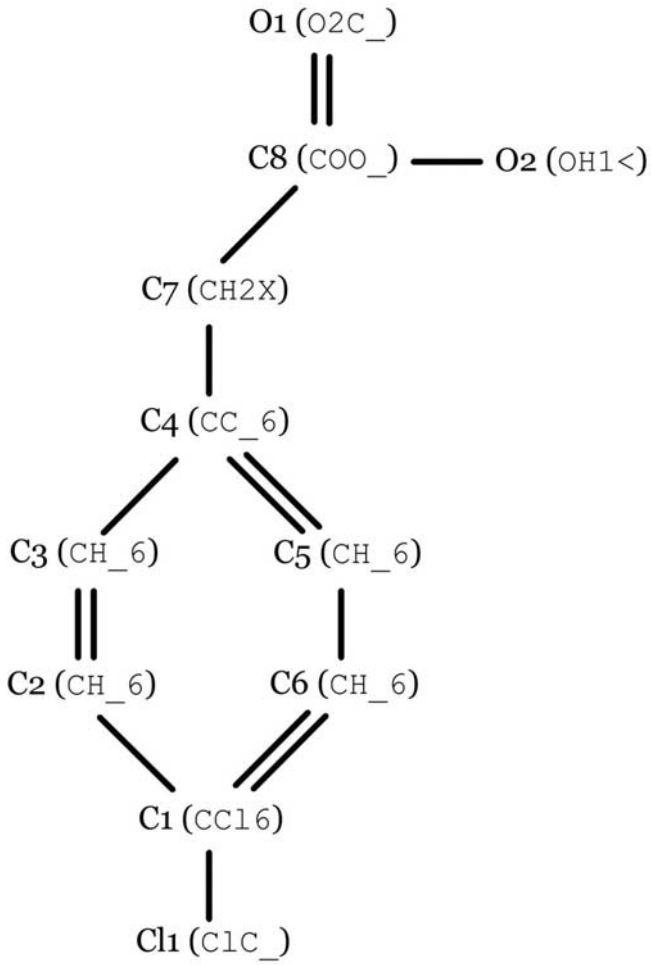


Figure 9: *PURY algorith example.* A sample description of how the PURY algorithm evaluates 4-chloro phenyl acetic acid (C8H7O2Cl) - (CSD reference AHATAE). Atoms are written with their name and derived atom class in parentheses. The ring consisting of six atoms (C1, C2, C3, C4, C5 and C6) and is found to be planar. All ring atoms are also planar so it is considered aromatic. The carbon atoms get a "6" on position four since they are all members of 6-membered aromatic ring. Ring atoms with bonded hydrogen atoms (C2, C3, C5 and C6) get "H" at position two, while other atoms obtain the name of the bound atom at positions two and three: "Cl" for chlorine for (C1), and "C_" for carbon for C4. Methylene carbon atom (C7) is assumed to have two hydrogen atoms bonded, and since the bond distance between atoms C4 and C7 corresponds to a single bond and the angle is below the sp2 threshold, it gets the class name "**CH2X**". The carbonyl carbon atom (C8) has two bonded oxygen atoms and it is assigned the special class "**COO_**". Oxygen atom (O2) with bonded hydrogen atom forms two single bonds and gets class name "**OH1<**", while the other oxygen atom (O1) forms only a single bond with the carbon atom and its distance suggests a double bond. Since oxygen is organic, its bond type is written on position two. C_ is written on the third and fourth places. The chlorine atom (Cl1) forms only one covalent bond so it is a bonding atom, which is a single character organic element, and is written on position three. The atom gets the class name "**ClC_**".

3.1.7 Generation of a data base of geometric restraints

Storage lists of all possible covalent and coordinate bond and angle terms are generated from the connectivity list. A storage list comprises a list of all appearances of each combination of atom classes for each particular geometric term. In addition to atom names and classes storage list entries also contain CSD reference code. The storage list of bonds thus contains pairs of atoms, the storage list of angles contains combinations of all bonded atoms and the storage list of improper terms is generated from all atoms bonded to three or more neighbours, among which at least 3 non-hydrogen atoms should be present. The values of

improper terms are always set positive, which means that the PURY analysis does not differentiate between R- and S- chiral centres. The storage lists of dihedral angles are generated by storing every possible combination of neighbours of all bonded atoms. Those containing hydrogen atoms at both ends are excluded.

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

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

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

Table 1: *PURY parameter examples.* Output examples for bond, angle, improper and dihedral terms for use in macromolecular refinement. The output shows atom classes, equilibrium values, corresponding force constants and multiplicity values where appropriate. Sigma values from which force values were calculated are also shown.

Entry	Class1	Class2	Class3	Class4	Force constant	Multiplicity	Average value	Sigma
Bond	CH2X	S__<			1643.6		1.818	0.0190
Angle	CH2X	S__<	CH3X		733.8		100.9	1.6274
Improper	CC_6	CH_6	CH_6	CH2X	1211.3	0	0.00	1.123
Dihedral	CH2X	CH2X	S__<	CH3X	7.41	3	90.0	20.000

3.2 Equipment

Most of the work was done on computer with an AMD Athlon 3500+ processor and 2GB of RAM running SUSE 9.2 Linux distribution. Some ab-initio calculations were done on Apple G5 computer with a dual 2GHz processor and 1GB of RAM running OSX. Around 9000 lines of the code were written in C and around 8500 lines of the code were written in PERL. The PURY core program which parses every single structural file was written in C programming language for its speed and efficiency. Compiling was done using open source gcc 3.3.4 C and FORTRAN language compiler. The user interfaces, supporting programs and www server interfaces were written in Perl 5.8.5 programming language. The server was setup on Apache web server 1.3.9. For optimum speed performance of the PURY server the SQLite format was selected for storing the parameter database.The data was extracted from Cambridge Structural Database version 5.28 using ConQuest tool (Allen, 1979). The whole database generation process takes approximately 24 hours for 165000 CSD structures on Athlon64 3500MHz with 2GB of RAM and SATA hard drives. The PURY parameters were tested on MAIN macromolecular modelling and refinement program (Turk, 1992).

4 Results

The atomic co-ordinates of the structures used for creating the PURY database were extracted from CSD v5.28. The structures were selected using CSD's ConQuest (Bruno et al., 2002) browser tool using filters “no errors” and a crystallographic R-value below 5 %. A total of 162,540 entries contained almost 10 million atom positions, 10,529,799 bond lengths, 20,342,046 bond angles, 2,703,482 improper terms, and 8,440,410 dihedral angles. From these data 1971 different atom classes were derived and parameter lists yielding 32,634 bond lengths, 236,821 bond angles, 64,133 improper and 229,821 dihedral angle restraints were generated. The numbers reporting the amount of the data on the web server may differ since PURY is evolving.

From the total of 1971 atom classes, 270 were assigned to carbon, obviously the most chemically versatile and frequently appearing chemical element. The second on the list is nitrogen with 159 classes, followed by phosphorous with 136. Sulphur and oxygen atoms, represented by 88 and 69 classes respectively, complete the group of "organic" atoms. The halogens F, Cl, I, and Br are represented by 51, 37, 29, and 27 atom classes. There are 18 Si classes and 2 hydrogen classes. In addition, there are 1091 atom classes representing the rest of the periodic table.

4.1 Assessment of the data – quality and reliability

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

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

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

The vast diversity of the data presented in CSD leads to generation of the parameters that are not statistically significant. Our decision was to include the data with low redundancy in the database despite awareness that this data may compromise the quality of the database. However the data are still better than lack of it and with this data the database does cover a substantially larger chemical space.

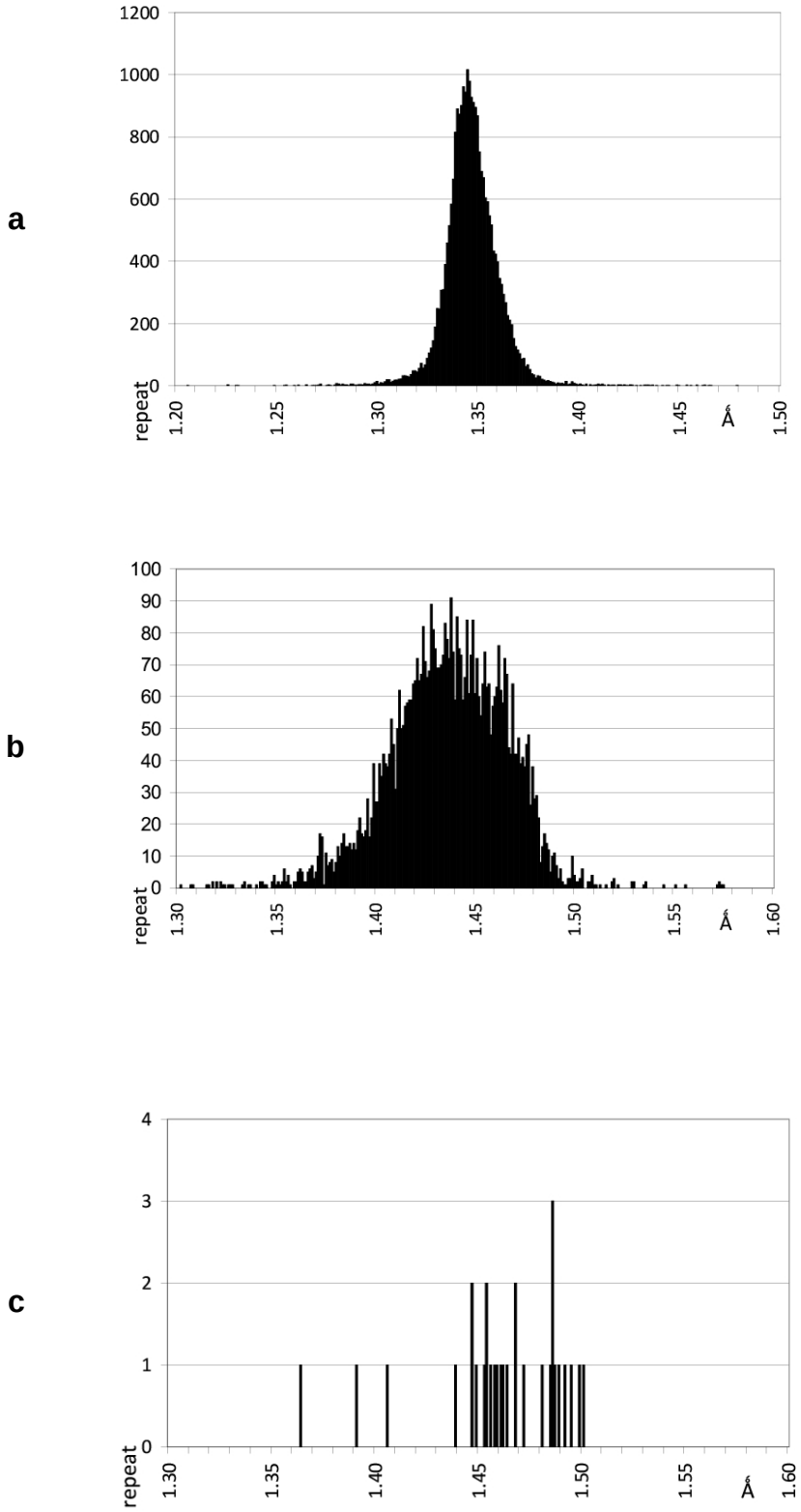


Figure 10: *Histograms of bond distances between selected atom classes.* a) Bond between sp2-hybridised carbon atom in a six-membered aromatic ring "CF_6" and a fluorine atom "F_C_". b) Bond between sp3-hybridised carbon atom with one bonded hydrogen atom "CH1X" and a sp3-hybridised oxygen atom "O__<". c) Bond between a sp2-hybridised carbon atom in a 5-membered aromatic ring with bonded oxygen atom "CO_5" and a sp2-hybridised carbon atom in a 5-membered aromatic ring with bonded chlorine atom "CCl5".

Figure 10 illustrates the connection between accuracy, precision and number of repetitions. Three covalent bond cases have been chosen, one from each population class: the first highly, the second moderately, and the third sparsely populated, belonging to "CF_6 – F_C_" (fluorine atom bonded to carbon atom in a 6 membered aromatic ring), "CH1X – O__<" (aliphatic carbon oxygen bond - ether), "CO_5 - CCl5" (carbon atoms from a 5 membered aromatic ring, the first with oxygen atom attached and the second's covalent partner outside the ring being a chlorine atom). The peak value of the first case has over 1000 repetitions and corresponds to the average 1.347 Å, while the peak value of the second case with over 90 repetitions corresponds to the average value 1.437 Å. In the third case the highest peak has only three repetitions at 1.487 Å, not really corresponding to the average at 1.462 Å. The precision of the terms is reflected in their sigma values, which are 0.0147 Å, 0.032 Å and 0.031 Å for the first, second and the third case. The shape of the first histogram leaves an impression of a highly accurate term. For the histogram of the third term it is obvious that the term is underrepresented, although its minimum and maximum bond lengths are closer than those of the other two terms; even the standard deviation is sharp, suggesting that its precision may not differ much from that of the middle term. This is a warning that indicates a more general phenomenon. Transfer of geometric restraints from similar fragments to unknown structures in the absence of the corresponding experimental structure or statistical validation of the term may be less reliable than anticipated. Only 12% of hetero compounds have a matching structural deposit in CSD (R. Taylor, personal communication).

The cases presented in **Figure 11** illustrate the behaviour for bonding, dihedral and improper angles. The bond angle of an amide fragment ("CH1X - C__Y - NH1Y") (**Figure 11a**) involved in a peptide bond has a clearly defined maximum peak which corresponds to the average value of 116.3º, and a smaller peak positioned at the other side of the 120º. The double peak of the bonding angle is reminiscent of the proline residue analysis, where coupling has been observed between the bonding and dihedral angles (Lamzin et al. 1995, Engh & Huber, 2001); however, the population of the lower peak is too low to allow firm conclusions to be drawn. The amide fragment dihedral (CH1X – NH1Y – C__Y – O2C) has the major peak at 0 deg and a much less populated one at 180º (**Figure 11c**) – an indication of the appearance of the trans and cis amide (peptide) bond conformations. The freely rotatable bond around the two sp3 hybrid carbons, dihedral CH1X-CH1X-CH1X-CH1X (**Figure 11d**), exhibits a distinct peak at 0º, in addition to the period of 3, indicating the presence of the cis conformation. The absence of peaks at +/- 120º suggests that the period of 6 is not justified, whereas the absence of any width of the 0º peak is indicative of the use of constraint during the structure refinement. This case illustrates that dihedral angle terms describing conformations of freely rotatable single bonds may not be represented well by the ideal value. Alternatively, one may use an all atom model with all hydrogen atoms included or not use any specific term for such dihedral angles (the energy barrier is set to 0) and use the 1-4 non-bonding interaction terms instead (as in the X-PLOR TOP_19 parameter set). PURY can deliver restraints for both types of restraint.

Improper angles on the other hand exhibit the least ambiguous behaviour. In "CC_6 - CH2X - CH_6 - CH_6" **(Figure 11b)** the planarity of the CG atom of residues such as TYR or PHE has a peak at 0º with sigma of 1.0º. (It needs to be emphasized, however, that the histograms of data for improper restraints are symmetric as the consequence of the way they are sampled.)

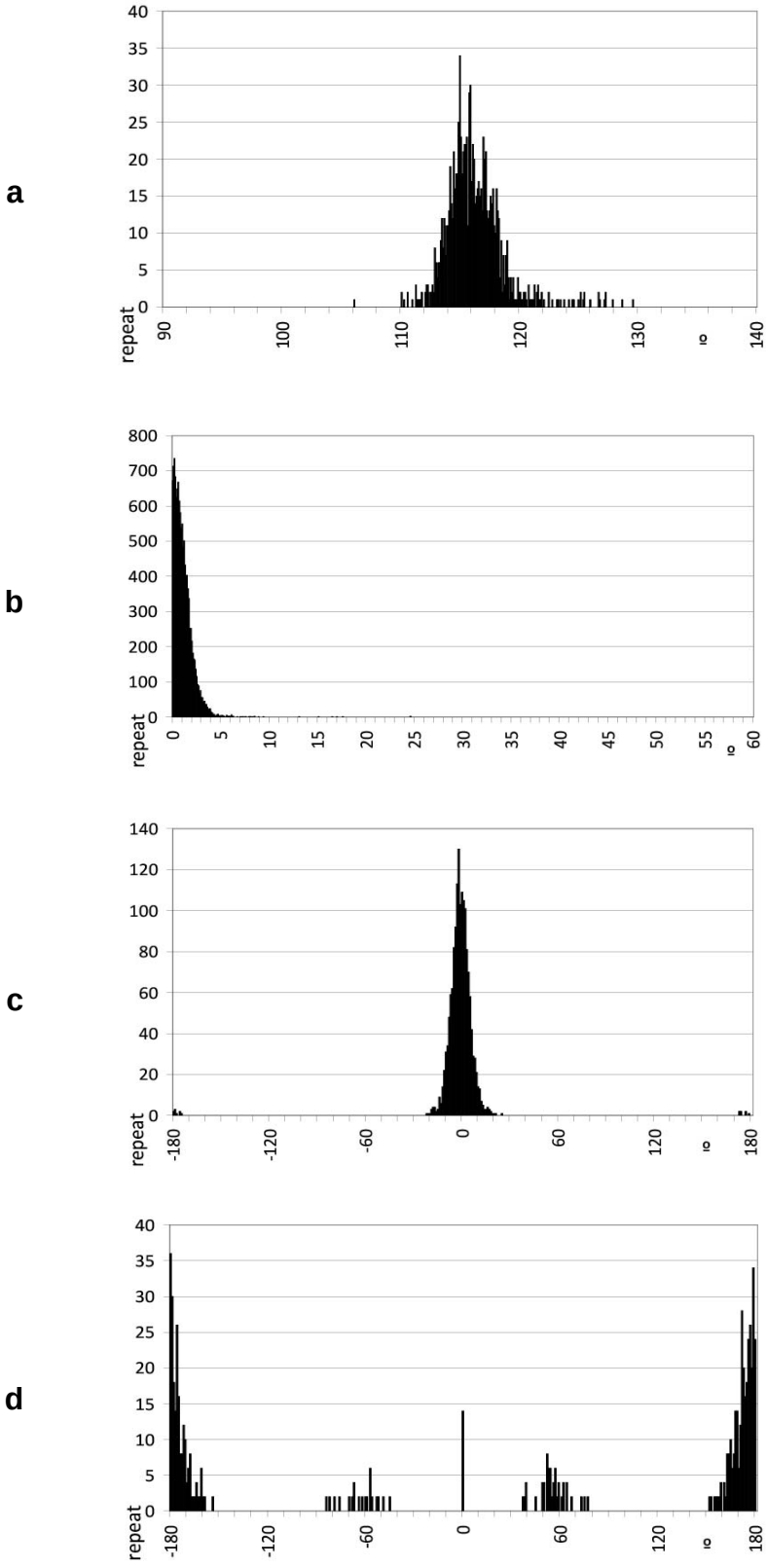


Figure 11: *Histograms of selected parameters.* a) A peptide bond angle including carbon alpha atom "CH1X", sp2-hybridized planar carbon atom "C__Y" and sp2-hybridised planar nitrogen atom with one bonded hydrogen atom "NH1Y". b) A tyrosine or phenylalanine-like improper angle around a sp2-hybridised carbon atom in a six-membered aromatic ring "CC_6" including beta carbon atom "CH2X" and two sp2-hybridised carbon atoms in a six-membered aromatic ring with bonded hydrogen atoms "CH_6". c) A dihedral angle through a planar peptide bond having only single peak "CH1X - NH1Y - C__Y - O2C_". d) A dihedral including 4 sp3-hybridised carbon atoms with freely rotatable single middle bond which has many energy minima "CH1X - CH1X - CH1X - CH1X".

4.2 Lengths of covalent bonds involving hydrogen atoms

Figure 12 shows the distribution of bond lengths of a hydrogen atom bound to an oxygen atom (hydroxyl group), to a phenyl ring carbon, to an amide, and to an aliphatic ring carbon. The presence of several high, narrow peaks on the background of an extremely broad distribution of bond lengths, ranging from 0.7 Å with several outliers even beyond 1.3 Å, in all four cases demonstrates that the positioning of hydrogen atoms is ambiguous. CSD does not provide experimental data to verify in which cases the positioning and refinement of hydrogen atom positions is supported by the diffraction data, the CSD filters do not allow us to analyze the reliability of the structures to the resolution at which they were determined. They do, however, allow structures to be chosen according to the radiation source.

Figure 13 shows equivalent bond lengths of structures determined only by neutron diffraction. In the analysed release there are 920 such deposits. They contain 33339 atoms and 66540 covalent bonds. In the total of 1692 bond lengths parameters from the neutron data, there are 121 cases of bonds with hydrogen atoms. In contrast to the whole set of CSD structures analysed, the most populated bond length of hydrogen atoms from neutron structures exhibit a single peak only. The positions of these peaks are marked on the corresponding figures derived from the complete CSD (**Figure 12**).

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

Table 3: *Neutron data.* Bond length of hydrogen atoms from neutron derived structures. We have selected only the terms represented by more than 100 repeats. The columns one and two show PURY atom classes forming the bond. The column 3three shows bond average obtained from neutron data, whereas the values in parentheses show average from the whole CSD. The column four shows corresponding sigma values while column five shows ratios between whole CSD parameter and the neutron derived one. The column six shows the ratios between the number of representatives from the whole CSD and the number of neutron data representatives.

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

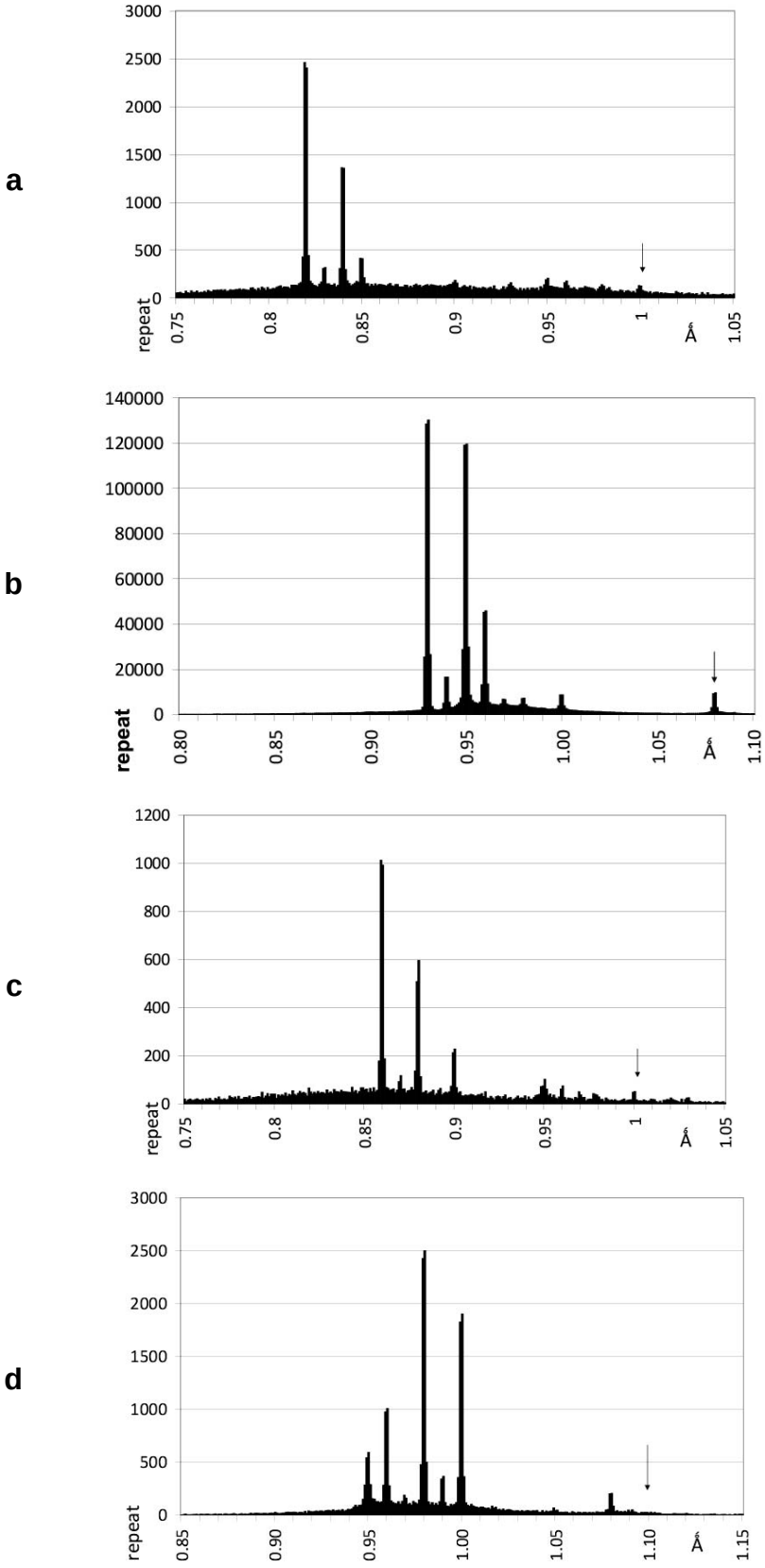


Figure 12: *Histograms of bond distances between hydrogen atoms.* a) The bond between a sp3-hybridised oxygen atom with one bonded hydrogen atom "**OH1<**" and a hydrogen atom "**HP__**". b) The bond between sp2-hybridised carbon atom in a six-membered aromatic ring with one bonded hydrogen atom "**CH_6**" and a hydrogen atom "**HC__**". c) The bond between a sp2-hybridised planar nitrogen atom with one bonded hydrogen atom "**NH1Y**" and a hydrogen atom "**HP__**". d) The bond between a sp3-hybridised carbon atom in six membered non-aromatic ring with one bonded hydrogen atom "**C61X**" and a hydrogen atom "**HC__**". The arrows mark the peaks obtained from structures determined by neutron radiation.

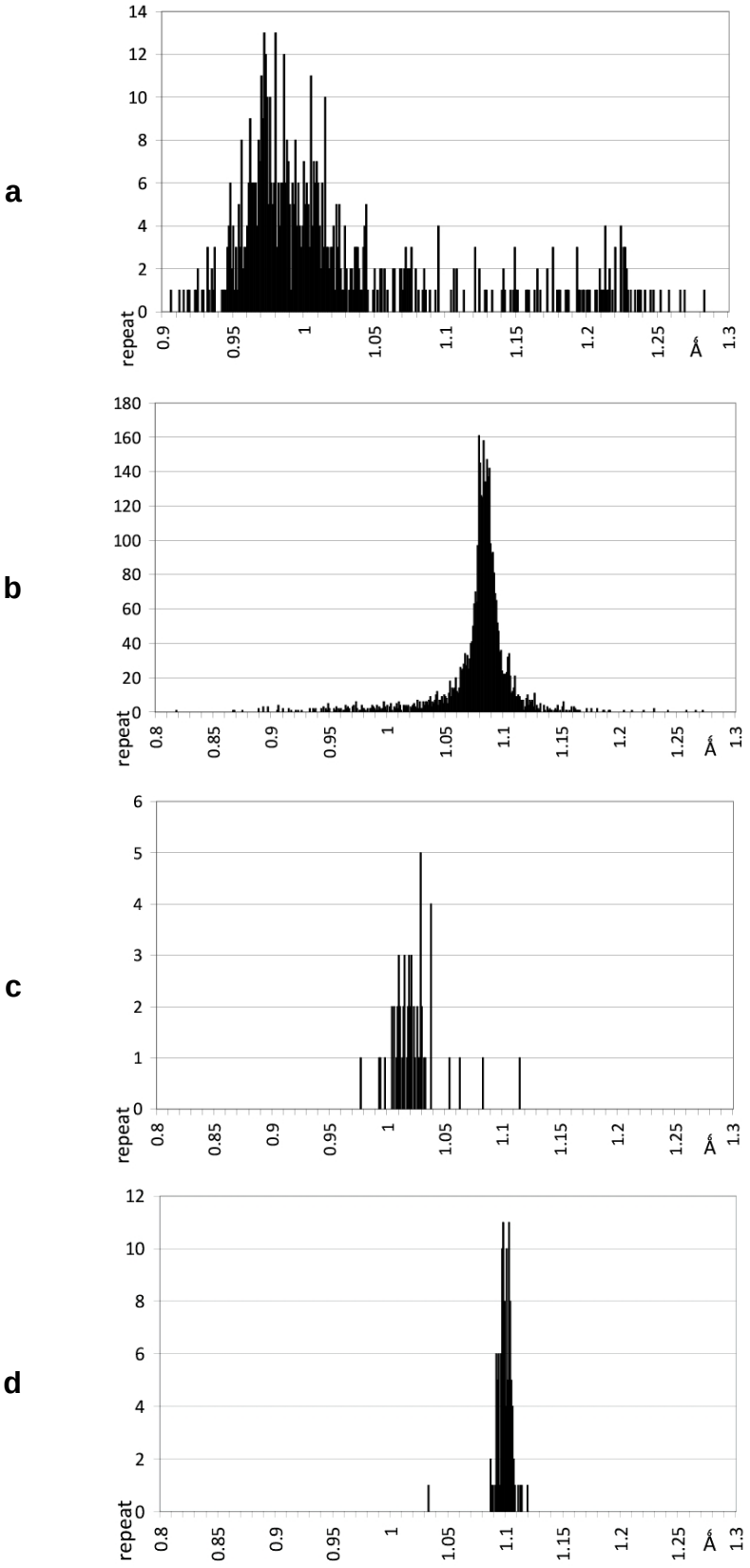


Figure 13: *Histograms of bond distances between hydrogen atoms and selected atoms.* The data from an analysis done on structures determined using neutron radiation source is presented. a) The bond between a sp3-hybridised oxygen atom with one bonded hydrogen atom **"OH1<"** and a hydrogen atom "**HP__"**. b) The bond between a sp2-hybridised carbon atom in a six-membered aromatic ring with one bonded hydrogen atom **"CH_6"** and a hydrogen atom **"HC__"**. c) The bond between a sp2-hybridised planar nitrogen atom with one bonded hydrogen atom **"NH1Y"** and a hydrogen atom **"HP__"**. d) The bond between a sp3-hybridised carbon atom in six membered non-aromatic ring with one bonded hydrogen atom **"C61X"** and a hydrogen atom **"HC__"**.

4.3 The PURY www server

The database is accessible through the WWW interface (http://pury.ijs.si), which enables geometric restraint parameters for 3-dimensional structures of molecules or fragments to be downloaded. A user has to upload the co-ordinate file of the 3-dimensional model or submit a SMILES string and download the resulting geometric restraints and topology files from the server (**Figure 14**).

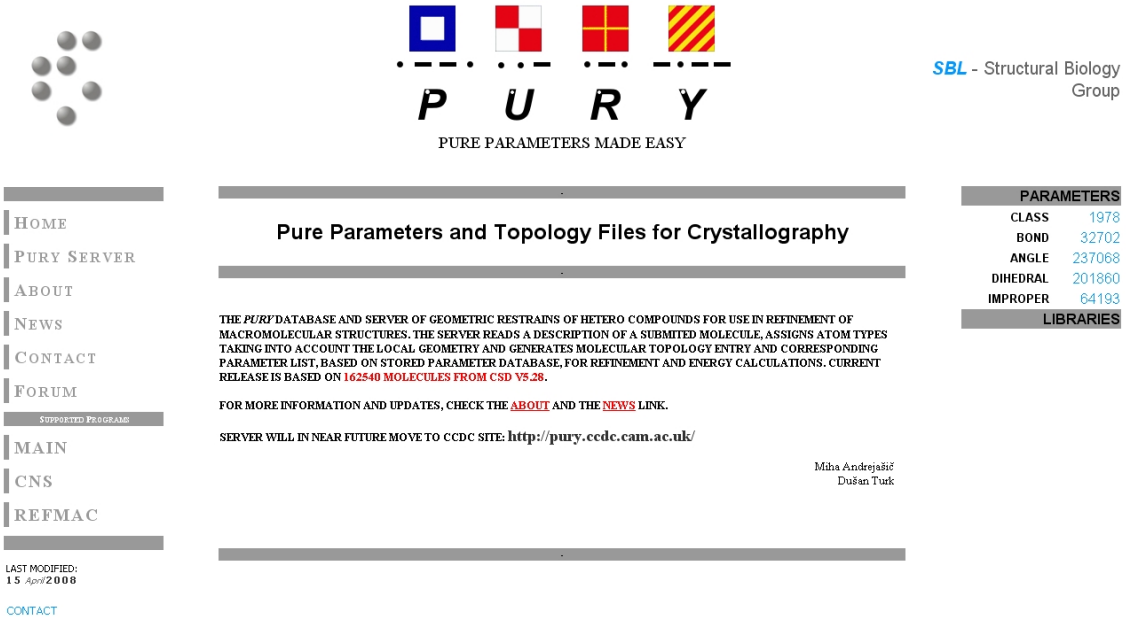


Figure 14: *Entry page to PURY server*

Currently only the PDB format for 3D molecular structure is supported for the upload. Alternatively, the starting geometry of the compound may be created on the server by interactive 3D graphical program JME (P. Ertl, Novartis). The output topology and parameter files are in formats readable by MAIN, X-PLOR/CNS, REFMAC and PHENIX macromolecular refinement programs. For REFMAC a modified ener_lib.dic special class library with PURY added classes and corresponding covalent and van der Waals radius has to be used. (SHELX ins files read in and out is on the way.)

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

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

5 Discussion

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

- By comparing a few experimental structures we have checked the consistency and the stability of the derived terms.
- By comparing the variability of the bonding terms of different parts of chemical space we have tried to assess the level of accuracy provided by the PURY parameters.
- By cross validation of the macromolecular crystal structures refined against PURY and EH parameter sets we have tried to assess the suitability of the PURY parameter set for refinement.
- By comparing PURY parameterisation with an expert derived parameter set on a clearly defined subset of chemical space (EH parameters for amino acid residues) we have tried to assess the limitations of the PURY approach and make some suggestions for EH set improvement.

5.1 Comparison with the CSD experimental data

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

The subset of CSD 1388 structures was selected by two criteria: they had to contain only C, H, N and O atoms and have B or E character on the third position in their reference code (refcode). Their bonds and angles were validated against PURY bond and angle parameters **(Figure 15a and 15b).** The histograms show that majority of bond deviations fall within 0.01 Å with the average bond RMSD of 0.008 Å. Most of angle deviations fall within 2º with the average RMSD of 1.55º.

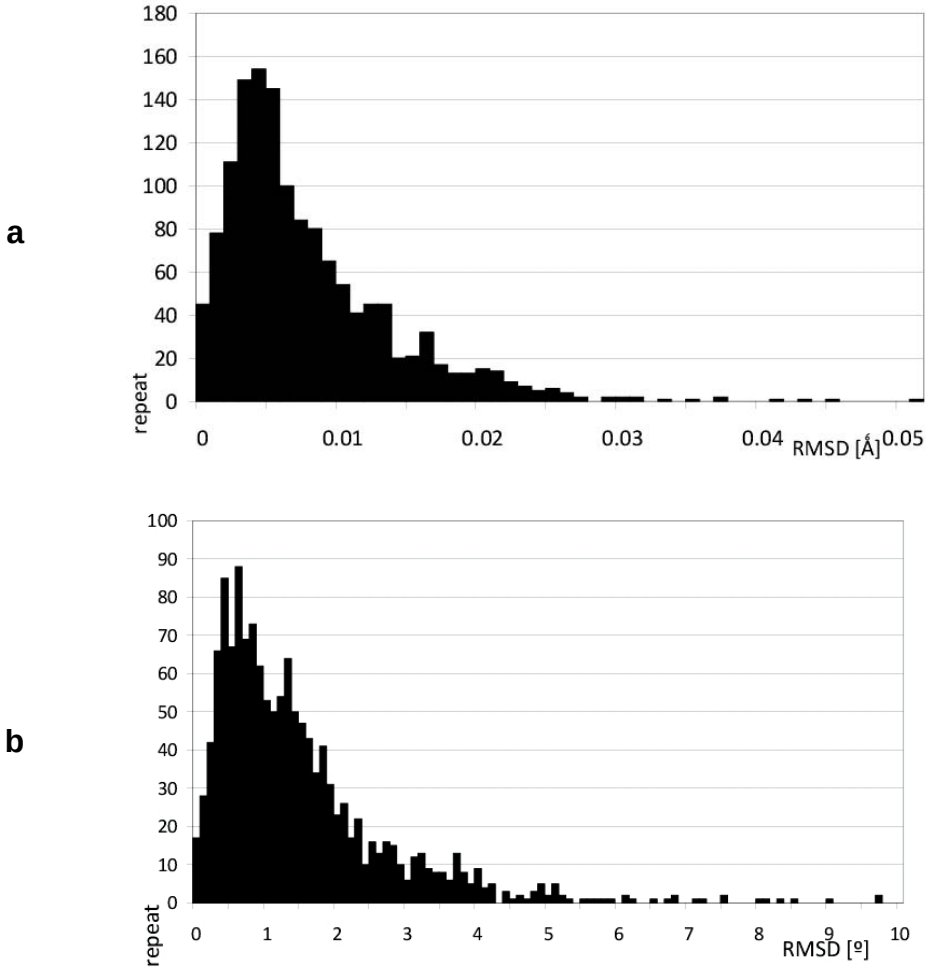


Figure 15: *Histograms of bond and angle RMSD distribution 1388 CSD structures validated with PURY parameters.* a) Bond RMSD distribution. b) Angle RMSD distribution. RMSDs for each structure separately were calculated with MAIN. The bin thicknesses are 0.01 Å and 0.1º for bonds and angles RMSDs respectively.

The three additional experimental structures were first validated using PURY geometry parameters. Comparison revealed that the RMSD for a bond of the ABIYUF structures with 0.02 Å and RMSD for angles of the ABIYUF and CEPTIA 1.38° and 1.65° correspond to tight acceptance criteria, whereas bond deviations of 0.03 Å for CEPTIA and ENAMEL and angle deviation 2.22º for the ENAMEL lie within the broadly acceptable boundaries (**Table 4, Figure 16).**

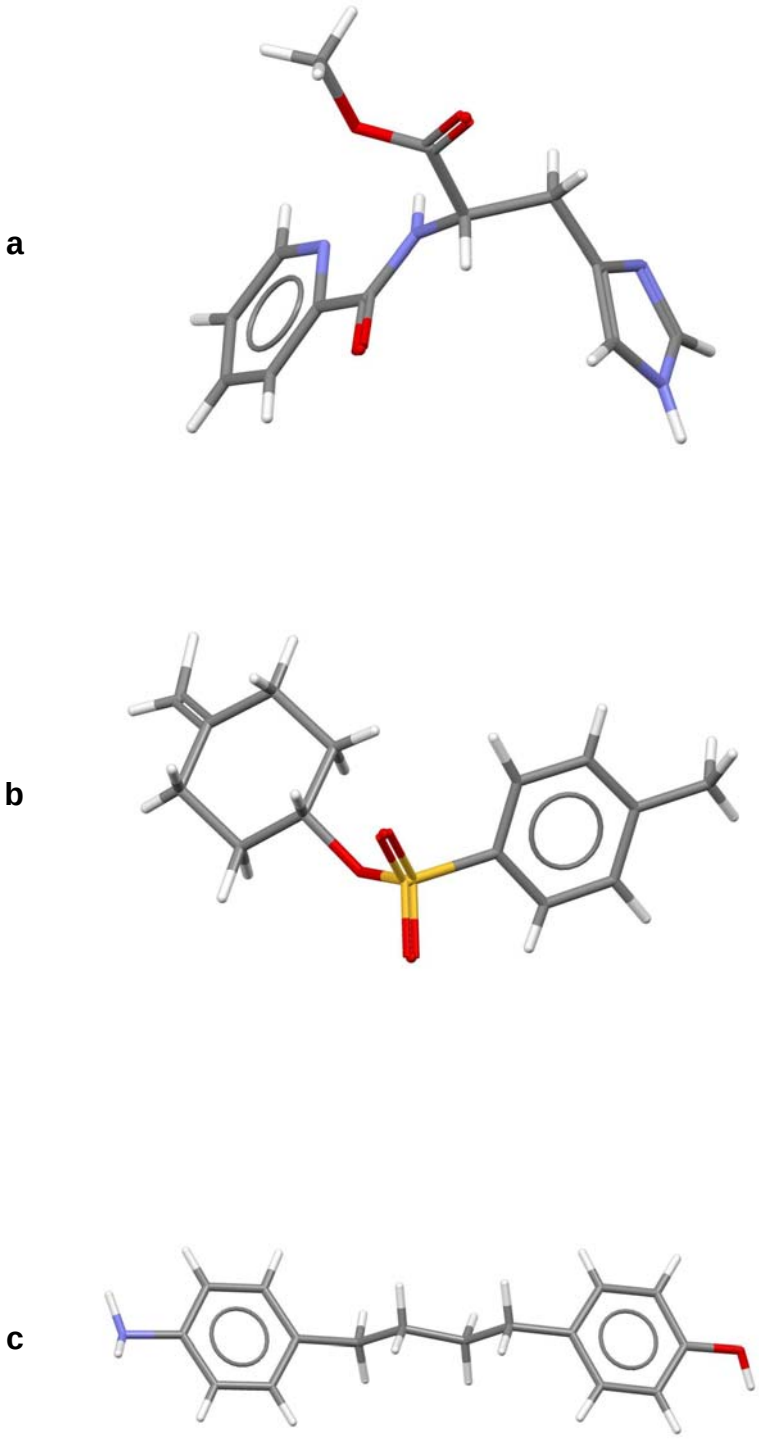


Figure 16: *CSD v5.28 selected entries.* ABIYUF a), CETPIA b) and ENAMEL c) used for geometric comparison with PURY parameters. Figures were made with CSD Mercury (Macrae et al., 2002).

Table 4: *Validation of structures ABIYUF, CETPIA and ENAMEL.* The first pair of lines shows R-factor and temperature of experiment. The second group of lines shows bond RMS values of experimental, GAMESS, eLBOW (PHENIX) and PURY models as validated with PURY parameters. The third group of lines shows angle RMS values of experimental, GAMESS, eLBOW and PURY models as validated with PURY parameters. The last group of lines shows coordinates RMS difference between models. All PURY models were energy minimised for 1000 steps using MAIN. The optimum geometric search for GAMESS models was performed with *ab initio* calculations at HF/6-31 level until the density change between two consecutive runs was less than $1.0*10^{-5}$ using the GAMESS (US) package (Schmidt et al., 1993, Gordon and Schmidt, 2005) from 22. November 2004 on G5 dual 2.0GHz with 1GB RAM running OSX. The optimum geometric search for eLBOW models was performed using eLBOW from PHENIX version 1.3 RC2 using –opt switch.

	ABIYUF	CETPIA	ENAMEL
Exp R-factor	4.84	6.8	4.18
Temperature	283 - 303 K	283 - 303 K	105 K
Bond RMS			
EXP. model	0.02 Å	0.03 Å	0.03 Å
GAMESS model	0.02 Å	0.002 Å	0.02 Å
eLBOW model		0.08 Å	0.12 Å
PURY model	0.003 Å	0.002 Å	0.002 Å
Angle RMS			
EXP. model	1.38 º	1.65 º	2.22 º
GAMESS model	1.84 º	1.65 º	2.54 º
eLBOW model		3.462 º	8.284 º
PURY model	0.707 º	0.429 º	0.324 º
Coordinates RMS			
EXP/PURY (max value)	0.040 Å (0.081 Å)	0.142 Å (0.361 Å)	0.083 Å (0.151 Å)
EXP/PURY (max value)	0.040 Å (0.081 Å)	0.142 Å (0.361 Å)	0.083 Å (0.151 Å)
GAMESS/PURY (max value)	1.175 Å (2.323 Å)	0.142 Å (0.361 Å)	0.168 Å (0.470 Å)
eLBOW/PURY (max value)		2.444 Å (5.770 Å)	1.729 Å (4.300 Å)
eLBOW/GAMESS (max value)		2.217 Å (5.430 Å)	1.754 Å (4.383 Å)
eLBOW/EXP (max value)		2,453 Å (5.716 Å)	1.731 Å (4.362 Å)
EXP/GAMESS (max value)	1.175 Å (2.305 Å)	0.142 Å (0.361 Å)	0.143 Å (0.361 Å)

When all three structures were energy minimised in MAIN until convergence was reached (gradient < 1.0 kcal/mol) using PURY geometric restraints, the RMS deviations dropped drastically, indicating that PURY parameters are self consistent. Also the RMSDs of the energy minimised models superimposed on the experimental structures (0.04 Å, 0.14 Å, 0.08 Å) revealed that conformations of the PURY minimised models remained essentially unchanged.

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

In addition we have optimised the selected structures with the program eLBOW from PHENIX program suite (Adams et al., 2002). ENAMEL and CEPTIA structures delivered deviations of bond and angle terms several times larger than others when compared to PURY and experimental structures. The ABIYUF minimisation, however, failed to run. The large differences of RMS of co-ordinates are again the result of conformational differences.

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

5.2 Reliability parameters for different parts of chemical space

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

Table 5: *Average sigma values.* Average sigma values for bonds and angles in PURY chemical subsets, EH set and PDB hetero molecules derived sets.

	PURY – amino acids	PURY – non-metal	PURY - all	EH	HET – non-metal	HET - all
Bonds	0.027	0.043	0.064	0.022	0.073	0.091
Angles	2.52	2.59	5.66	1.84	5.96	6.36

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

5.3 Validation of PURY restraints in refining macromolecular structures

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

Table 6: Refinement statistics and cross validation of the crystal structure of cathepsin B, **1SP4** (Štern et al., 1999), two structures of beta-lactamases **2Q9M, 2Q9N** (Plantan, 2007) and crystal structure of the SARS-corona virus ORF7A accessory protein, **1XAK** (Nelson et al., 2005). All four structures were refined with the program MAIN using all structure factors in the available resolution span. The crystallographic refinement target was set to 0.01 Å for the RMSD bond deviations. The structures were first distorted with a 0.3 Å kick and then refined against PURY and EH target values until the gradient reached the value of 5 energy units.

PDB ID	**1SP4**	**2Q9M**	**2Q9N**	**1XAK**
Resolution (Å)	2.20	2.05	2.20	1.80
PURY R-factor (R-free)	19.5	21.3 (24.9)	25.4 (29.8)	23.9 (27.5)
EH R-factor (R-free)	19.3	21.2 (24.8)	25.1 (29.7)	23.1 (28.4)
Bonds RMS				
PURY/PURY	0.0112 Å	0.0108 Å	0.0109 Å	0.0093 Å
EH/EH	0.0111 Å	0.0113 Å	0.0110 Å	0.0129 Å
PURY/EH	0.0141 Å	0.0144 Å	0.0133 Å	0.0125 Å
EH/PURY	0.0149 Å	0.0158 Å	0.0114 Å	0.0167 Å
Angles RMS				
PURY/PURY	1.649º	1.744º	1.786º	1.559º
EH/EH	1.578º	1.701º	1.807º	1.744º
PURY/EH	1.656º	1.798º	1.796º	1.676º
EH/PURY	1.858º	1.960º	1.846º	1.904º
coor RMS (PURY/EH)	0.026 Å	0.020 Å	0.005 Å	0.082 Å
max coor RMS (PURY/EH)	0.244 Å	0.196 Å	0.030 Å	0.312 Å

5.4 Detailed comparison with an expert derived parameter set

We have shown that the PURY parameter set performs essentially equivalently to the EH set, however, a direct comparison of individual terms would be informative as it exposes a few limitations and provides hints for future development. The number of PURY classes (30) covering 20 amino acid residues differs from that of the EH classes (35). The translation from one set to the other is not “bidirectional”, since quite often a single EH class is described by a few PURY classes and vice versa (**Table 7)**. Although the comparison of classes is indicative, the true value of such a comparison can only be revealed by comparison of the geometrical parameters they define.

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

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

Table 8 shows direct translation between EH and PURY bond terms with their corresponding average values and sigma values. Only a selected set of comparisons between the two parameter sets follow.

Table 8: *Bond by bond translation of Engh-Huber parameter set into PURY derived equivalents*

EH 1	EH 2	AVG [Å]	SIGMA [Å]	PURY 1	PURY 2	AVG [Å]	SIGMA [Å]
C5W	CW	1.433	0.018	CC_5	C__*	1.442	0.0318
CW	CW	1.409	0.017	C__*	C__*	1.412	0.0417
C	CH1E	1.525	0.021	C__Y	CH1X	1.521	0.0164
C	CH1E	1.525	0.021	C__Y	C51X	1.515	0.0153
C5	CH2E	1.497	0.014	CC_5	CH2X	1.500	0.0184
C5W	CH2E	1.498	0.031	CC_5	CH2X	1.500	0.0184
CF	CH2E	1.502	0.023	CC_6	CH2X	1.506	0.0160
CY	CH2E	1.512	0.022	CC_6	CH2X	1.506	0.0160
C	CH2E	1.516	0.025	C__Y	CH2X	1.505	0.0250
CN	CH2E	1.503	0.019	COO_	CH2X	1.514	0.0157
C	CH2G	1.516	0.018	C__Y	CH2X	1.505	0.0250
C5W	CR1E	1.365	0.025	CC_5	CH_5	1.400	0.0306
CW	CR1E	1.398	0.016	C__*	CH_6	1.401	0.0208
CW	CR1W	1.394	0.021	C__*	CH_6	1.401	0.0208
CF	CR1E	1.384	0.021	CC_6	CH_6	1.389	0.0170
CY	CR1E	1.389	0.021	CC_6	CH_6	1.389	0.0170
CY2	CR1E	1.378	0.024	CO_6	CH_6	1.389	0.0213
C5	CR1H	1.354	0.011	CC_5	CH_5	1.400	0.0306
C5	CR1E	1.356	0.011	CC_5	CH_5	1.400	0.0306
C	N	1.341	0.016	C__Y	N5_Y	1.364	0.0323
C	NC2	1.326	0.018	C__G	NH2G	1.322	0.0164
C5	NH1	1.378	0.011	CC_5	NH_5	1.352	0.0304
CW	NH1	1.370	0.011	C__*	NH_5	1.376	0.0239
C	NH1	1.329	0.014	C__G	NH1G	1.341	0.0271
C	NH1	1.329	0.014	C__Y	NH1Y	1.327	0.0335
C	NH2	1.328	0.021	C__Y	NH2Y	1.320	0.0230
C5	NR	1.371	0.017	CC_5	N__5	1.340	0.0307
C	O	1.231	0.020	C__Y	O2C_	1.222	0.0293
CN	O	1.208	0.023	COO_	O2C_	1.215	0.0182
C	OC	1.249	0.019	COO_	O-1_	1.255	0.0242
CY2	OH1	1.376	0.021	CO_6	OH1<	1.355	0.0210
C	OH1	1.304	0.022	COO_	OH1<	1.306	0.0192
CH1E	CH1E	1.540	0.027	CH1X	CH1X	1.534	0.0219
CH1E	CH2E	1.530	0.020	CH1X	CH2X	1.522	0.0256
CH1E	CH3E	1.521	0.033	CH1X	CH3X	1.518	0.0297
CH1E	N	1.466	0.015	C51X	N5_Y	1.471	0.0154
CH1E	NH1	1.458	0.019	CH1X	NH1Y	1.468	0.0206
CH1E	NH3	1.491	0.021	CH1X	NH3X	1.488	0.0186
CH1E	OH1	1.433	0.016	CH1X	OH1<	1.425	0.0198
CH2E	CH2E	1.520	0.030	CH2X	CH2X	1.507	0.0398
CH2P	CH2E	1.492	0.050	C52X	C52X	1.491	0.0513
CH2P	CH2P	1.503	0.034	C52X	C52X	1.491	0.0513
CH2E	CH3E	1.513	0.039	CH2X	CH3X	1.494	0.0613
CH2P	N	1.473	0.014	C52X	N5_Y	1.465	0.0284
CH2G	NH1	1.451	0.016	CH2X	NH1Y	1.464	0.0305
CH2E	NH1	1.460	0.018	CH2X	NH1G	1.461	0.0118
CH3E	NH1	1.460	0.018	CH3X	NH1Y	1.457	0.0244
CH2E	NH3	1.489	0.030	CH2X	NH3X	1.480	0.0233
CH2G	NH3	1.489	0.030	CH2X	NH3X	1.480	0.0233
CH2E	OH1	1.417	0.020	CH2X	OH1<	1.421	0.0265
CH2E	S	1.822	0.020	CH2X	S__<	1.818	0.0190
CH2E	SM	1.803	0.034	CH2X	S__<	1.818	0.0190
CH2E	SH1E	1.808	0.033	CH2X	SH1<	1.828	0.0180
CH3E	SM	1.791	0.059	CH3X	S__<	1.796	0.0210
CR1E	CR1E	1.382	0.030	CH_6	CH_6	1.379	0.0202
CR1E	CR1W	1.400	0.025	CH_6	CH_6	1.379	0.0202
CR1W	CR1W	1.368	0.019	CH_6	CH_6	1.379	0.0202
CR1E	NH1	1.374	0.021	CH_5	NH_5	1.349	0.0252
CRH	NH1	1.345	0.020	CH_5	NH_5	1.349	0.0252
CRHH	NH1	1.321	0.010	CH_5	NH_5	1.349	0.0252
CR1H	NH1	1.374	0.011	CH_5	NH_5	1.349	0.0252
CR1E	NR	1.382	0.030	CH_5	N__5	1.338	0.0296
S	S	2.030	0.016	S__<	S__<	2.050	0.0468
CRH	NR	1.319	0.013	CH_5	N__5	1.338	0.0296

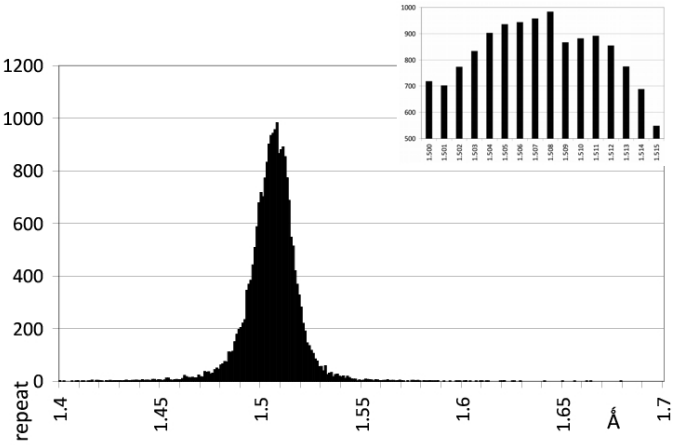
5.4.1 Tyrosine – phenylalanine CG atoms

PURY uses the same atom class "CC_6" to describe the "CG" atom of the phenylalanine and tyrosine residues, whereas EH uses two classes, "CF" and "CY". A consequence of this is that, in the EH set, there are two different parameters for the "CB - CG" atoms bond which describe the single bond by which the phenyl ring is attached to the alanine base, and two for the "CG – CD" atom bonds which describe the bond in the aromatic ring and the corresponding angles.

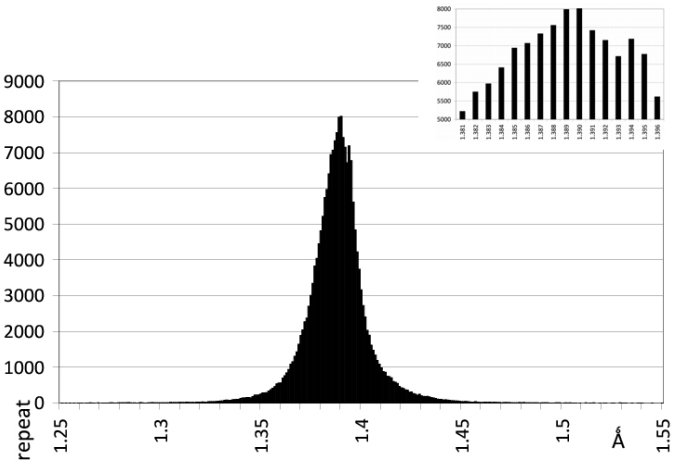
The PURY mean bond value for "CB - CG" is 1.508 Å (0.019 Å), which is between the EH target values 1.502 Å (0.023 Å) and 1.512 Å (0.022 Å) for PHE and TYR residues respectively. Both EH values lie within the 1 sigma range of the PURY target – the differences are –0.006 Å and +0.004 Å. Also the 0.01 Å difference between the two EH values lies well within 1 sigma. The histogram of bond lengths shown in **Figure 17a** indicates the presence of a double peak in the "CC_6 - CH2X" bonds corresponding to the EH target values. The histogram of distances between "CG - CD" bonds shown in **Figure 17b** also shows a double peak. Interestingly, although only one peak matches the EH values (1.389 Å) corresponding to the TYR bond, the other peak appears higher (1.394 Å) and not lower than the EH value for phenylalanine bond (1.384 Å), suggesting that the target for this bond length restraint should be re-evaluated.

In contrast, the histograms for angles shown in **Figures 17c** and **Figure 17d,** corresponding to the phenylalanine and tyrosine "CB - CG - CD(1,2)", "CD1 - CG - CD2" angle value distributions, exhibit no double peaks. The EH parameters for the "CB - CG - CD (1, 2)" angle differ by 0.5º, and the PURY angle is in the middle of the two. Interestingly, the histogram of "CH_6 - CC_6 - CH_6" (**Figure 17d**) reveals the presence of a peak at exactly 120º, probably an indication of the constrained angle value used during refinement. The absence of this peak on the "CB - CG - CD (1, 2)" histogram (**Figure 17c**) is probably not apparent since the restraint and the mean are equal.

a



b



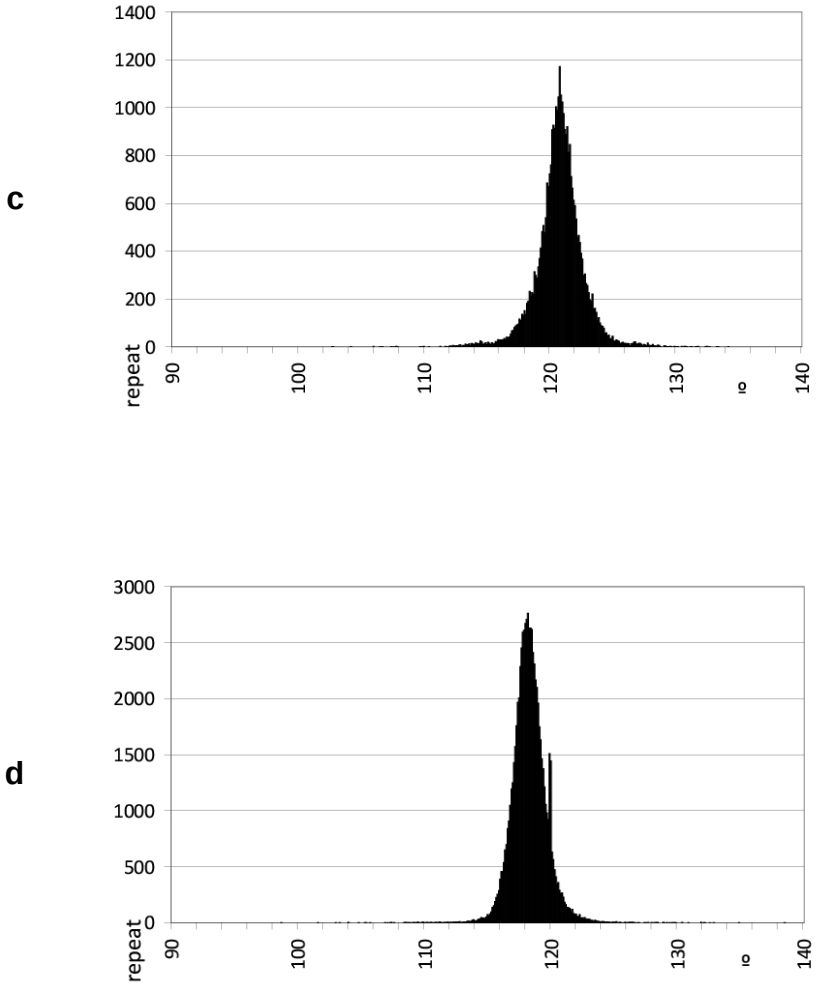


Figure 17: *Histograms of bond and bond angle parameters involving CG atom of phenyl alanine and tyrosine residues.* a) A CB - CG bond, which describes the single bond by which phenyl ring is attached to the alanine base **"CH2X - CC_6"**. Insert shows zoomed in peak. b) A CG - CD bond, which describes the bond in the aromatic ring **"CC_6 - CH_6"**. Insert shows zoomed in peak. c) A CB - CG - CD angle, which describes the angle by which phenyl ring is attached to the alanine base **"CH2X - CC_6 - CH_6"**. d) A CD1 - CG - CD2 bond, which describes the angle in the aromatic ring **"CH_6 - CC_6 - CH_6"**.

This comparison thus questions the need for the two different classes for the “CG” atom of PHE and TYR residues in refinement of macromolecular structures.

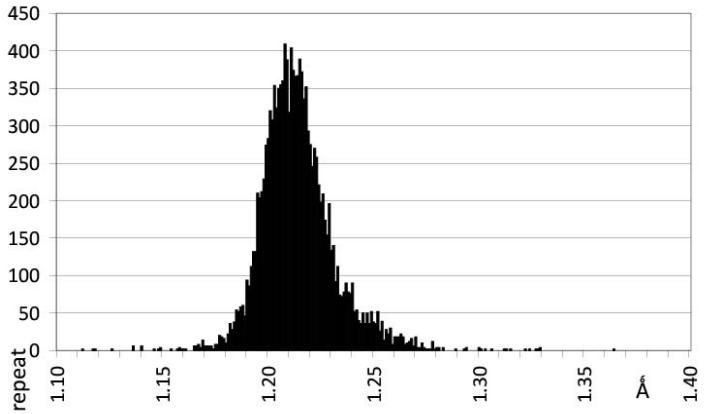
5.4.2 Carboxylic group

The EH parameter set uses two different atom classes for designation of carboxylic acid group carbon atom. For charged carboxylic group they provide **"C"** atom class which is the same atom class as used in peptide main chain. Other atom class is **"CN"** which is provided for uncharged carboxylic acid group. The corresponding bond involving above mentioned atom classes are carbon - oxygen bond designated as **"C - OC",** with average bond distance of 1.249 Å corresponding to charged carboxylic group, carbon - oxygen double bond designated as **"C - O",** with average bond distance of 1.231 Å corresponding to main chain peptide bond, carbon - oxygen single bond designated as **"C - OH1",** with average bond distance of 1.304 Å corresponding to neutral carboxylic group and carbon - oxygen double bond designated as **"CN - O",** with average bond distance of 1.208 Å corresponding to neutral carboxylic group. The EH parameter does not provide corresponding carbon - oxygen single bond termed **"CN - OH1"** designed for neutral carboxylic group. The "CN - O" term is describing pure double carbon - oxygen bond for neutral carboxylic

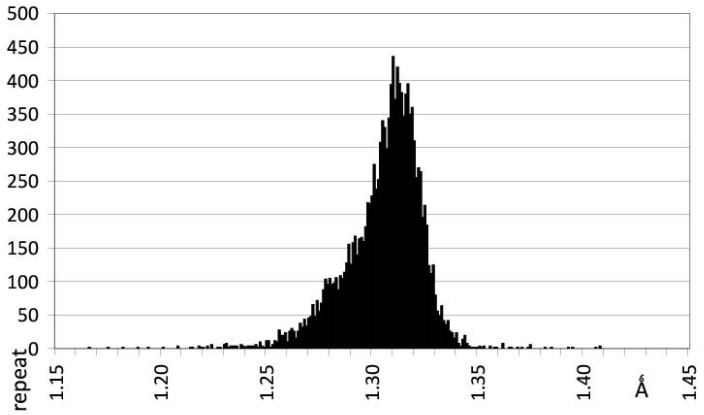
group while length of "C -O" bond term is elongated due to rigidity and partial double-bond characteristic of peptide bond. In PURY parameter set, three bonding parameters are used to describe the charged and neutral carboxylic groups of the GLU and ASP residues: “COO_ - O-1_” (1.255 Å) for charged and “COO_ - OH1<” (1.306 Å) and “COO_ - O2C_” (1.215 Å) for neutral carboxylic groups while separate atom class and bond terms are used for main chain description.

The histogram of the bond distances "COO_ - O-1_" is symmetrical (**Figure 18a**), suggesting that two equivalent oxygen atoms are bonded to the carbon atom. Its average value of 1.255 Å comes right in the middle of the averages for the double (1.215 Å) and single (1.306 Å) bond distances of the neutral carboxylic group (**Figure 18b and 18c**). Both angle terms shown in **Figures 18d and 18e** show single peaks with equivalent target values, suggesting that, for the angle parameters, there is no noticeable difference between the neutral and charged carboxylic group.

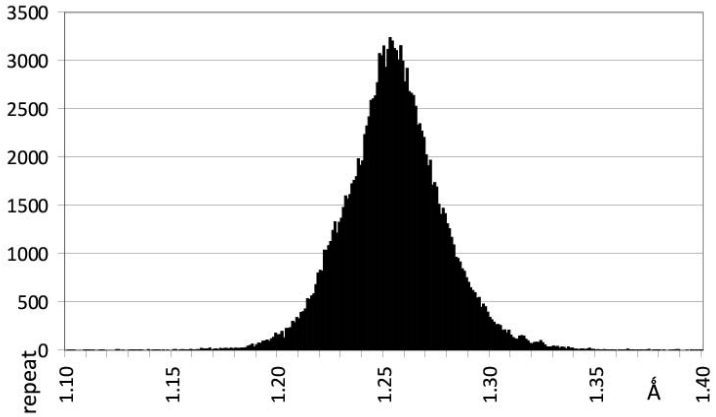
a



b



c



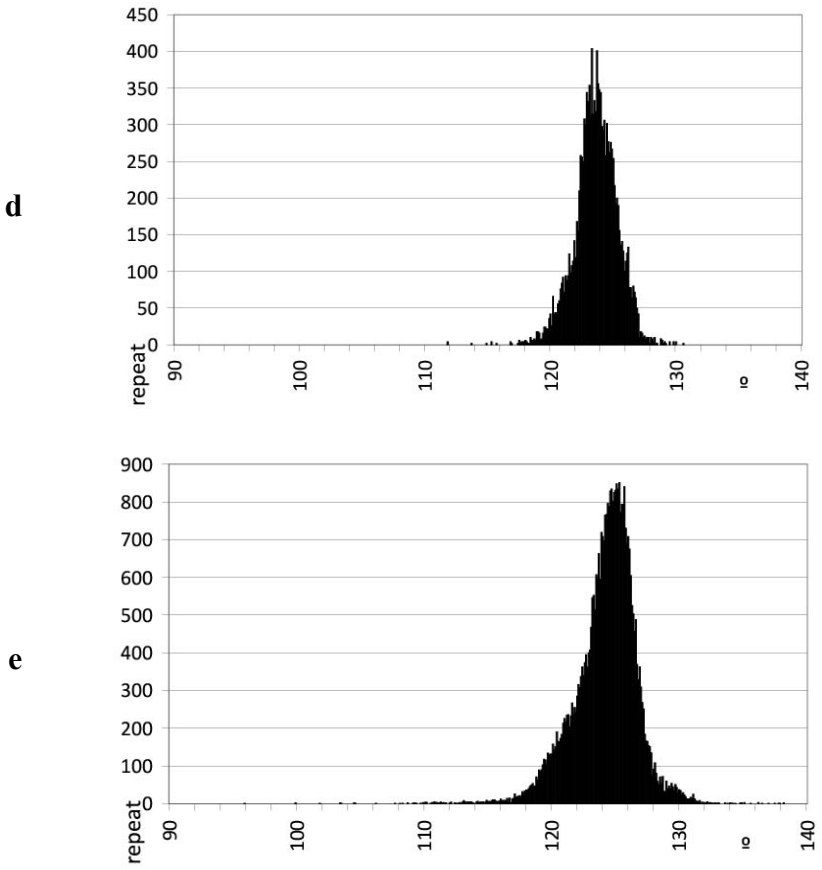


Figure 18: *Histograms of bond and bond angle parameters involved in the carboxylic group.* a) A double bond between a sp2-hybridised carbon **"COO_"** atom and a sp2-oxygen atom **"O2C_"**. b) A single bond between a sp2-hybridised carbon atom **"COO_"** and a sp3-oxygen atom with one bonded hydrogen atom **"OH1<"**. c) A single bond between a sp2-hybridised carbon atom **"COO_"** and a sp3-oxygen atom without explicitly bonded hydrogen atom and with partial charge distribution in carboxylic group **"O-1<"**. d) An angle describing carboxylic group with explicitly defined hydrogen atoms **"O2C_ - COO_ - OH1<"**. e) An angle describing carboxylic group without explicitly defined hydrogen atoms and with partial charge **"O-1< - COO_ - O-1<"**.

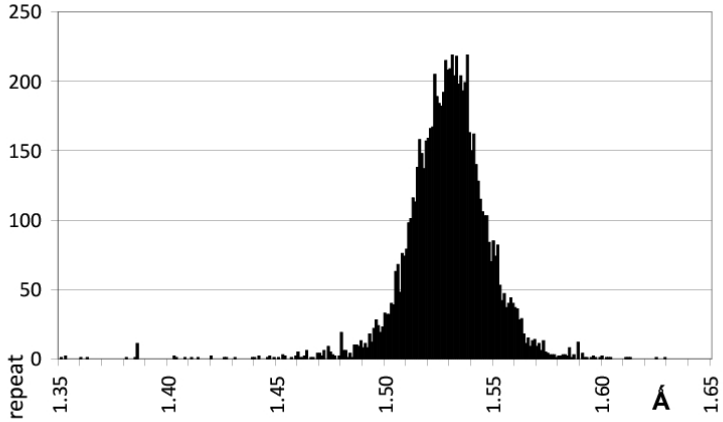
The analysis is based on the assumption that the structural data are correct – that the hydrogen atoms are present when the carboxylic group is not charged. The broad bottom of the charged group bonding distance (**Figure 18a**) and non-symmetrical distributions of the bond distances of the neutral groups (**Figure 18b and 18c**), however, leave the impression that not all carboxylic groups are necessarily correctly assigned. Nevertheless the histograms suggest that present atom classes and corresponding parameters for neutral carboxylic groups should be extended in the standard EH parameter set for use in the refinement of protein structures.

5.4.3 Proline

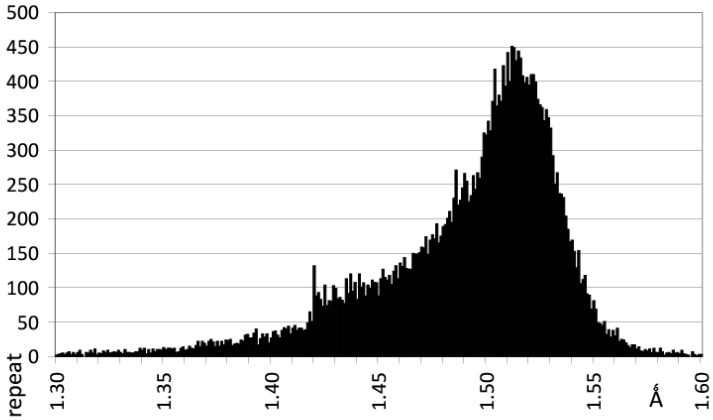
The PURY bond length for "CA - CB" bond in the proline residue ("C51X - C52X", average = 1.529 Å, sigma = 0.021 Å) is broader than the one used in the EH set, where it is kept constant for all amino acids (average = 1.530 Å, sigma= 0.020 Å), however PURY differentiates between the proline ring "CA - CB" atoms ("C51X - C52X") and the non-ring "CA - CB" atoms ("CH1X - CH2X"). As already noted (Engh and Huber, 2001), the proline "CB - CG" bond with an average 1.492 Å has a large sigma value (sigma = 0.05 Å). Also PURY analysis delivered similar values (average = 1.491 Å, sigma = 0.051 Å). However, the histogram of "C52X - C52X" bonds between sp3 atoms in a 5 member ring is skewed with the peak at 1.518 Å which lies away from the average value. The broadening and the skewed shape indicate that the variability is a result of ring puckering, during which the two atoms approach each other (**Figures 19a and 19b).**

The double peak of the bonding angle is reminiscent of the proline residue analysis, where coupling was observed between the bonding and dihedral angles. The current PURY approach cannot differentiate between cis and trans prolines, as noted by Engh and Huber [Engh and Huber, 2001], suggesting that an expert parameter set performs better. The bond length assignment is actually a problem of the concept of atom class assignment based on chemical environment, which cannot differentiate between geometric arrangements such as cis and trans peptide bonds. A simple solution to this problem is to use different topology library entries in combination with different atom classes for describing prolines in their cis and trans conformations.

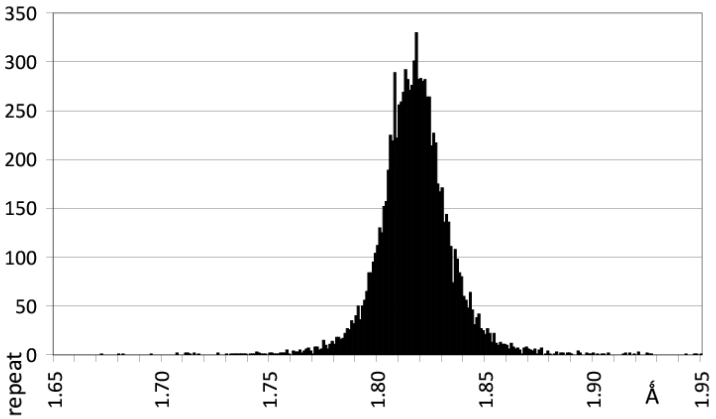
a



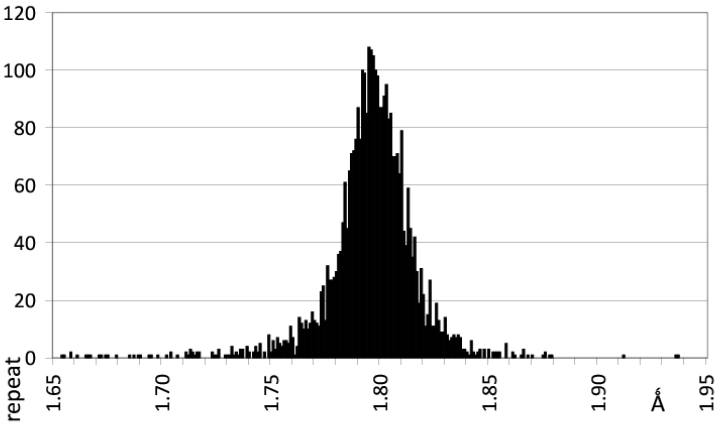
b



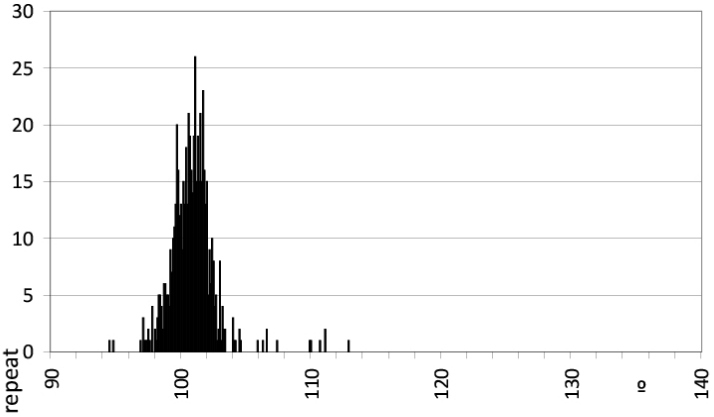
c


350
300
250
200
150
100
50
0
repeat
1.65
1.70
1.75
1.80
1.85
1.90
Å
1.95

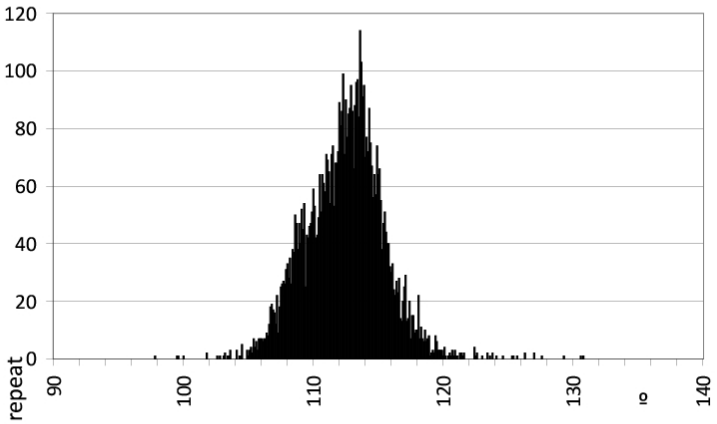
d


120
100
80
60
40
20
0
repeat
1.65
1.70
1.75
1.80
1.85
1.90
Å
1.95

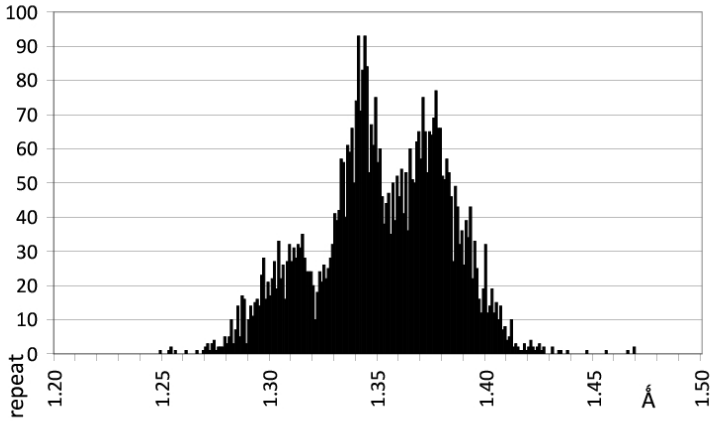
e


30
25
20
15
10
5
0
repeat
90
100
110
120
130
º
140

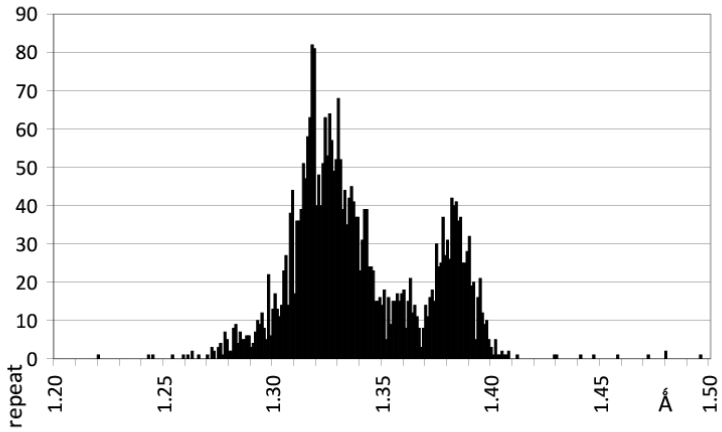
f


120
100
80
60
40
20
0
repeat
90
100
110
120
130
140

g


100
90
80
70
60
50
40
30
20
10
0
repeat
1.20
1.25
1.30
1.35
1.40
1.45
Å
1.50

h


90
80
70
60
50
40
30
20
10
0
repeat
1.20
1.25
1.30
1.35
1.40
1.45
Å
1.50

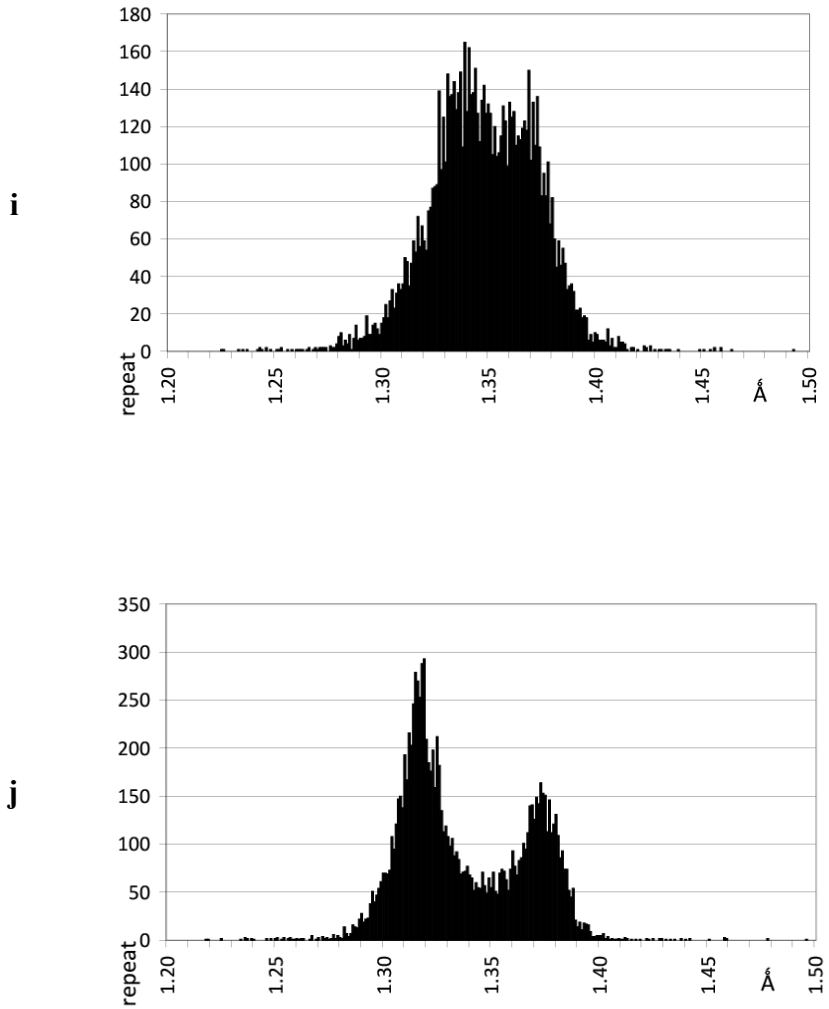


Figure 19: *Histogram distributions of bond and bond angle parameters involved various amino acid residues*. a) proline CA – CB **"C51X - C52X"**, b) proline CB – CG **"C52X - C52X"**, c) methionine CG – SD **"CH2X - S__<"**, d) methionine SD – CE **"S__< - CH3X"**, e) methionine CG - SD – CE **"CH2X - S__< - CH3X"**, f) methionine CB - CG – SD **"CH2X -CH2X -S__<"**, g) histidine CG – ND1 **"CC_5 - NH_5"**, h) histidine NE2 – CG **"CC_5 - N__5"**, i) histidine CE1 – NE2 **"CH_5 - NH_5"**, j) histidine CE1 - ND1 **"CH_5 - N__5"**.

5.4.4 Methionine

Parameters involving the sulphur atom in methionine EH and PURY parameters differ. The EH "CG - SD" bond target with 1.803 Å is 0.015 Å shorter (almost 1 sigma 0.019) than the PURY value (1.819 Å), whereas the SD – CE are much more alike – 1.791 Å and 1.796 Å for EH and PURY respectively. The parameters for angles CB-CG-SD and CG-SD-CE are also similar. Interestingly though, the PURY force constants are evidently higher than those for EH in all terms except the CG-CD-CE angle, where they are approximately equal. The bond and angle histograms involving methionine parameters are shown in **Figures 19c, 19d, 19e and 19f**. Due to the larger number of structures used in PURY analysis (when compared to the presumably smaller number used to derive EH targets) we suggest modifying the geometric restraints (in particular forces) for the methionine residue. In addition PURY parameters for restraints of the seleno-methionine residues are provided in **Table 9**.

Table 9: *Comparison of methionine geometric parameters from PURY and EH sets*. Atom classes involved in bonds, angles and dihedral angles describing side chain of methionine and corresponding average values and force constants are shown. Specific parameters for the seleno-methionine residue are shown below.

Entry	Class1	Class2	Class3	Class4	Force constant	Multiplicity	Average value
EH methionine specific parameters							
Bond	CH2E	SM			512.111		1.803
Bond	CH3E	SM			170.066		1.791
Angle	CH2E	SM	CH3E		401.534		100.900
Angle	CH2E	CH2E	SM		215.936		112.700
Dihedral	X	CH2E	SM	X	3.60	3	0.000
Dihedral	X	CH2E	CH2E	X	4.80	3	180.000
PURY methionine specific parameters							
Bond	CH2X	S__<			1643.603		1.818
Bond	CH3X	S__<			1324.387		1.796
Angle	CH2X	S__<	CH3X		733.833		100.904
Angle	CH2X	CH2X	S__<		217.735		112.487
Dihedral	CH2X	CH2X	S___<	CH3X	7.41	3	60.000
Dihedral	CH2X	CH2X	CH2X	S__<	4.71	3	60.000
PURY Se-methionine specific parameters							
Bond	CH2X	Se_<			2492.665		1.960
Bond	CH3X	Se_<			502.996		1.943
Angle	CH2X	Se_<	CH3X		1070.409		97.976
Angle	CH2X	CH2X	Se_<		342.319		113.030
Dihedral	CH2X	CH2X	Se__<	CH3X	7.27	3	60.000
Dihedral	CH2X	CH2X	CH2X	Se_<	4.72	3	60.000

5.4.5 Histidine

The PURY and EH parameter sets differ most in the histidine residue terms. Both atom class assignments differentiate between protonated and non-protonated ND1 and NE2 atoms, however, PURY atom class assignment does not differentiate between CD2 and CE1 atoms, which are both recognized as "CH_5" (carbon atom within a five member planar ring with bonded hydrogen atom). As a consequence, four different bonding targets of the EH set merge into one PURY target, which lies somewhere in the middle. The situation is similar with the bond length of non protonated histidines, which merge into a single PURY target value. The longer and shorter bonding distances are an indication of single and double bond character of the ring bonds and cannot be appropriately elaborated with the current PURY concept (**Figures 19g-j**). Namely, only a single target value can be derived for bonds between two atom classes, yet the bonds between two atom classes can be single or double. For example, in a putative case in which sp2 hybrid atoms are bonded to each other in a chain, single and double bonds alternate, yet they are both bonds between the same atom classes. Clearly, the chance that such a case occurs within a small set is rather small, however, when a large pool of data such as CSD has been analysed, the occurrence of such cases is not that uncommon.

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

6 Conclusions

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

The analysis, in particular the multiple maxima and non-symmetrical histograms of geometrical terms, has exposed two essential questions: First, is the current atom class assignment scheme indeed recognizing all appropriate atom classes (which important issues have we missed)? Second, how reliable are the data presented in CSD? We hope that the use of PURY and the coming validation of all structures deposited in CSD against the PURY database will expose further potential miss assignments while the use of PURY for validation of small molecule structures may draw more attention to details of the structures such as charged states and protonation and thereby increase the reliability of the data being deposited. Unfortunately, the lack of the structure factors prevents remediation of the structures already in CSD.

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

7 Acknowledgements

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

8 References

Adams, P. D., Pannu, N. S., Read, R. J. & Brunger, A. T. (1997). Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement, Proc. Natl Acad. Sci. USA, **94**, 5018-5023.

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K. and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Cryst., D58, 1948-1954.

Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., et al (1979). The Cambridge Crystallographic Data Centre: computer-based search, retrival, analysis and display of information. Acta Cryst. B35, 2331-2339.

Allen, F.H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst*. B**58**, 380-388.

Allinger, N. (1977): Conformational Analysis 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms .J. Am. Chem. Soc. **99**, 8127-8134

Amemiya Y., Matsushita T., Nakagawa A., Satow Y., Miyahara J. and Chikawa J.-I. (1988). Design and performance of an imaging plate system for X-ray diffraction study. Nucl. Instrum. Methods Phys. Res. A. **266**, 645-653.

Barry C.D. and McAlister J.P. (1982). Computational crystallography. Ed. D. Sayre. Oxford. Clarendon Press. 274.

Bergfors T. (1999) Protein Crystallization Techniques, Strategies, and Tips: A Laboratory Manual, International University Line, La Jolla USA

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235-242.

Bernstein, F.C., T.F. Koetzle, G.J. Williams, E.F. Meyer Jr, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112: 535.

Blow D.M. and Rossmann, M.G. (1961): The single isomorphous replacement method. Acta Cryst **14** 1195-1202

Blundell TL, Cutfield JF, Cutfield SM, Dodson EJ, Dodson GG, Hodgkin DC, Mercola DA, Vijayan M. (1971) Atomic Positions in Rhombohedral 2-Zinc Insulin Crystals. *Nature* **231**(5304): 506-511.

Booth A.D. (1946a). A differential Fourier method for refining atomic parameters in crystal structure analysis. Trans. Faraday Soc. 42, 444-448.

Booth A.D. (1946b). The simultaneous differential refinement of co-ordinates and phase angles in X-ray Fourier synthesis. Trans. Faraday Soc. 42, 617-619.

Bragg W.L., 1913, *The Diffraction of Short Electromagnetic Waves by a Crystal*, *Proc. Cambridge Phil. Soc.*, **17**, 43-57.

Bricogne G. and Irwin J.J. (1996) Proceedings of the CCP4 study weekend. Macromolecular refinement. Ed. E. Dodson, M. Moore, A. Ralph and S. Bailey. 85-92. Warrington.

Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M. (1983): *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem. **4**, 197-217

Brünger, A.T., Kuryan, J. and Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. Science, 235, 458-460.

Brunger A.T. (1992). The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature **255**, 472-474.

Brunger, A. T. & Adams, P. D. (2002). Molecular dynamics applied to X-ray structure refinement. Acc. Chem. Res. **35**, 404-412.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszevski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography and NMR System: a new software suite for macromolecular structure determination. *Acta Cryst.*, D**54**, 905-921.

Bruno, I.J., Cole, J.C., Edgington, P.R., Kessler, M., Macrae, C.F., McCabe, P., Pearson, J. and Taylor, R. (2002). New Software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Cryst.* B**58**, 389-397.

Busing W.R., Martin K.O. and Levy H.A. (1962). Report ORNL-TM-305. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

Connolly M.L. and Olson A.J. (1985). GRANNY, a companion to GRAMPS for the real-time manipulation of macromolecular models. J. Comput. Chem. 9, 1-6.

CORINA, Molecular Networks GmbH Computerchemie, Erlangen, Germany (http://www.mol-net.de)

Dauter Z., Lamzin V.S. and Wilson K.S. (1997). The Benefits of Atomic Resolution. *Curr Opin Struct Biol* **7**, 681-688.

Diamond R. (1971). A real-space refinement procedure for proteins. Acta Cryst. A**27**, 436-452.

Diamond R. (1982). Computational crystallography. Ed. D. Sayre. Oxford. Clarendon Press. 318-325.

Dixon, S. L. & Merz, K. M. (1996). Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.* **104**, 6643-6649.

Dixon, S. L. & Merz, K. M. (1997). Fast, accurate semiempirical molecular orbital calculations for

macromolecules. *J. Chem. Phys.* **107**, 879-893.

Eikenberry E.F., Tate M.W., Bilderback D.H. and Gruner S.M. (1992). Inst. Phys. Conf. Ser. **121**, 273-280.

Engh, R.A. and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta. Cryst.*, A**47**, 392-400.

Engh, R.A. and Huber, R. (2001). Structure quality and target parameters. *International Tables for Crystallography, Volume F: Crystallography of Biological Macromolecules*, edited by M.G. Rossman and E. Arnold. pp. 382-392. Dordrecht, The Netherlands: Kluwer.

Ertl, P. JME - Java Molecular Editor, Novartis (http://www.molinspiration.com/jme/)

Eyck L.F.T. and Watenpaugh K.D. (2001). *International Tables for Crystallography, Volume F: Crystallography of Biological Macromolecules*, edited by M.G. Rossman and E. Arnold. pp. 369-374. Dordrecht, The Netherlands: Kluwer.

Ferrin T.E., Huang C.C., Jarvis L.E. and Langridge R. (1988). The MIDAS display system. J. Mol. Graphics **6**, 13-27.

Garman E.F. and Mitchell E.P. (1996). Glycerol concentrations required for cryoprotection of 50 typical protein crystallization solutions. J. Appl. Cryst. **29**, 584-587.

Giege R., Moras D. and Thierry J.-C. (1977). Yeast transfer RNAAsp: A new high-resolution X-ray diffracting crystal form of a transfer RNA. J. Mol. Biol. **115**, 91-96.

Gordon, M.S. and Schmidt, M.W. (2005). Advances in electronic structure theory: GAMESS a decade later. *Theory and Applications of Computational Chemistry: the first forty years*, edited by C.E. Dykstra, G. Frenking, K.S. Kim and G.E. Scuseria, pp. 1167-1189. Amsterdam; Elsevier.

Gotz G., Meszaros E. and Vali G. (1991). Atmospheric particles and nuclei. Budapest, Akademiai Kiado, 142.

Greaves, R.B., Vagin, A.A. and Dodson, E.J. (1999). Automated production of small-molecule dictionaries for use in crystallographic refinements. *Acta Cryst.* D**55**, 1335-1339.

Hampel A., Labananskas M., Conners P.G., Kirkegard L., Raj Bhandary U.L., Sigler P.B. and Bock R.M. (1968). Single Crystals of Transfer RNA from Formylmethionine and Phenylalanine Transfer RNA's. Science **162**. 1384-1386.

Hendrickson W.A. (1985). Stereochemically Restrained Refinement of Macromolecular Structures. Methods Enzymol. **115**, 252-270.

Hendrickson W.A., Horton J.R. and LeMaster D.M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J. 9, 1665-1672.

Hendrickson, W.A. (1991): Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science. **254** 51-58.

Hodgkin D.C. (1935). X-Ray Single Crystal Photographs of Insulin. Nature 135: 591–592.

Hodgkin D.C., Kamper J., Mackay M. and Pickworth J. (1956) Structure of Vitamin B12 . Nature, **178**, 64

Hope H., Frolow F. von Bohlen K., Makowski I., Kratky C., Halfon Y., Danz H., Webster P., Bartels K.S., Wiottmann H.G. and Yonath A. (1989). Cryocrystallography of ribosomal particles. Acta Cryst B**45**, 190-199.

Hope H. (1990). Crystallography of Biological Macromolecules at Ultra-Low Temperature. Annu. Rev. Biophys. Biophys. Chem. **19**, 107-126.

Hoppe W. (1957). Abstracts of Papers. Acta Cryst **10**, 750-751.

Hoppe W. and Gassmann J. (1968). Phase correction, a new method to solve partially known structures. Acta Cryst. B**24**, 97-107.

Hubbard R.E. (1986). Computer graphics and molecular modeling. Ed. R. Fletterick and M. Zoller. Cold Spring Harbor Press. 9-12.

Hughes E.W. (1941) The Crystal Structure of Melamine. J. Am. Chem. Soc. **63,** 1737-1752.

Hughes S.H. and Stock A.M. (2001) *International Tables for Crystallography, Volume F: Crystallography of Biological Macromolecules*, edited by M.G. Rossman and E. Arnold. pp. 65-79. Dordrecht, The Netherlands: Kluwer.

Jack, A. & Levitt, M.(1978). Refinement of large structures by simultaneous minimization of energy and *R* factor. Acta Cryst. A**34**, 931-935.

Jaskolski, M,. Gilski, M., Dauter, Z. and Wlodawer, A. (2007). Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta Cryst.*, D**63,** 611-620.

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin. Proc. Natl Acad. Sci. USA, **97**, 3171-3176.

Jerusalmi D and Steitz T.A. (1997). Use of organic cosmotropic solutes to crystallize flexible proteins: application to T7 RNA polymerase and its complex with the inhibitor T7 lysozyme. J. Mol. Biol. 274, 748-756.

Jones T.A. (1978). A graphics model building and refinement system for macromolecules. J. Appl. Cryst. **11**, 268-272.

Jullien M, Crosio M.P. and Baudet-Nessler S. (1994). Evidence for a dimeric intermediate on the crystallization pathway of ribonuclease A. Acta Cryst D**50**, 398-403.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. (1958): A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis . Nature **181**(4610): 662-626.

Kleywegt, G.J. and Jones, T.A. (1998). Databases in Protein Crystallography. Acta Cryst., D54, 1119-1131

Kleywegt, G.J., Henrick, K., Dodson, E.J. and van Aalten, D.M.F. (2003). Pound-Wise but Penny-Foolish: How Well Do Micromolecules Fare in Macromolecular Refinement? *Structure*, **11**, 1051-1059.

Kleywegt, G.J. (2007). Crystallographic refinement of ligand complexes. *Acta Cryst*., D**63**, 94-100

Konnert J.H. (1976). A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. Acta Cryst A**32**, 614-617.

Konnert J.H. and Hendrickson W.A. (1980) A restrained-parameter thermal-factor refinement procedure. Acta Cryst A**36**, 344-350.

Lamzin, V.S., Dauter, Z. and Wilson, K.S. (1995). Dictionary of Protein Stereochemistry. *J. Appl. Cryst.*, **28**, 338-340.

Lee, T. S., York, D. M. & Yang, W. T. (1996). Linear-scaling semiemperical quantum calculations for macromolecules. J. Chem. Phys. **105**, 2744-2750.

Loewenstein J.E. and Cohen A.I. (1964). Dry mass, lipid content and protein content of the intact and zona-free mouse ovum. J. Embryol. Exp. Morph. Vol **12**, Part 1, 113-121.

Lyle H.J. (1985): Overview of refinement in macromolecular structure analysis. Methods in Enzymology 115 Ed Wyckoff H.W., Hirs C.H. and Timasheff S.N. Academic Press, 227-234

Macrae, C.F., Edgington, P.R., McCabe, P., Pidcock, E., Shields, G.P., Tazlor, R., Towler, M. and van de Streek, J. (2006). *Mercury*: visualization and analysis of crystal structures. *J. Appl. Cryst.*, **39**, 453-457.

MacKerell AD; Bashford D; Bellott M; Dunbrack RL; Evanseck JD; Field MJ; Fischer S; Gao J; Guo H; Ha S; Joseph-McCarthy D; Kuchnir L; Kuczera K; Lau FTK; Mattos C; Michnick S; Ngo T; Nguyen DT; Prodhom B; Reiher WE; Roux B; Schlenkrich M; Smith JC; Stote R; Straub J; Watanabe M; Wiorkiewicz-Kuczera J; Yin D; Karplus M (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, **102**, 3586-3616.

Muirhead H, Perutz MF. (1963) Structure Of Haemoglobin: A Three-Dimensional Fourier Synthesis of Reduced Human Haemoglobin at 5.5 Å Resolution. Nature (Aug 17) **199**: 633-638.

Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.*, D**53**, 240-255.

Nelson, C.A., Pekosz, A., Lee, C.A., Diamond, M.S. and Fremont, D.H. (2005). Structure and Intracellular Targeting of the SARS-Coronavirus Orf7a Accessory Protein. Structure . **13**. 75-85.

Nilsson, K., Lecerof, D., Sigfridsson, E. and Ryde, U. (2003). An automatic method to generate force-field parameters for hetero-compounds. *Acta Cryst,*. D**59,** 274-289

O'Donnell T.J. and Olson A.J. (1981). GRAMPS - A graphics language interpreter for real-time, interactive, three-dimensional picture editing and animation. Comput. Graphics **15**. 133-142.

Otwinowski Z. and Minor W., " Processing of X-ray Diffraction Data Collected in Oscillation Mode ", *Methods in Enzymology*, Volume **276**: Macromolecular Crystallography, part A, p.307-326, 1997,C.W. Carter, Jr. & R. M. Sweet, Eds., Academic Press (New York).

Parkinson, G., Vojtechovsky, J., Clowney, L., Brunger, A.T. and Berman, H.M. (1996). New Parameters for the Refinement of Nucleic Acid Containing Structures.*Acta Cryst.*, D**52**, 57-64.

Pannu, N. S. & Read, R. J. (1996). Improved structure refinement through maximum likelihood. *Acta Cryst.* A**52**, 659-668.

Pearlman, D,. Case, D., Caldwell, J., Ross, W., Cheatham, T. III, DeBolt, S., Ferguson, D., Seibel, G. and Kollman, P. (1995): AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput. Phys. Commun. **91**, 1-41.

Petsko G.A. (1975). Protein crystallography at sub-zero temperatures: Cryo-protective mother liquors for protein crystals. J. Mol. Biol. 96, 381-392.

Phillips W.C. (1985). X-ray sources. Methods Enzymol. 114. 300-316.

Plantan, I., Selič, L., Mesar, T., Štefanič, P., Oblak, M., Preželj, A., Hesse, L., Andrejašič, M., Vilar, M., Turk, D., Kocijan, A., Prevec, T., Vilfan, G., Kocjan, D., Čopar, A., Urleb, U. and Šolmajer, T.. (2007). 4-substituted trinems as broad spectrum [beta]-lactamase inhibitors. J. med. chem., vol. 50, 17, 4113-4121.

Read, R. J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst.* A**42**, 140-149.

Read, R. J. (1990). Structure-factor probabilities for related structures. *Acta Cryst.* A**46**, 900-912.

Richard B., Bonnete F., Dym O. and Zaccai G. (1995). Archaea, a laboratory manual. Cold Spring Harbour Laboratory Press, 149-154.

Rossmann M.G. and Blow D.M. (1962). The detection of subunits within the crystallographic asymmetric unit. Acta Cryst **15**, 24-31.

Rossmann M.G. (1972) *The molecular replacement method.* Gordon & Breach . New York.

Rossman, M.G. (2001): Historical Background. *Internation Tables for Crystallography, Volume F: Crystallography of Biological Macromolecules*, edited by M.G. Rossman and E. Arnold. pp. 4-9. Dordrecht, The Netherlands: Kluwer.

Ryde, U. & Nilsson, K. (2003a). Quantum refinement - a combination of quantum chemistry and protein crystallography. J. Mol. Struct. (Theochem), 632, 259-275.

Ryde, U. & Nilsson, K. (2003b). Quantum chemistry can improve protein crystal structures locally. J. Am. Chem. Soc. **125**, 14232-14233.

Ryde, U. & Nilsson, K. (2003c). Quantum refinement — a method to determine protonation and oxidation states of metal sites in protein crystal structures. J. Inorg. Biochem. **96**, 39.

Ryde, U., Olsen, L. & Nilsson, K. (2002). Quantum chemical geometry optimisations in proteins using crystallographic raw data. J. Comput. Chem. **23**, 1058-1070.

Salemme F.R., Genieser L. Finzel B.C., Hilmer R.M. and Wendoloski J.J. (1988). Molecular factors stabilizing protein crystals. J. Cryst. Growth 9**0**, 273-282.

Sanger F and Tuppy H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates Biochem J. **49,** 463–481.

Schüttelkopf, A.W. and van Aalten, D.M.F. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Cryst., D60, 1355-1363.

Sheldrick, G.M. and Schneider, T.R. (1997). SHELXL: high-resolution refinement. *Methods Enzymol.*, **277**, 319-343.

Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S., Windus, T.L., Dupuis, M. and Montgomery, J.A. (1993), General Atomic and Molecular Electronic Structure System *J. Comput. Chem.*, **14**, 1347-1363.

Štern, I., Schaschke, N., Moroder, L. and Turk, D. (2004). Crystal structure of NS-134 in complex with bovine cathepsin B: a two-headed epoxysuccinyl inhibitor extends along the entire active-site cleft. *Biochem. j.*, **381**, 511-517.

Stryer, L. (1996): Biochemistry 4th Ed. W.H. Freeman and Co., New York, 45-65.

Tronrud, D.E. (1992) Conjugate-direction minimization: an improved method for the refinement of macromolecules. *Acta Cryst.* A**48**, 912-916

Tronrud, D.E. (1997) TNT refinement package. Methods Enzymol., 277, 306-318.

Tronrud, D.E. (2004) Introduction to macromolecular refinement. *Acta Cryst.* D**60**, 2156-2168

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Cryst.* A**43**, 489-491.

Turk, D. (1992). *Weiterentwicklung eines Programms fuer Molekuelgraphik und Elektrondichte-Manipulation und seine Anwendung auf verschiedene Protein-Strukturaufklaerungen.* Ph.D. Thesis, Technische Universitaet, Muenchen.

Vagin, A.A., Steiner, R.A., Lebedev, A.A., Potterton, L., McNicholas, S., Long, F. and Murshudov, G.N. (2004). *REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Cryst.*, D**60**, 2184-2195.

Vali G. (1995). Biological ice nucleation and its applications. Ed. Lee R.E. Jr., Warren G.J. and Gusta L.V. St. Paul. APS Press 1-28.

Watenpaugh K.D., Sieker L.C., Herriott J.R. and Jensen L.H. (1972). Cold Spring Harbour Symp. Quant. Biol. **36**, 359-367.

Watenpaugh K.D., Sieker L.C., Herriott J.R. and Jensen L.H. (1972). Refinement of the model of a protein: rubredoxin at 1.5 Å resolution. *Acta Cryst* B**29**, 943-956.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II. J. Mol. Biol. 180, 301-329.

Wlodek, S., Skillman, A.G. *and Nicholls,* A. (2006). Automated ligand placement and refinement with a combined force field and shape potential. *Acta Cryst,*. D**62,** 741-749

Wright W.V. (1982). Computational crystallography. Ed. D. Sayre. Oxford. Clarendon Press. 294-302.

Yang, W. (1991). Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett.* **66**, 1438-1441.

Yang, W. T. & Lee, T. S. (1995). A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.* **103**, 5674-5678.

Yang, W., Hendrickson, W. A., Crouch, R. D. & Satow, Y**.** (1990). Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein. *Science* **249**, 1398-1405.

Yoshimatsu M. and Kozaki S. (1977). X-ray optics Ed. H.-J. Queisser, Berlin, Springer. Chapter 2.

Yu N., Yennawar H.P. and Kenneth M. Merz Jr (2005) Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics. *Acta Cryst.* **D**61, 322-332.

Index of Figures

Index of Tables

Index of Algorithms

Appendix

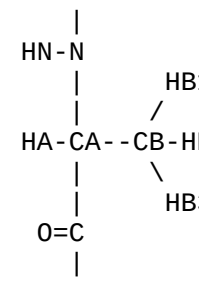
Appendix A - Amino acids abbreviations

Table 10: *Amino acid abbreviation table.*

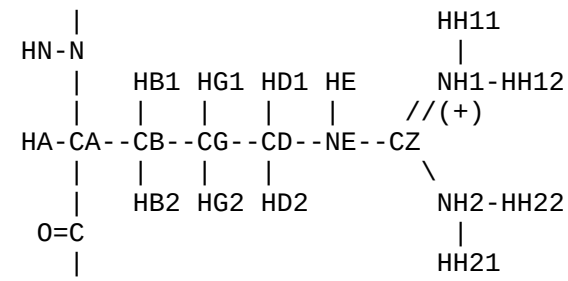
Name	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Appendix B - Amino acids atom naming conventions

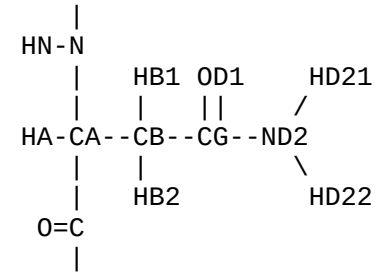
Alanine - ALA



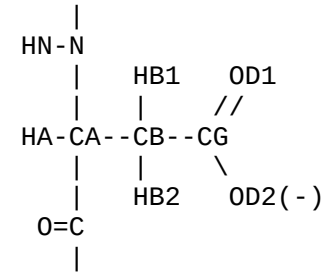
Arginine - ARG



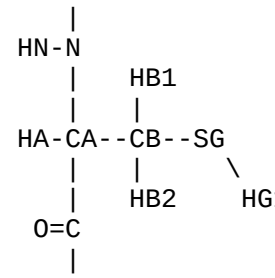
Asparagine - ASN



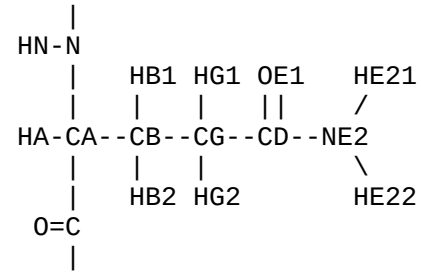
Aspartic acid - ASP



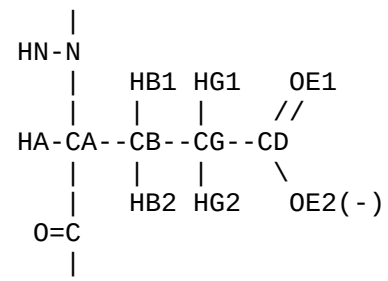
Cysteine - CYS



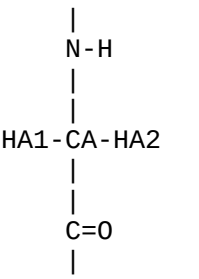
Glutamine - GLN



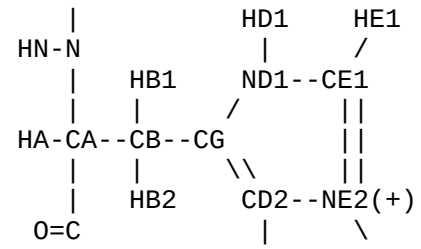
Glutamic acid - GLU



Glycine - GLY



Histidine - HIS



```
    |          HD2    HE2
```

Isoleucine - ILE

```
    |    HG21 HG22
 HN-N      | /
    |     CG2--HG23
    |    /
 HA-CA--CB-HB    HD1
    |    \       /
    |     CG1--CD--HD2
  O=C    / \     \
    | HG11 HG12  HD3
```

Leucine - LEU

```
    |        HD11 HD12
 HN-N          | /
    |   HB1   CD1--HD13
    |   |    /
 HA-CA--CB--CG-HG
    |   |    \
    |   HB2   CD2--HD23
  O=C          | \
    |        HD21 HD22
```

Lysine - LYS

```
    |
 HN-N
    |   HB1 HG1 HD1 HE1    HZ1
    |   |   |   |   |     /
 HA-CA--CB--CG--CD--CE--NZ--HZ2
    |   |   |   |   |     \
    |   HB2 HG2 HD2 HE2    HZ3
  O=C
    |
```

Methionine - MET

```
    |
 HN-N
    |   HB1 HG1     HE1
    |   |   |       |
 HA-CA--CB--CG--SD--CE--HE3
    |   |   |       |
    |   HB2 HG2     HE2
  O=C
    |
```

Phenylalanine - PHE

```
    |        HD1  HE1
 HN-N          |    |
    |   HB1  CD1--CE1
    |   |    //     \\
 HA-CA--CB--CG      CZ--HZ
    |   |    \  __  /
    |   HB2  CD2--CE2
  O=C          |    |
```

```
    |        HD2  HE2
```

Proline - PRO

```
      HD1 HD2
    |   \ /
    N---CD   HG1
    |     \  /
    |      CG
    |     /  \
 HA-CA--CB   HG2
    |   / \
    | HB1 HB2
  O=C
    |
```

Serine - SER

```
    |
 HN-N
    |   HB1
    |   |
 HA-CA--CB--OG
    |   |     \
    |   HB2    HG1
  O=C
    |
```

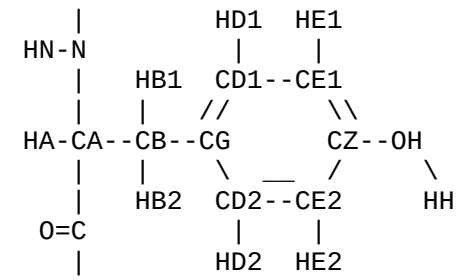
Threonine - THR

```
    |
 HN-N
    |     OG1--HG1
    |    /
 HA-CA--CB-HB
    |    \
    |     CG2--HG21
  O=C    / \
    | HG21 HG22
```

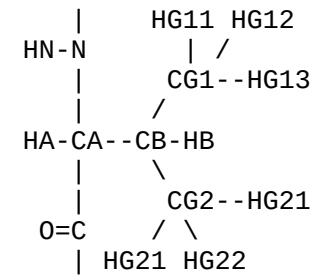
Tryptophan - TRP

```
    |                  HE3
 HN-N                   |
    |   HB1            CE3
    |   |             /  \\
 HA-CA--CB---CG-----CD2   CZ3-HZ3
    |   |    ||     ||     |
    |   HB2  CD1    CE2   CH2-HH2
  O=C       /   \   / \  //
    |     HD1    NE1   CZ2
                 |     |
                HE1   HZ2
```

Tyrosine - TYR



Valine - VAL



Appendix C - PURY database generation

PURY database generation is a streamline of processes which from data deposited in the crystallographic database such as CSD (Allen et al., 1979) extract the information about topology and geometry of the deposited structures and builds the database of geometric restraints. This database is then accessed and exploited through the www server. Below the layout of the process, programs and script control, input and output files and the data structures are described. While describing PURY environment the following conventions are used: **Scripts and programs** are marked in bold, ***subroutines*** are shown in italics and bold and variables are shown in italics.

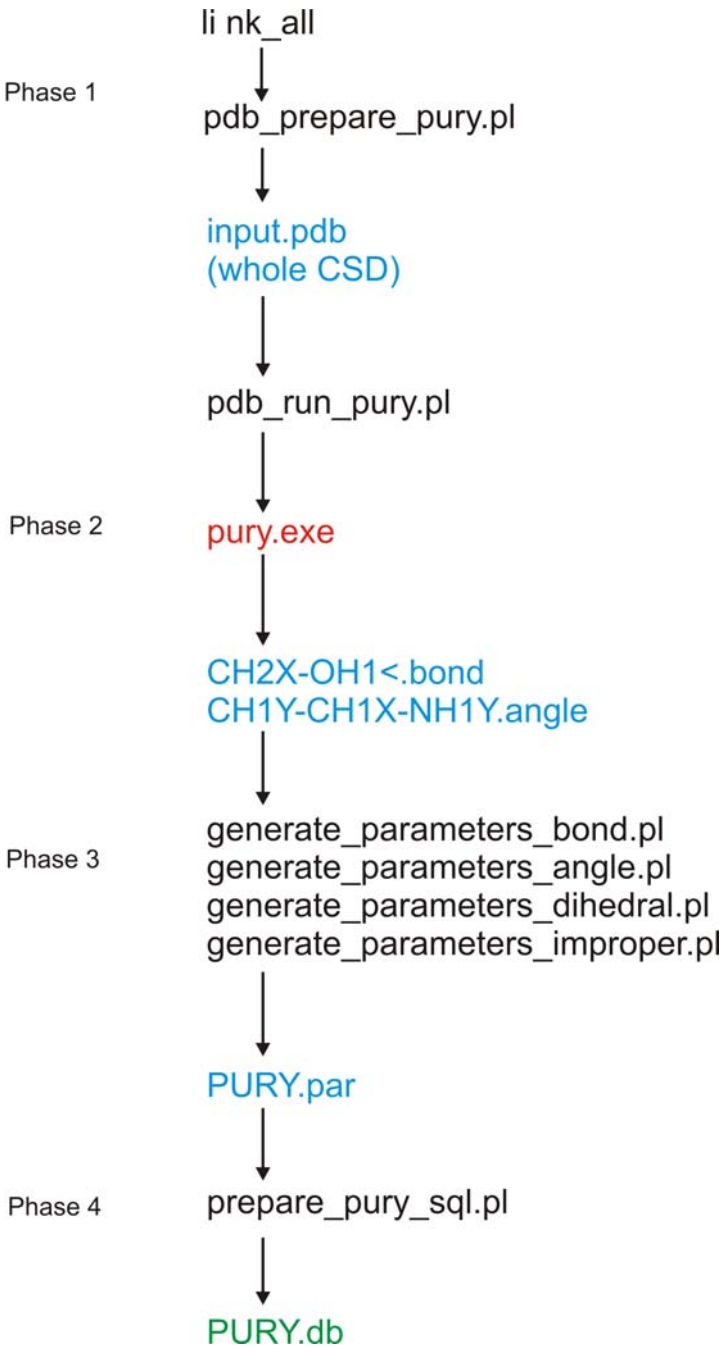


Figure 20: *PURY database generation flowchart.* From PDB molecules to the parameter database. Phases of the process and connections between various parts of PURY program. Colours: black - PERL scripts, red - PURY executable, blue - input/output files, green - SQLite parameter database

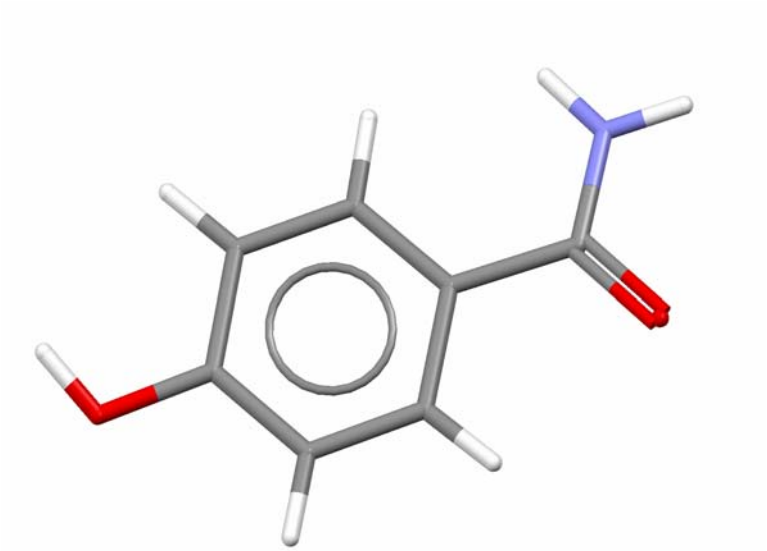


Figure 21: *VIDMAX.pdb.* The structure is selected from CSD v5.29 to demonstrate the PURY procedures. The figure was made with CSD Mercury (Macrae et al., 2002).

Phase 1

For database generation the environment was prepared which included specific directory tree to accommodate files generated by PURY. The environment is prepared by a script called **link_all** to which a single keyword is passed. The keyword presents the root name to store the parameters. In our case **link_all** *DEMO* was used which calls the **pdb_prepare_pury.pl** script (marked with the letter **P** in Table 15) with the option of the root name. The script creates proper directories and the required log files. The script printout is shown in Phase Print 1 and the directory tree layout in Print 2.

Phase Print 1: PURY environment preparation script. The printout made by link_all.

```
>link_all DEMO

Linking files
Preparing environment
FFU : Checking environment

FFU : FFC Executable ..... OK
FFU : TEMP ............... OK
FFU : Creating environment files

FFU : count file ........ OK
FFU : workdir file ........ OK
FFU : status file ........ OK
FFU : failed file ........ OK
FFU : batch file ........ OK
FFU : workfile file ........ OK
FFU : used file ........ OK
FFU : ffc_exe file ........ OK
FFU : log file ........ OK
FFU : OUT ............... OK

FFU : STAT bond ................ OK
FFU : STAT angl ................ OK
FFU : STAT dihe ................ OK
FFU : STAT impr ................ OK
FFU : CHARTS ............... OK

Linking pury.exe
All done for project DEMO
```

Print 2: *PURY prior to extraction of the directory tree. The printout shows the directory tree generated for PURY including log files. The printout was made with LINUX tree command using -a and --dirsfirst switches.*

```
.
|-- CHARTS
|-- ffc_OUT
|   `-- DEMO.par
|-- ffc_STAT_angle
|-- ffc_STAT_bond
|-- ffc_STAT_dihedral
|-- ffc_STAT_improper
|-- ffc_TEMP
|   |-- .ffc_batch
|   |-- .ffc_count
|   |-- .ffc_failed
|   |-- .ffc_log
|   |-- .ffc_status
|   |-- .ffc_usedfile
|   |-- .ffc_workdir
|   |-- .ffc_workfile
|   |-- DEMO.angle
|   |-- DEMO.bond
|   |-- DEMO.class
|   |-- DEMO.dihedral
|   |-- DEMO.improper
|   `-- DEMO.nonbonded
|-- .ffc_log
|-- .ffc_status
|-- generate_parameters_dihe.pl ->
/ajda/miha/PURY/last_known_working_version/create_parameters_dihe.pl
|-- generate_parameters_impr.pl ->
/ajda/miha/PURY/last_known_working_version/create_parameters_impr.pl
|-- generate_parameters_angle.pl ->
/ajda/miha/PURY/last_known_working_version/generate_parameters_angle.pl
|-- generate_parameters_bond.pl ->
/ajda/miha/PURY/last_known_working_version/generate_parameters_bond.pl
|-- pdb_prepare_pury.pl ->
/ajda/miha/PURY/last_known_working_version/pdb_prepare_pury.pl
|-- pdb_run_pury.pl -> /ajda/miha/PURY/last_known_working_version/pdb_run_pury.pl
|-- pdb_run_pury_connect.pl ->
/ajda/miha/PURY/last_known_working_version/pdb_run_pury_connect.pl
'-- pury.exe -> /ajda/miha/PURY/last_known_working_version/pury.exe

7 directories, 25 files
```

Phase 2

The generation of the database is divided in three steps and is run with **pdb_run_pury.pl** script. in Table 15 marked with letter E. The script is run with two options for the root name and the location of the stored files. The first phase of the database generation is one molecule at the time analysed by the PURY executable of all PDB files submitted.

PURY core program

The PURY core program is written in C and the ring finding routine is written in FORTRAN programming language. The PURY executable is compiled from 5 files: **working.c, funcPury.h, varPury.h, elementPury.h** and **ringsPury.f**. The program is controlled by the following switches shown in Table 11.

Table 11: *PURY input switches.* The first column shows the available switches used for running PURY executable. The second column shows the values that can follow the switches if any. The third column gives a description of the switch.

Switch	Value	Description
-r	NAME	Defines root name value which is used for naming all storage and parameter files (eq.: NAME.par)
-f	Filename.pdb	Selected file to be parsed by PURY
-e		Runs PURY in extraction mode with provided file
-t		Runs PURY in topology generation mode with provided file
-m	Molecule_name RES	Additional switch used in topology mode - prepares RES_top.main and RES_par.main files
-b		Used for bonded residues (obsolete - transferred into Perl part)
-s		Runs PURY with some statistics
-o	0, 1, 2	Runs PURY with various levels of output
-w	Work_file.pdb	Storing file name for bug tracking purposes

Working.c is the core program which contains all subroutines needed for parsing the input, for processing the data and for providing the raw output which is later processed by Perl scripts. The subroutine ***get_args*** reads and parses the input switches which later direct the flow of the program. The subroutine ***read_pdb*** reads standard PDB files by reading *ATOM*, *HETATM* and *CONECT* keywords. The routine internally stores the whole structure in arrays of structural variables named atom and connect. *Atom struct variable* contains several members of basic types float, int and char. The basic members of *atom struct variable* are shown in Table 12. *Connect struct variable* contains only a single member which is an array for storing bonded atom numbers.

Table 12: *Atoms array variables.* The first column shows the variable name, the second one the variable type, the third one gives a description of the variable and the possible choices. The fourth column shows example values.

Members	Type	Description	Example value
atom_num	Int	atom number as extracted from PDB file	1, 2, 3, ...
x, y, z	float	atom coordinates	1.123, 4.678, -0.456
name	char	atom name	CA, MN, OD1, ...
element	char	element symbol as extracted from atom name	H, C, Cl, O, ...
short_class	char	4-letter atom class generated by PURY	CH1X, HP__, Fe_8, ...
valence	Int	number of bonded atoms (explicitly defined plus assigned hydrogen atoms)	0, 1, 2, ...
ring_type	Int	ring type (aromatic, aliphatic or chain)	0, 1, 2
ring_size	Int	ring size (up to 12 members)	5, 6, 7, ...
hydrogens	Int	Number of explicitly and implicitly bonded hydrogen atoms	0, 1, 2, ...
Oxygens	Int	Number of bonded oxygen atoms	0, 1, 2, ...

After storing the descriptions of atom names, elements, coordinates and connectivity subroutines ***find_rings, calculate_bonds, calculate_angles, calculate_dihedrals*** *and* ***calculate_impropers*** calls follow. Again, each subroutine has a special *array of struct variables* for storing specific derived data. Sizes of the arrays are defined at the compile time by the *ATOMS* variable which is set to 500000 atoms. Consequently, *ATOMS* variable times 2 of bonds are set, times 3 of angles and times 2 of the dihedral

and the improper arrays are set. The variables in arrays and their members are shown in Table 13.

Table 13: PURY structural elements variables. The first column shows the array name, the second one the array variable names, the third one the variable types, the fourth column contains the basic explanation with possible choices, whereas the fifth one gives example values.

Variable	Members	Type	Description	Example value
bond_size	bond	float	bond length	1.389
	atom1	int	Atom ID	1
	atom2	int	Atom ID	2
angle_size	angle	float	angle dimension in degrees	109.400
	angle_dist	float	Angle dimension as distance in Angstroms between atoms 1 and 3 for SHELX compatibility	2.234
	atom1	int	Atom ID	13
	atom2	int	Atom ID	10
	atom3	int	Atom ID	14
dihedral_size	dihedral	float	dihedral size in degrees	179.023
	used	int	on/off switch weather dihedral has been used in topology generation	0 or 1
	atom1	int	Atom ID	1
	atom2	int	Atom ID	2
	atom3	int	Atom ID	3
	atom4	int	Atom ID	5
improper_size	improper	float	improper size in degrees	35.980
	real_value	float	improper size in degrees but with negative or positive sign (compatible with MAIN)	-35.980
	shelx_improper	float	improper value as chiral volume in $Å^3$ for SHELX compatibility	1.345
	atom1	int	Atom ID	4
	atom2	int	Atom ID	6
	atom3	int	Atom ID	7
	atom4	int	Atom ID	9
ring	members	array of int	number of atoms in ring	6
	size	int	size of ring	6
	type	int	ring type (aliphatic, aromatic or bridged)	2
	Distance	Float	maximum distance in Angstroms between two most distant planes of the ring - determining planarity	0.234

The next subroutine called is named ***check_organic_and_add_hydrogen***. This subroutine tags organic type atoms, calculates valences and determines hybridization of organic type atoms. This subroutine evaluates geometry of atoms based on internally stored cutoff values. Cutoff values for bonds and angles are stored inside **varPury.h**. Bond cutoff values are marked with element names and bond types (*V1_C1_C1*) while angle cutoff values are marked with hybridization type of atom (*SP3_MAX*).

Number of bonds defines the hybridization state of the atom. The evaluation is strongly based on the number of bonds that atom forms. Atoms with more than one bond are easier to describe and define the hybridization state. Atoms forming only one bond are evaluated last and re-evaluated 2 more times to fix hybridization states if necessary by examining close chemical environment more extensively. Already parsed and defined atoms are used as guide lines to assign hybridization to yet undeclared atoms. In case of there are no hydrogen atoms present, the subroutine tries to attach hydrogen atoms to heavy atoms according to their guessed hybridization state.

Subroutine ***shortTopology*** generates atom classes which are unique to every atom in different chemical environment. Based on connectivity table and the information gathered earlier (number of bonded oxygen atoms, number of bonded hydrogen atoms, ring involvement and hybridization) atom classes are generated in 6 steps (special classes, organic aromatic ring classes, organic ring classes, organic classes, metalo acids and metals). Special classes are assigned first including *COO_* - carboxyl acid carbon atom, *NH1G* - guanidinium group nitrogen atom, *NO2_* - nitro group nitrogen atom, *SO4_* - sulfuric acid sulphur atom and various other atom classes. Second step is assignment of atom classes to organic type atoms involved in aromatic rings. Steps for assignment atom classes to non aromatic rings and non ring organic type atoms follow. Steps to assign atom classes to atoms involved in metalo acids and metals themselves follow. By the end of this subroutine every atom passed to PURY has unique atom class signature. Extraction running example is shown in Phase Print 3.

Phase Print 3: *PURY extraction run.* The printouts were selected while running PURY executable in the extraction mode with VIDMAX.pdb file.

```
VIDMAX.pdb
Running extraction...
Using 1 files
Current file: VIDMAX.pdb
Total: 1 of 1
..//VIDMAX.pdb
/people/miha/MojiDokumenti/doktorat/PURY_test/PURY_base/pury.exe -s -r DEMO -e
..//VIDMAX.pdb -o 2 -w VIDMAX.pdb
0 0
Setting up environment
O
O
N
C
C
C
C
C
C
C
H
H
H
H
H
H
H

***** End of file *****

Total lines in file = 17

Atoms: 17
********************* Coordinates ********************

No     Type   Name   Class          X      Y      Z
   1   O       O1          1  0 1    3.364    1.188    4.844
   2   O       O2          1  0 2    5.032    2.156   -1.165
   3   N       N1          1  0 3    2.167    3.067    4.515
   4   C       C1          1  0 3    3.009    2.126    4.091
   5   C       C2          1  0 3    3.530    2.197    2.700
   6   C       C3          1  0 3    4.503    1.287    2.312
   7   C       C4          1  0 3    4.524    2.187    0.101
   8   C       C5          1  0 3    5.004    1.277    1.025
   9   C       C6          1  0 3    3.072    3.120    1.771
  10   C       C7          1  0 3    3.566    3.108    0.472
  11   H       H1          0  0 1    1.789    2.948    5.359
  12   H       H2          0  0 1    1.876    3.724    3.972
  13   H       H3          0  0 1    2.388    3.786    2.016
  14   H       H4          0  0 1    3.229    3.724   -0.202
  15   H       H5          0  0 1    4.555    2.683   -1.716
```

```
  16   H       H6          0  0 1    5.686    0.662    0.763
  17   H       H7          0  0 1    4.790    0.644    2.976

********************** End of data *******************

Atoms: 17
******************** Connectivity table ***************

   1      4
   2      7     15
   3      4     11     12
   4      1      3      5
   5      4      6      9
   6      5      8     17
   7      2      8     10
   8      6      7     16
   9      5     10     13
  10      7      9     14
  11      3
  12      3
  13      9
  14     10
  15      2
  16      8
  17      6

******************** End of data *****************

       Data OK!

       Starting list generation!

Finished with bonds...
Number of bonds: 17
No errors reported
NOF_RINGS: 1
  RLOOPS> end nrings 1
Ring diff 0.032730 member[1] 8
Aromatic ring 0 size 6
Finished with angles...
Finished with dihedrals...
Finished with impropers...

       Digestion completed!

Number of atoms:     17
Number of bonds:     17
Number of angles:    25
Number of dihedrals: 14
Number of impropers: 8
Number of rings:     1

Start SPECIAL CLASSES

Special classes - DONE
Aromatic Rings - DONE
0 Type 0 element O class O2C_
1 Type 0 element O class OH1<
2 Type 0 element N class NH2Y
3 Type 0 element C class C__Y
Organics - DONE
Rings - DONE
Metal Acids - DONE
Metals - DONE
REST - DONE

Atoms: 17
********************** Coordinates ******************

No     Type   Name   Class         X      Y      Z
```

```
   1   O       O1     O2C_ 1 22 1    3.364    1.188    4.844
   2   O       O2     OH1< 1 21 2    5.032    2.156   -1.165
   3   N       N1     NH2Y 1 20 3    2.167    3.067    4.515
   4   C       C1     C__Y 1  2 3    3.009    2.126    4.091
   5   C       C2     CC_6 1  6 3    3.530    2.197    2.700
   6   C       C3     CH_6 1  6 3    4.503    1.287    2.312
   7   C       C4     CO_6 1  6 3    4.524    2.187    0.101
   8   C       C5     CH_6 1  6 3    5.004    1.277    1.025
   9   C       C6     CH_6 1  6 3    3.072    3.120    1.771
  10   C       C7     CH_6 1  6 3    3.566    3.108    0.472
  11   H       H1     HP__ 0 62 1    1.789    2.948    5.359
  12   H       H2     HP__ 0 62 1    1.876    3.724    3.972
  13   H       H3     HC__ 0 61 1    2.388    3.786    2.016
  14   H       H4     HC__ 0 61 1    3.229    3.724   -0.202
  15   H       H5     HP__ 0 62 1    4.555    2.683   -1.716
  16   H       H6     HC__ 0 61 1    5.686    0.662    0.763
  17   H       H7     HC__ 0 61 1    4.790    0.644    2.976

********************** End of data ********************

      Writting the database!

Writting bond data...
Writting angle data...
Writting dihedral data...
Writting improper data...
Writting new database...
Class:    9
Reading and writting class database...
      >>> Class Database written <<<

Adding bond database...
Bonds:    17
      >>> Bond Database written <<<

Adding angle database...
Angles:    25
      >>> Angle Database written <<<

Adding improper database...
Impropers:    8
      >>> Improper Database written <<<

Adding dihedral database...
Dihedrals:    14
      >>> Dihedral Database written <<<

Writting new database...
Nonbonded:    9
Reading and writting nonbonded database...
      >>> Nonbonded Database written <<<

Combining all databases
DONE>>> Library generated in 0 seconds
Done in 0 seconds
```

Phase 3

Subroutines ***new_write_(bond, angle, dihedral, improper)_to_file*** are called storing all parsed bond, angle, dihedral and improper values in separate files with corresponding atom classes and file name from which data was extracted. The results are lists of all bonding terms. Table 14 shows in which directories, files and how data is stored in parameter generation mode.

Table 14: *PURY database generation files and directories.* The first column shows the directory names, the second one the example names of the files containing the bonding parameters and the third column shows example values stored in the files; the values are the term and the corresponding CSD refcode from which the term was extracted, and the atom names of atoms involved in the term.

Directory	File	Data
ffc_STAT_bond	CH_6-CC_6.bond	1.387578;VIDMAX.pdb= C2 - C3
ffc_STAT_angle	CH_6-CH_6-CC_6.angle	121.311287;2.413442;VIDMAX.pdb= C2 - C3 - C5
ffc_STAT_dihedral	O2C_-C__Y-CH1X-CH2X.dihedral	120.879;ABEUBE.pdb
ffc_STAT_improper	CC_6-CH_6-C__Y-CH_6.improper	0.511673;VIDMAX.pdb

In database generation phase PERL scripts start generating *DEMO.par* file which is text based parameter library. Scripts **generate_parameters_(bond, angle, dihedral, improper).pl** marked with letter **D** in Table 15 are called. Each script reads file with bonded values, calculates their average and sigma values. The force value is calculated from the sigma value. The whole database is now stored in single TXT file shown in example of bond term in Print 4.

Print 4: *Bond example database entry.* An example of the bond entry for CH_6-CC_6 bond in DEMO.par database. The stored values are termed keyword, atom class 1, atom class 2, force value, average value, sigma value and number of repeats.

```
bond   CC_6 CH_6    120.816      1.387 !      0.0700 2
```

Phase 4

The text based parameter database file *DEMO.par* is later converted into SQLite database using **prepare_sql_pury.pl** script marked with **S** in Table 15. With conversion into SQLite format the parameter extraction and generation procedure is complete. The Print 5 shows final directory tree at the end of the process.

Table 15: *PURY database generation scripts.* The first column shows the script names, the second column the part of the process they are involved in, the third column shows optional values to be passed to scripts. The fourth column provides the basic explanation.

Script	Tag	Options	Description
pdb_prepare_pury.pl	P	ROOTNAME	Generates environment for PURY
pdb_run_pury.pl	E	ROOTNAME /directory/with/files	Runs PURY in extraction mode with provided files
generate_parameters_bond.pl	D	ROOTNAME	Creates bond part of database
generate_parameters_angle.pl	D	ROOTNAME	Creates angle part of database
generate_parameters_dihe.pl	D	ROOTNAME	Creates dihedral part of database
generate_parameters_impr.pl	D	ROOTNAME	Creates improper part of database
prepare_pury_sql.pl	S	(ROOTNAME)	Transforms TXT PURY parameter database to more efficient SQLite based database.

Print 5: *PURY final directory tree.* The printout shows a directory tree generated for PURY including log files. The printout was made with LINUX tree command using -a and --dirsfirst switches.

```
.
|-- CHARTS
|-- ffc_OUT
|   |-- DEMO.angle
|   |-- DEMO.bond
|   |-- DEMO.class
|   |-- DEMO.dihedral
|   |-- DEMO.improper
|   |-- DEMO.nonbonded
|   `-- DEMO.par
|-- ffc_STAT_angle
|   |-- CC_6-C__Y-NH2Y.angle
|   |-- CC_6-C__Y-O2C_.angle
|   |-- CH_6-CC_6-CH_6.angle
|   |-- CH_6-CC_6-C__Y.angle
|   |-- CH_6-CH_6-CC_6.angle
|   |-- CH_6-CO_6-CH_6.angle
|   |-- CH_6-CO_6-OH1<.angle
|   |-- CO_6-CH_6-CH_6.angle
|   |-- HC__-CH_6-CC_6.angle
|   |-- HC__-CH_6-CH_6.angle
|   |-- HC__-CH_6-CO_6.angle
|   |-- HP__-NH2Y-C__Y.angle
|   |-- HP__-NH2Y-HP__.angle
|   |-- HP__-OH1<-CO_6.angle
|   `-- NH2Y-C__Y-O2C_.angle
|-- ffc_STAT_bond
|   |-- CC_6-C__Y.bond
|   |-- CH_6-CC_6.bond
|   |-- CH_6-CH_6.bond
|   |-- CH_6-CO_6.bond
|   |-- CO_6-OH1<.bond
|   |-- C__Y-NH2Y.bond
|   |-- C__Y-O2C_.bond
|   |-- HC__-CH_6.bond
|   |-- HP__-NH2Y.bond
|   `-- HP__-OH1<.bond
|-- ffc_STAT_dihedral
|   |-- CH_6-CC_6-C__Y-NH2Y.dihedral
|   |-- CH_6-CC_6-C__Y-O2C_.dihedral
|   |-- CH_6-CH_6-CC_6-CH_6.dihedral
|   |-- CH_6-CH_6-CC_6-C__Y.dihedral
|   |-- CH_6-CH_6-CO_6-OH1<.dihedral
|   |-- CH_6-CO_6-CH_6-CH_6.dihedral
|   `-- CO_6-CH_6-CH_6-CC_6.dihedral
|-- ffc_STAT_improper
|   |-- CC_6-CH_6-C__Y-CH_6.improper
|   |-- CH_6-CC_6-HC__-CH_6.improper
|   |-- CH_6-CH_6-HC__-CO_6.improper
|   |-- CO_6-CH_6-OH1<-CH_6.improper
|   |-- C__Y-CC_6-O2C_-NH2Y.improper
|   `-- NH2Y-C__Y-HP__-HP__.improper
|-- ffc_TEMP
|   |-- .ffc_batch
|   |-- .ffc_count
|   |-- .ffc_failed
|   |-- .ffc_log
|   |-- .ffc_status
|   |-- .ffc_usedfile
|   |-- .ffc_workdir
|   |-- .ffc_workfile
|   |-- DEMO.angle
|   |-- DEMO.angle_0
|   |-- DEMO.bond
|   |-- DEMO.bond_0
```

```
|   |-- DEMO.class
|   |-- DEMO.dihedral
|   |-- DEMO.dihedral_0
|   |-- DEMO.improper
|   |-- DEMO.improper_0
|   `-- DEMO.nonbonded
|-- .ffc_log
|-- .ffc_status
|-- bond.log
|-- charge.tmp
/ajda/miha/PURY/last_known_working_version/create_parameters_class.pl
|-- generate_parameters_dihe.pl ->
/ajda/miha/PURY/last_known_working_version/create_parameters_dihe.pl
|-- generate_parameters_impr.pl ->
/ajda/miha/PURY/last_known_working_version/create_parameters_impr.pl
|-- dir.tree
|-- generate_parameters_angle.pl ->
/ajda/miha/PURY/last_known_working_version/generate_parameters_angle.pl
|-- generate_parameters_bond.pl ->
/ajda/miha/PURY/last_known_working_version/generate_parameters_bond.pl
|-- organic.run
|-- pdb_prepare_pury.pl ->
/ajda/miha/PURY/last_known_working_version/pdb_prepare_pury.pl
|-- pdb_run_pury.pl -> /ajda/miha/PURY/last_known_working_version/pdb_run_pury.pl
|-- pdb_run_pury_connect.pl ->
/ajda/miha/PURY/last_known_working_version/pdb_run_pury_connect.pl
|-- prepare.log
|-- pury.exe -> /ajda/miha/PURY/last_known_working_version/pury.exe
`-- run.log

7 directories, 80 files
```

Appendix D - PURY topology generation

For automatic access and generation of topology and parameter files to users, the PURY WWW server (http://pury.ijs.si and http://pury.ccdc.cam.ac.uk) was created. It is constructed of several HTML pages and CGI scripts. The main functionality for preparing TOP and PAR files is done with CGI script **execute.cgi** and the script **serverTopo.pl** which is run by CGI script. The CGI script provides means of input and output for communication with user. The **serverTopo.pl** script is used to process the file, to determine connectivity, to distinguish the protein from hetero molecules, to call PURY executable and to generate PAR and TOP files. The server flowchart is shown in Figure 21.

Phase 1

User deposits the PDB file, draws the molecule using JME editor or deposits the SMILES string via the web server interface. To demonstrate the process of the topology preparation, the CSD VIDMEX file is shown, which was uploaded via the server interface.

Phase 2

The input is processed or uploaded by **execute.cgi** script. The script creates a project directory in which all files are stored. The directory and the file structure of a demonstration project is shown in Print 6. All the project directories are assigned by consecutive numbers. **Execute.cgi** script executes **serverTopol.pl** which further processes data.

Phase 3

ServerTopo.pl reads the uploaded PDB file, reads or calculates the connectivity and runs **PURY executable** in the topology generation mode using the *-t* switch. When **serverTopo.pl** executes PURY executable, a similar output is obtained as in the log file shown in Phase Print 3. In the topology generation mode for writing the output ***new_write_topo_to_file***, a subroutine is called for storing all the data describing the parsed molecule into a single file name *RES_new.tmp*. *RES* can be any three-letter name used to describe a residue.

Phase 4

A prepared temporary file is shown in Print 7. *RES_new.tmp* is processed by serverTopo.pl. The script accesses and reads **PURY.db** according to the bonding terms generated in previous steps. From the database average values, sigma values and force values are extracted. The extracted values from **PURY.db** and the data read from *RES_new.tmp* are used for the TOP and PAR files generation for supported refinement programs. The script has a unique subroutine for every refinement program. MAIN output is generated by the ***write_main_topo*** subroutine, CNS output by the ***write_cns_topo*** subroutine and REFMAC output by the ***write_refmac_topo*** subroutine. The MAIN specific files are shown in Print 8 and Print 9, the CNS specific files in Print 10 and Print 11 and the REFMAC file in Print 12. Output files can be downloaded by users via a prepared HTML page.

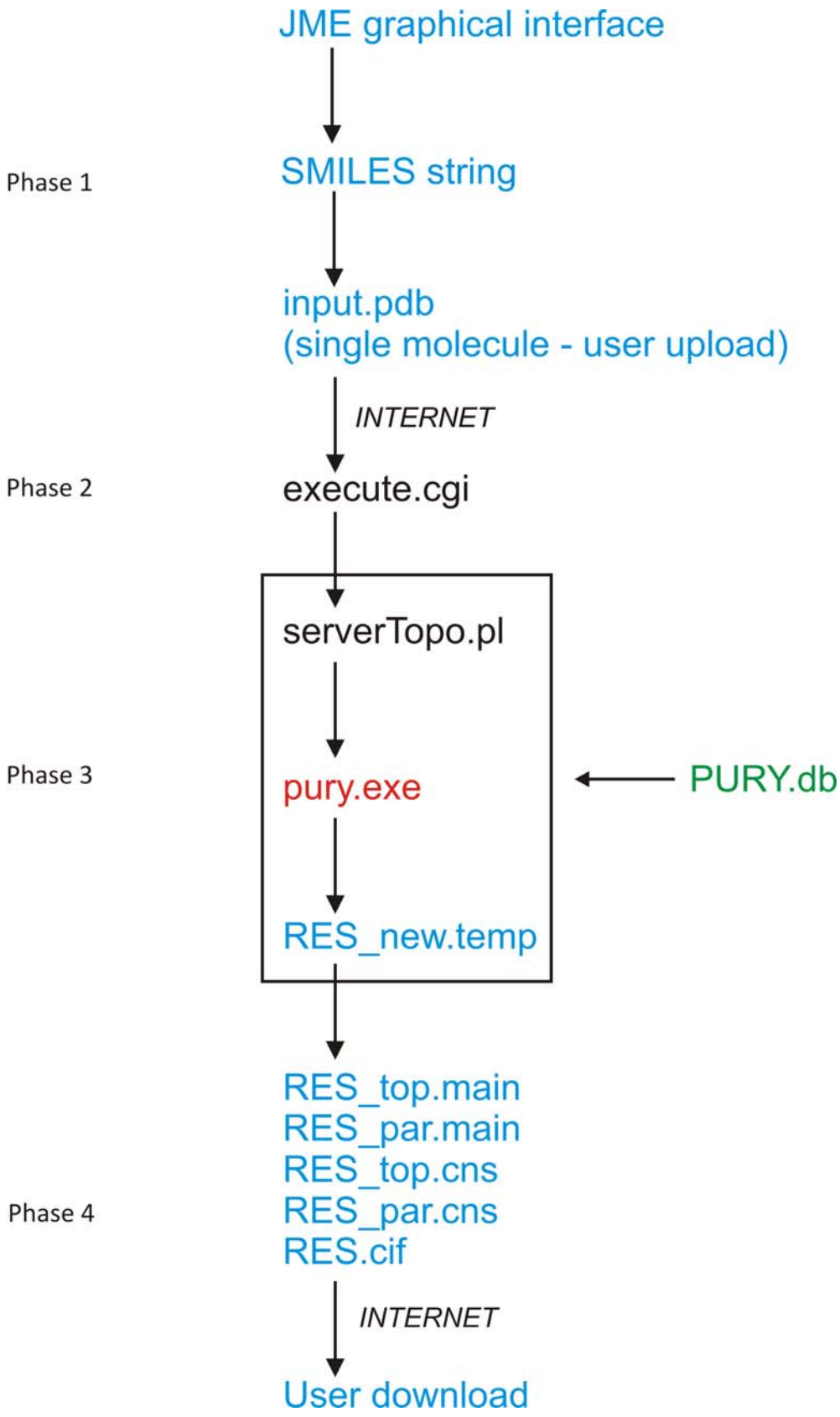


Figure 22: *PURY topology server flowchart.* From a user uploaded PDB molecule to the topology and the corresponding parameter files for various refinement programs. Colours: black - PERL scripts, red - PURY executable, blue - input/output, green - SQLite parameter database

Print 6: *Project 636 work space.* The first line shows part of the server directory tree, the rest are the files in this directory.

```
./jobs/636

.
|-- UNK.cif
|-- UNK.par
|-- UNK.pdb
|-- UNK_new.tmp
|-- UNK_par.cns
|-- UNK_par.main
|-- UNK_top.cns
|-- UNK_top.main
|-- VIDMAX.pdb
|-- VIDMAX.tmp
|-- bond.log
|-- charge.tmp
|-- data.raw
|-- end.txt
|-- get_top_par_pury_UNK.com
|-- libcheck.bat
|-- libcheck.doc
|-- libcheck.log
|-- libcheck.run
|-- log.backup
|-- log.perlfile
|-- organic.run
|-- temp_UNK.cif
|-- temp_UNK.pdb
`-- temp_UNK.ps
```

Print 7: *VIDMAX temporary file*. The temporary file contains all the structural elements of PDB file with generated atom classes and calculated experimental values. The data are later used for parameter extraction and TOP and PAR file generation.

```
clas O2C_  15.9994
clas OH1<  17.0074
clas NH2Y  16.0226
clas C__Y  12.0112
clas CC_6  12.0112
clas CH_6  13.0191
clas CO_6  12.0112
clas HP__   1.0080
clas HC__   1.0080

desc VIDMAX
mole UNK

atom  O1  1 UNK    1 O2C_    0.000    0.000    0.000
atom  O2  2 UNK    1 OH1<    1.668    0.968   -6.009
atom  N1  3 UNK    1 NH2Y   -1.197    1.879   -0.329
atom  C1  4 UNK    1 C__Y   -0.355    0.938   -0.753
atom  C2  5 UNK    1 CC_6    0.166    1.009   -2.144
atom  C3  6 UNK    1 CH_6    1.139    0.099   -2.532
atom  C4  7 UNK    1 CO_6    1.160    0.999   -4.743
atom  C5  8 UNK    1 CH_6    1.640    0.089   -3.819
atom  C6  9 UNK    1 CH_6   -0.292    1.932   -3.073
atom  C7  10 UNK    1 CH_6    0.202    1.920   -4.372
atom  H1  11 UNK    1 HP__   -1.575    1.760    0.515
atom  H2  12 UNK    1 HP__   -1.488    2.536   -0.872
atom  H3  13 UNK    1 HC__   -0.976    2.598   -2.828
atom  H4  14 UNK    1 HC__   -0.135    2.536   -5.046
atom  H5  15 UNK    1 HP__    1.191    1.495   -6.560
atom  H6  16 UNK    1 HC__    2.322   -0.526   -4.081
atom  H7  17 UNK    1 HC__    1.426   -0.544   -1.868

bond  O1  1 UNK    1  C1  4 UNK    1    1.254
bond  O2  2 UNK    1  C4  7 UNK    1    1.364
bond  O2  2 UNK    1  H5  15 UNK    1    0.899
```

```
bond  N1  3 UNK    1  C1  4 UNK    1    1.332
bond  N1  3 UNK    1  H1  11 UNK    1    0.932
bond  N1  3 UNK    1  H2  12 UNK    1    0.901
bond  C1  4 UNK    1  C2  5 UNK    1    1.487
bond  C2  5 UNK    1  C3  6 UNK    1    1.388
bond  C2  5 UNK    1  C6  9 UNK    1    1.387
bond  C3  6 UNK    1  C5  8 UNK    1    1.381
bond  C3  6 UNK    1  H7  17 UNK    1    0.968
bond  C4  7 UNK    1  C5  8 UNK    1    1.383
bond  C4  7 UNK    1  C7  10 UNK    1    1.380
bond  C5  8 UNK    1  H6  16 UNK    1    0.955
bond  C6  9 UNK    1  C7  10 UNK    1    1.390
bond  C6  9 UNK    1  H3  13 UNK    1    0.986
bond  C7  10 UNK    1  H4  14 UNK    1    0.973

angl  O1  1 UNK    1  C1  4 UNK    1  N1  3 UNK    1  121.077
angl  O1  1 UNK    1  C1  4 UNK    1  C2  5 UNK    1  119.878
angl  O2  2 UNK    1  C4  7 UNK    1  C5  8 UNK    1  118.410
angl  O2  2 UNK    1  C4  7 UNK    1  C7  10 UNK    1  121.545
angl  N1  3 UNK    1  C1  4 UNK    1  C2  5 UNK    1  119.045
angl  C1  4 UNK    1  N1  3 UNK    1  H1  11 UNK    1  117.016
angl  C1  4 UNK    1  N1  3 UNK    1  H2  12 UNK    1  121.848
angl  C1  4 UNK    1  C2  5 UNK    1  C3  6 UNK    1  118.420
angl  C1  4 UNK    1  C2  5 UNK    1  C6  9 UNK    1  122.852
angl  C2  5 UNK    1  C3  6 UNK    1  C5  8 UNK    1  121.311
angl  C2  5 UNK    1  C3  6 UNK    1  H7  17 UNK    1  116.859
angl  C2  5 UNK    1  C6  9 UNK    1  C7  10 UNK    1  120.184
angl  C2  5 UNK    1  C6  9 UNK    1  H3  13 UNK    1  120.811
angl  C3  6 UNK    1  C2  5 UNK    1  C6  9 UNK    1  118.722
angl  C3  6 UNK    1  C5  8 UNK    1  C4  7 UNK    1  119.470
angl  C3  6 UNK    1  C5  8 UNK    1  H6  16 UNK    1  121.290
angl  C4  7 UNK    1  O2  2 UNK    1  H5  15 UNK    1  110.957
angl  C4  7 UNK    1  C5  8 UNK    1  H6  16 UNK    1  119.233
angl  C4  7 UNK    1  C7  10 UNK    1  C6  9 UNK    1  120.257
angl  C4  7 UNK    1  C7  10 UNK    1  H4  14 UNK    1  118.469
angl  C5  8 UNK    1  C3  6 UNK    1  H7  17 UNK    1  121.799
angl  C5  8 UNK    1  C4  7 UNK    1  C7  10 UNK    1  120.041
angl  C6  9 UNK    1  C7  10 UNK    1  H4  14 UNK    1  121.246
angl  C7  10 UNK    1  C6  9 UNK    1  H3  13 UNK    1  119.002
angl  H1  11 UNK    1  N1  3 UNK    1  H2  12 UNK    1  120.520

dihe  O1  1 UNK    1  C1  4 UNK    1  N1  3 UNK    1  H1  11 UNK    1   -6.129
dihe  O1  1 UNK    1  C1  4 UNK    1  N1  3 UNK    1  H2  12 UNK    1 -177.175
dihe  O1  1 UNK    1  C1  4 UNK    1  C2  5 UNK    1  C3  6 UNK    1   -6.160
dihe  O1  1 UNK    1  C1  4 UNK    1  C2  5 UNK    1  C6  9 UNK    1  172.910
dihe  O2  2 UNK    1  C4  7 UNK    1  C5  8 UNK    1  C3  6 UNK    1 -179.856
dihe  O2  2 UNK    1  C4  7 UNK    1  C5  8 UNK    1  H6  16 UNK    1    1.062
dihe  O2  2 UNK    1  C4  7 UNK    1  C7  10 UNK    1  C6  9 UNK    1 -179.668
dihe  O2  2 UNK    1  C4  7 UNK    1  C7  10 UNK    1  H4  14 UNK    1    2.229
dihe  N1  3 UNK    1  C1  4 UNK    1  C2  5 UNK    1  C3  6 UNK    1  173.897
dihe  N1  3 UNK    1  C1  4 UNK    1  C2  5 UNK    1  C6  9 UNK    1   -7.033
dihe  C1  4 UNK    1  C2  5 UNK    1  C3  6 UNK    1  C5  8 UNK    1  178.324
dihe  C1  4 UNK    1  C2  5 UNK    1  C3  6 UNK    1  H7  17 UNK    1    0.308
dihe  C1  4 UNK    1  C2  5 UNK    1  C6  9 UNK    1  C7  10 UNK    1 -177.834
dihe  C1  4 UNK    1  C2  5 UNK    1  C6  9 UNK    1  H3  13 UNK    1    1.537
dihe  C2  5 UNK    1  C1  4 UNK    1  N1  3 UNK    1  H1  11 UNK    1  173.814
dihe  C2  5 UNK    1  C1  4 UNK    1  N1  3 UNK    1  H2  12 UNK    1    2.768
dihe  C2  5 UNK    1  C3  6 UNK    1  C5  8 UNK    1  C4  7 UNK    1   -0.288
dihe  C2  5 UNK    1  C3  6 UNK    1  C5  8 UNK    1  H6  16 UNK    1  178.773
dihe  C2  5 UNK    1  C6  9 UNK    1  C7  10 UNK    1  C4  7 UNK    1   -0.622
dihe  C2  5 UNK    1  C6  9 UNK    1  C7  10 UNK    1  H4  14 UNK    1  177.428
dihe  C3  6 UNK    1  C2  5 UNK    1  C6  9 UNK    1  C7  10 UNK    1    1.233
dihe  C3  6 UNK    1  C2  5 UNK    1  C6  9 UNK    1  H3  13 UNK    1 -179.396
dihe  C3  6 UNK    1  C5  8 UNK    1  C4  7 UNK    1  C7  10 UNK    1    0.919
dihe  C4  7 UNK    1  C5  8 UNK    1  C3  6 UNK    1  H7  17 UNK    1  177.630
dihe  C4  7 UNK    1  C7  10 UNK    1  C6  9 UNK    1  H3  13 UNK    1  180.000
dihe  C5  8 UNK    1  C3  6 UNK    1  C2  5 UNK    1  C6  9 UNK    1   -0.785
dihe  C5  8 UNK    1  C4  7 UNK    1  O2  2 UNK    1  H5  15 UNK    1  169.998
dihe  C5  8 UNK    1  C4  7 UNK    1  C7  10 UNK    1  C6  9 UNK    1   -0.469
dihe  C5  8 UNK    1  C4  7 UNK    1  C7  10 UNK    1  H4  14 UNK    1 -178.572
dihe  C6  9 UNK    1  C2  5 UNK    1  C3  6 UNK    1  H7  17 UNK    1 -178.801
```

```
dihe  C7  10 UNK    1  C4  7 UNK    1  O2  2 UNK    1  H5  15 UNK    1  -10.791
dihe  C7  10 UNK    1  C4  7 UNK    1  C5  8 UNK    1  H6  16 UNK    1 -178.162
dihe  H3  13 UNK    1  C6  9 UNK    1  C7  10 UNK    1  H4  14 UNK    1   -1.955
dihe  H6  16 UNK    1  C5  8 UNK    1  C3  6 UNK    1  H7  17 UNK    1   -3.309

impr  N1  3 UNK    1  C1  4 UNK    1  H1  11 UNK    1  H2  12 UNK    1    5.130
impr  C1  4 UNK    1  C2  5 UNK    1  N1  3 UNK    1  O1  1 UNK    1   -0.040
impr  C2  5 UNK    1  C3  6 UNK    1  C6  9 UNK    1  C1  4 UNK    1   -0.512
impr  C3  6 UNK    1  C2  5 UNK    1  C5  8 UNK    1  H7  17 UNK    1    1.190
impr  C4  7 UNK    1  C7  10 UNK    1  C5  8 UNK    1  O2  2 UNK    1    0.468
impr  C5  8 UNK    1  C3  6 UNK    1  C4  7 UNK    1  H6  16 UNK    1   -0.587
impr  C6  9 UNK    1  C2  5 UNK    1  C7  10 UNK    1  H3  13 UNK    1   -0.387
impr  C7  10 UNK    1  C6  9 UNK    1  C4  7 UNK    1  H4  14 UNK    1   -1.205
```

Print 8: *UNK_top.main.* A MAIN specific topology file generated upon the VIDMAX.pdb file upload.

```
CLASS O2C_ element  O
CLASS OH1< element  O
CLASS NH2Y element  N
CLASS C__Y element  C
CLASS CC_6 element  C
CLASS CH_6 element  C
CLASS CO_6 element  C
CLASS HP__ element  H
CLASS HC__ element  H

molecule UNK
group

atom O1 clas=O2C_ charge=-0.472 coor 0.000 0.000 0.000
atom O2 clas=OH1< charge=-0.520 coor 1.668 0.968 -6.009
atom N1 clas=NH2Y charge=-0.137 coor -1.197 1.879 -0.329
atom C1 clas=C__Y charge=0.020 coor -0.355 0.938 -0.753
atom C2 clas=CC_6 charge=0.000 coor 0.166 1.009 -2.144
atom C3 clas=CH_6 charge=-0.042 coor 1.139 0.099 -2.532
atom C4 clas=CO_6 charge=-0.000 coor 1.160 0.999 -4.743
atom C5 clas=CH_6 charge=-0.038 coor 1.640 0.089 -3.819
atom C6 clas=CH_6 charge=-0.041 coor -0.292 1.932 -3.073
atom C7 clas=CH_6 charge=-0.042 coor 0.202 1.920 -4.372
atom H1 clas=HP__ charge=0.201 coor -1.575 1.760 0.515
atom H2 clas=HP__ charge=0.201 coor -1.488 2.536 -0.872
atom H3 clas=HC__ charge=0.164 coor -0.976 2.598 -2.828
atom H4 clas=HC__ charge=0.163 coor -0.135 2.536 -5.046
atom H5 clas=HP__ charge=0.205 coor 1.191 1.495 -6.560
atom H6 clas=HC__ charge=0.165 coor 2.322 -0.526 -4.081
atom H7 clas=HC__ charge=0.168 coor 1.426 -0.544 -1.868

bond O1 C1 bond O2 C4 bond O2 H5 bond N1 C1 bond N1 H1
bond N1 H2 bond C1 C2 bond C2 C3 bond C2 C6 bond C3 C5
bond C3 H7 bond C4 C5 bond C4 C7 bond C5 H6 bond C6 C7
bond C6 H3 bond C7 H4

dihe O1 C1 N1 H1 dihe O1 C1 C2 C3 dihe C2 C3 C5 C4
dihe C2 C6 C7 C4 dihe C3 C2 C6 C7 dihe C3 C5 C4 C7
dihe C5 C3 C2 C6 dihe C5 C4 O2 H5 dihe C5 C4 C7 C6


impr N1 C1 H1 H2 impr C1 N1 C2 O1 impr C2 C6 C3 C1
impr C3 C2 C5 H7 impr C4 C7 C5 O2 impr C5 C4 C3 H6
impr C6 C7 C2 H3 impr C7 C4 C6 H4

inte O1 C1 1.254 N1 121.077 H1 -6.129
inte O2 C4 1.364 C5 118.410 C3 -179.856
inte N1 C1 1.332 C2 119.045 C3 173.897
inte C1 C2 1.487 C3 118.420 C5 178.324
inte C2 C1 1.487 N1 119.045 H1 173.814
inte C3 C2 1.388 C1 118.420 O1 -6.160
inte C4 C5 1.383 C3 119.470 C2 -0.288
inte C5 C3 1.381 C2 121.311 C1 178.324
```

```
inte C6 C2 1.387 C1 122.852 O1 172.910
inte C7 C6 1.390 C2 120.184 C1 -177.834
inte H1 N1 0.932 C1 117.016 O1 -6.129
inte H2 N1 0.901 C1 121.848 O1 -177.175
inte H3 C6 0.986 C2 120.811 C1 1.537
inte H4 C7 0.973 C4 118.469 O2 2.229
inte H5 O2 0.899 C4 110.957 C5 169.998
inte H6 C5 0.955 C4 119.233 O2 1.062
inte H7 C3 0.968 C2 116.859 C1 0.308
```

Print 9: *UNK_par.main*. MAIN specific parameter file generated upon upload of VIDMAX.pdb file.

```
class    O2C_ element  O
class    OH1< element  O
class    NH2Y element  N
class    C__Y element  C
class    CC_6 element  C
class    CH_6 element  C
class    CO_6 element  C
class    HP__ element  H
class    HC__ element  H

!BONDS TOP
bond    O2C_    C__Y  687.326    1.222
bond    O2C_    C__Y  687.326    1.222
bond    OH1<    CO_6 1347.773    1.355
bond    OH1<    HP__   53.174    0.889
bond    NH2Y    C__Y 1118.346    1.320
bond    NH2Y    HP__  104.167    0.889
bond    C__Y    CC_6  879.348    1.486
bond    CC_6    CH_6 2052.018    1.389
bond    CH_6    CH_6 1446.803    1.379
bond    CH_6    HC__  289.588    0.956
bond    CO_6    CH_6 1306.157    1.389

!ANGLES
angle    O2C_    C__Y    NH2Y  457.739   122.350
angle    O2C_    C__Y    CC_6  211.087  120.885
angle    OH1<    CO_6    CH_6  287.579  119.848
angle    NH2Y    C__Y    CC_6  290.817  118.900
angle    C__Y    NH2Y    HP__  119.074  119.255
angle    C__Y    CC_6    CH_6  387.623  120.623
angle    CC_6    CH_6    CH_6 1104.635  120.639
angle    CC_6    CH_6    HC__  595.954  119.361
angle    CH_6    CC_6    CH_6  965.717  118.369
angle    CH_6    CH_6    CO_6 1152.059  119.926
angle    CH_6    CH_6    HC__  477.778  119.918
angle    CO_6    OH1<    HP__   84.304  109.554
angle    CO_6    CH_6    HC__  537.000  119.746
angle    CH_6    CO_6    CH_6  797.215  120.221
angle    HP__    NH2Y    HP__   64.667  119.553

!DIHEDRALS
dihedral    O2C_    C__Y    NH2Y    HP__     5.00    3   180.00
dihedral    O2C_    C__Y    NH2Y    HP__     5.00    3   180.00
dihedral    O2C_    C__Y    CC_6    CH_6   127.77    2     0.00
dihedral    OH1<    CO_6    CH_6    CH_6   821.80    2   180.00
dihedral    OH1<    CO_6    CH_6    HC__     5.00    3   180.00
dihedral    NH2Y    C__Y    CC_6    CH_6    34.04    2     0.00
dihedral    C__Y    CC_6    CH_6    CH_6   904.26    2   180.00
dihedral    C__Y    CC_6    CH_6    HC__     5.00    3   180.00
dihedral    CC_6    C__Y    NH2Y    HP__     5.00    3   180.00
dihedral    CC_6    CH_6    CH_6    CO_6   933.70    2     0.00
dihedral    CC_6    CH_6    CH_6    HC__     5.00    3   180.00
dihedral    CH_6    CC_6    CH_6    CH_6   973.15    2     0.00
dihedral    CH_6    CC_6    CH_6    HC__     5.00    3   180.00
dihedral    CH_6    CH_6    CO_6    CH_6   937.94    2     0.00
dihedral    CO_6    CH_6    CH_6    HC__     5.00    3   180.00
dihedral    CH_6    CO_6    OH1<    HP__     5.00    3   180.00
dihedral    CH_6    CO_6    CH_6    HC__     5.00    3   180.00
```

```
dihedral    HC__    CH_6    CH_6    HC__     5.00    3   180.00

!IMPROPERS
improper    NH2Y    C__Y    HP__    HP__    38.87    0     0.00
improper    NH2Y    HP__    C__Y    HP__    38.87    0     0.00
improper    C__Y    CC_6    NH2Y    O2C_  1700.00    0     0.00
improper    C__Y    NH2Y    CC_6    O2C_  1700.00    0     0.00
improper    CC_6    CH_6    CH_6    C__Y  1073.07    0     0.00
improper    CH_6    CC_6    CH_6    HC__   646.82    0     0.00
improper    CH_6    CH_6    CC_6    HC__   646.82    0     0.00
improper    CO_6    CH_6    CH_6    OH1<  1700.00    0     0.00
improper    CH_6    CH_6    CO_6    HC__   619.63    0     0.00
improper    CH_6    CO_6    CH_6    HC__   619.63    0     0.00

!NONBONDED
nonbonded    O2C_    0.127    2.281    0.079    1.426
nonbonded    OH1<    0.127    2.281    0.079    1.426
nonbonded    NH2Y    0.190    2.281    0.119    1.426
nonbonded    C__Y    0.096    3.040    0.060    1.900
nonbonded    CC_6    0.096    3.040    0.060    1.900
nonbonded    CH_6    0.096    3.040    0.060    1.900
nonbonded    CO_6    0.096    3.040    0.060    1.900
nonbonded    HP__    0.040    1.120    0.025    0.700
nonbonded    HC__    0.040    1.120    0.025    0.700
```

Print 10: *UNK_top.cns.* CNS specific topology file generated upon the VIDMAX.pdb file upload.

```
set echo=false end

checkversion 1.1

MASS   O2C_  15.9994
MASS   OH1<  17.0074
MASS   NH2Y  16.0226
MASS   C__Y  12.0112
MASS   CC_6  12.0112
MASS   CH_6  13.0191
MASS   CO_6  12.0112
MASS   HP__   1.0080
MASS   HC__   1.0080

RESIdue UNK
GROUp

ATOM O1   TYPE=O2C_  CHARge=-0.472    END
ATOM O2   TYPE=OH1<  CHARge=-0.520    END
ATOM N1   TYPE=NH2Y  CHARge=-0.137    END
ATOM C1   TYPE=C__Y  CHARge=0.020    END
ATOM C2   TYPE=CC_6  CHARge=0.000    END
ATOM C3   TYPE=CH_6  CHARge=-0.042    END
ATOM C4   TYPE=CO_6  CHARge=-0.000    END
ATOM C5   TYPE=CH_6  CHARge=-0.038    END
ATOM C6   TYPE=CH_6  CHARge=-0.041    END
ATOM C7   TYPE=CH_6  CHARge=-0.042    END
ATOM H1   TYPE=HP__  CHARge=0.201    END
ATOM H2   TYPE=HP__  CHARge=0.201    END
ATOM H3   TYPE=HC__  CHARge=0.164    END
ATOM H4   TYPE=HC__  CHARge=0.163    END
ATOM H5   TYPE=HP__  CHARge=0.205    END
ATOM H6   TYPE=HC__  CHARge=0.165    END
ATOM H7   TYPE=HC__  CHARge=0.168    END

BOND O1   C1
BOND O2   C4
BOND O2   H5
BOND N1   C1
BOND N1   H1
BOND N1   H2
BOND C1   C2
BOND C2   C3
```

```
BOND C2   C6
BOND C3   C5
BOND C3   H7
BOND C4   C5
BOND C4   C7
BOND C5   H6
BOND C6   C7
BOND C6   H3
BOND C7   H4

DIHEdral O1   C1   N1   H1
DIHEdral O1   C1   N1   H2
DIHEdral O1   C1   C2   C3
DIHEdral O1   C1   C2   C6
DIHEdral O2   C4   C5   C3
DIHEdral O2   C4   C5   H6
DIHEdral O2   C4   C7   C6
DIHEdral O2   C4   C7   H4
DIHEdral N1   C1   C2   C3
DIHEdral N1   C1   C2   C6
DIHEdral C1   C2   C3   C5
DIHEdral C1   C2   C3   H7
DIHEdral C1   C2   C6   C7
DIHEdral C1   C2   C6   H3
DIHEdral C2   C1   N1   H1
DIHEdral C2   C1   N1   H2
DIHEdral C2   C3   C5   C4
DIHEdral C2   C3   C5   H6
DIHEdral C2   C6   C7   C4
DIHEdral C2   C6   C7   H4
DIHEdral C3   C2   C6   C7
DIHEdral C3   C2   C6   H3
DIHEdral C3   C5   C4   C7
DIHEdral C4   C5   C3   H7
DIHEdral C4   C7   C6   H3
DIHEdral C5   C3   C2   C6
DIHEdral C5   C4   O2   H5
DIHEdral C5   C4   C7   C6
DIHEdral C5   C4   C7   H4
DIHEdral C6   C2   C3   H7
DIHEdral C7   C4   O2   H5
DIHEdral C7   C4   C5   H6
DIHEdral H3   C6   C7   H4
DIHEdral H6   C5   C3   H7


IMPRoper N1   C1   H1   H2
IMPRoper C1   N1   C2   O1
IMPRoper C2   C6   C3   C1
IMPRoper C3   C2   C5   H7
IMPRoper C4   C7   C5   O2
IMPRoper C5   C4   C3   H6
IMPRoper C6   C7   C2   H3
IMPRoper C7   C4   C6   H4

END {UNK}
```

Print 11: *UNK_par.cns.* CNS specific parameter file generated upon the VIDMAX.pdb file upload.

```
set echo=false end

checkversion 1.1


remark BONDS
bond O2C_ C__Y  687.326   1.222
bond O2C_ C__Y  687.326   1.222
bond OH1< CO_6 1347.773   1.355
bond OH1< HP__   53.174   0.889
bond NH2Y C__Y 1118.346   1.320
```

```
bond NH2Y HP__  104.167   0.889
bond C__Y CC_6  879.348   1.486
bond CC_6 CH_6 2052.018   1.389
bond CH_6 CH_6 1446.803   1.379
bond CH_6 HC__  289.588   0.956
bond CO_6 CH_6 1306.157   1.389

remark ANGLES
angle O2C_ C__Y NH2Y  457.739   122.350
angle O2C_ C__Y CC_6  211.087   120.885
angle OH1< CO_6 CH_6  287.579   119.848
angle NH2Y C__Y CC_6  290.817   118.900
angle C__Y NH2Y HP__  119.074   119.255
angle C__Y CC_6 CH_6  387.623   120.623
angle CC_6 CH_6 CH_6 1104.635   120.639
angle CC_6 CH_6 HC__  595.954   119.361
angle CH_6 CC_6 CH_6  965.717   118.369
angle CH_6 CH_6 CO_6 1152.059   119.926
angle CH_6 CH_6 HC__  477.778   119.918
angle CO_6 OH1< HP__   84.304   109.554
angle CO_6 CH_6 HC__  537.000   119.746
angle CH_6 CO_6 CH_6  797.215   120.221
angle HP__ NH2Y HP__   64.667   119.553

 remark DIHEDRALS
dihe O2C_ C__Y NH2Y HP__    5.0      3      180.0
dihe O2C_ C__Y NH2Y HP__    5.0      3      180.0
dihe O2C_ C__Y CC_6 CH_6  127.8      2        0.0
dihe OH1< CO_6 CH_6 CH_6  821.8      2      180.0
dihe OH1< CO_6 CH_6 HC__    5.0      3      180.0
dihe NH2Y C__Y CC_6 CH_6   34.0      2        0.0
dihe C__Y CC_6 CH_6 CH_6  904.3      2      180.0
dihe C__Y CC_6 CH_6 HC__    5.0      3      180.0
dihe CC_6 C__Y NH2Y HP__    5.0      3      180.0
dihe CC_6 CH_6 CH_6 CO_6  933.7      2        0.0
dihe CC_6 CH_6 CH_6 HC__    5.0      3      180.0
dihe CH_6 CC_6 CH_6 CH_6  973.1      2        0.0
dihe CH_6 CC_6 CH_6 HC__    5.0      3      180.0
dihe CH_6 CH_6 CO_6 CH_6  937.9      2        0.0
dihe CO_6 CH_6 CH_6 HC__    5.0      3      180.0
dihe CH_6 CO_6 OH1< HP__    5.0      3      180.0
dihe CH_6 CO_6 CH_6 HC__    5.0      3      180.0
dihe HC__ CH_6 CH_6 HC__    5.0      3      180.0

reamark IMPROPERS
impr NH2Y C__Y HP__ HP__   38.9     0      0.00
impr C__Y NH2Y CC_6 O2C_ 1700.0     0      0.00
impr CC_6 CH_6 CH_6 C__Y 1073.1     0      0.00
impr CH_6 CC_6 CH_6 HC__  646.8     0      0.00
impr CO_6 CH_6 CH_6 OH1< 1700.0     0      0.00
impr CH_6 CO_6 CH_6 HC__  619.6     0      0.00
impr CH_6 CH_6 CC_6 HC__  646.8     0      0.00

remark NONBONDED
NONBonded    O2C_    0.1270    2.2810    0.0790    1.4260
NONBonded    OH1<    0.1270    2.2810    0.0790    1.4260
NONBonded    NH2Y    0.1900    2.2810    0.1190    1.4260
NONBonded    C__Y    0.0960    3.0400    0.0600    1.9000
NONBonded    CC_6    0.0960    3.0400    0.0600    1.9000
NONBonded    CH_6    0.0960    3.0400    0.0600    1.9000
NONBonded    CO_6    0.0960    3.0400    0.0600    1.9000
NONBonded    HP__    0.0400    1.1200    0.0250    0.7000
NONBonded    HC__    0.0400    1.1200    0.0250    0.7000
```

Print 12: *UNK.cif*. REFMAC specific topology and parameter file generated upon the VIDMAX.pdb file upload.

```
global_
_lib_name         mon_lib
_lib_version      4.3
```

```
_lib_update       11/06/03
# ------------------------------------------------
#
# ---   LIST OF MONOMERS ---
#
data_comp_list
loop_
_chem_comp.id
_chem_comp.three_letter_code
_chem_comp.name
_chem_comp.group
_chem_comp.number_atoms_all
_chem_comp.number_atoms_nh
_chem_comp.desc_level
UNK      UNK 'VIDMAX                                ' non-polymer        17  10 .
#
# --- DESCRIPTION OF MONOMERS ---
#
data_comp_UNK
#
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.type_energy
_chem_comp_atom.partial_charge
_chem_comp_atom.x
_chem_comp_atom.y
_chem_comp_atom.z
 UNK           O1     O    O2C_     -0.472      0.000    0.000    0.000
 UNK           C1     C    C__Y      0.020     -0.355    0.938   -0.753
 UNK           N1     N    NH2Y     -0.137     -1.197    1.879   -0.329
 UNK           H2     H    H         0.201     -1.513    1.842    0.503
 UNK           H1     H    H         0.201     -1.435    2.536   -0.882
 UNK           C2     C    CC_6      0.000      0.166    1.009   -2.144
 UNK           C6     C    CH_6     -0.041     -0.292    1.932   -3.073
 UNK           H3     H    H         0.164     -0.949    2.578   -2.819
 UNK           C7     C    CH_6     -0.042      0.202    1.920   -4.372
 UNK           H4     H    H         0.163     -0.123    2.552   -5.011
 UNK           C4     C    CO_6     -0.000      1.160    0.999   -4.743
 UNK           O2     O    OH1<     -0.520      1.668    0.968   -6.009
 UNK           H5     H    H         0.205      1.280    1.598   -6.503
 UNK           C5     C    CH_6     -0.038      1.640    0.089   -3.819
 UNK           H6     H    H         0.165      2.313   -0.542   -4.069
 UNK           C3     C    CH_6     -0.042      1.139    0.099   -2.532
 UNK           H7     H    H         0.168      1.469   -0.534   -1.896
loop_
_chem_comp_tree.comp_id
_chem_comp_tree.atom_id
_chem_comp_tree.atom_back
_chem_comp_tree.atom_forward
_chem_comp_tree.connect_type
 UNK       O1     n/a    C1     START
 UNK       C1     O1     C2     .
 UNK       N1     C1     H1     .
 UNK       H2     N1     .      .
 UNK       H1     N1     .      .
 UNK       C2     C1     C6     .
 UNK       C6     C2     C7     .
 UNK       H3     C6     .      .
 UNK       C7     C6     C4     .
 UNK       H4     C7     .      .
 UNK       C4     C7     C5     .
 UNK       O2     C4     H5     .
 UNK       H5     O2     .      .
 UNK       C5     C4     C3     .
 UNK       H6     C5     .      .
 UNK       C3     C5     H7     .
 UNK       H7     C3     .      END
 UNK       C2     C3     .    ADD
loop_
```

```
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.type
_chem_comp_bond.value_dist
_chem_comp_bond.value_dist_esd
 UNK      O1     C1        coval       1.222    0.029
 UNK      O2     C4        coval       1.355    0.021
 UNK      O2     H5        coval       0.889    0.105
 UNK      N1     C1        coval       1.320    0.023
 UNK      N1     H1        coval       0.889    0.075
 UNK      N1     H2        coval       0.889    0.075
 UNK      C1     C2        coval       1.486    0.026
 UNK      C2     C3        coval       1.389    0.017
 UNK      C2     C6        coval       1.389    0.017
 UNK      C3     C5        coval       1.379    0.020
 UNK      C3     H7        coval       0.956    0.045
 UNK      C4     C5        coval       1.389    0.021
 UNK      C4     C7        coval       1.389    0.021
 UNK      C5     H6        coval       0.956    0.045
 UNK      C6     C7        coval       1.379    0.020
 UNK      C6     H3        coval       0.956    0.045
 UNK      C7     H4        coval       0.956    0.045
loop_
_chem_comp_angle.comp_id
_chem_comp_angle.atom_id_1
_chem_comp_angle.atom_id_2
_chem_comp_angle.atom_id_3
_chem_comp_angle.value_angle
_chem_comp_angle.value_angle_esd
 UNK      O1     C1     N1      122.350    2.061
 UNK      O1     C1     C2      120.885    3.034
 UNK      O2     C4     C5      119.848    2.600
 UNK      O2     C4     C7      119.848    2.600
 UNK      N1     C1     C2      118.900    2.585
 UNK      C1     N1     H1      119.255    4.040
 UNK      C1     N1     H2      119.255    4.040
 UNK      C1     C2     C3      120.623    2.239
 UNK      C1     C2     C6      120.623    2.239
 UNK      C2     C3     C5      120.639    1.326
 UNK      C2     C3     H7      119.361    1.806
 UNK      C2     C6     C7      120.639    1.326
 UNK      C2     C6     H3      119.361    1.806
 UNK      C3     C2     C6      118.369    1.419
 UNK      C3     C5     C4      119.926    1.299
 UNK      C3     C5     H6      119.918    2.017
 UNK      C4     O2     H5      109.554    4.801
 UNK      C4     C5     H6      119.746    1.902
 UNK      C4     C7     C6      119.926    1.299
 UNK      C4     C7     H4      119.746    1.902
 UNK      C5     C3     H7      119.918    2.017
 UNK      C5     C4     C7      120.221    1.561
 UNK      C6     C7     H4      119.918    2.017
 UNK      C7     C6     H3      119.918    2.017
 UNK      H1     N1     H2      119.553    5.482
loop_
_chem_comp_tor.comp_id
_chem_comp_tor.id
_chem_comp_tor.atom_id_1
_chem_comp_tor.atom_id_2
_chem_comp_tor.atom_id_3
_chem_comp_tor.atom_id_4
_chem_comp_tor.value_angle
_chem_comp_tor.value_angle_esd
_chem_comp_tor.period
 UNK      tor1     O1     C1     N1     H1       180.000   20.000   3
 UNK      tor2     O1     C1     N1     H2       180.000   20.000   3
 UNK      tor3     O1     C1     C2     C3         0.000    1.985   2
 UNK      tor4     O1     C1     C2     C6         0.000    1.985   2
 UNK      tor5     O2     C4     C5     C3       180.000    1.068   2
 UNK      tor6     O2     C4     C5     H6       180.000   20.000   3
```

```
 UNK      tor7     O2     C4     C7     C6       180.000    1.068   2
 UNK      tor8     O2     C4     C7     H4       180.000   20.000   3
 UNK      tor9     N1     C1     C2     C3         0.000    3.086   2
 UNK      tor10    N1     C1     C2     C6         0.000    3.086   2
 UNK      tor11    C1     C2     C3     C5       180.000    1.034   2
 UNK      tor12    C1     C2     C3     H7       180.000   20.000   3
 UNK      tor13    C1     C2     C6     C7       180.000    1.034   2
 UNK      tor14    C1     C2     C6     H3       180.000   20.000   3
 UNK      tor15    C2     C1     N1     H1       180.000   20.000   3
 UNK      tor16    C2     C1     N1     H2       180.000   20.000   3
 UNK      tor17    C2     C3     C5     C4         0.000    1.023   2
 UNK      tor18    C2     C3     C5     H6       180.000   20.000   3
 UNK      tor19    C2     C6     C7     C4         0.000    1.023   2
 UNK      tor20    C2     C6     C7     H4       180.000   20.000   3
 UNK      tor21    C3     C2     C6     C7         0.000    1.009   2
 UNK      tor22    C3     C2     C6     H3       180.000   20.000   3
 UNK      tor23    C3     C5     C4     C7         0.000    1.022   2
 UNK      tor24    C4     C5     C3     H7       180.000   20.000   3
 UNK      tor25    C4     C7     C6     H3       180.000   20.000   3
 UNK      tor26    C5     C3     C2     C6         0.000    1.009   2
 UNK      tor27    C5     C4     O2     H5       180.000   20.000   3
 UNK      tor28    C5     C4     C7     C6         0.000    1.022   2
 UNK      tor29    C5     C4     C7     H4       180.000   20.000   3
 UNK      tor30    C6     C2     C3     H7       180.000   20.000   3
 UNK      tor31    C7     C4     O2     H5       180.000   20.000   3
 UNK      tor32    C7     C4     C5     H6       180.000   20.000   3
 UNK      tor33    H3     C6     C7     H4       180.000   20.000   3
 UNK      tor34    H6     C5     C3     H7       180.000   20.000   3
loop_
_chem_comp_chir.comp_id
_chem_comp_chir.id
_chem_comp_chir.atom_id_centre
_chem_comp_chir.atom_id_1
_chem_comp_chir.atom_id_2
_chem_comp_chir.atom_id_3
_chem_comp_chir.volume_sign
 UNK      chir_01  N1     C1     H1     H2       negativ
 UNK      chir_02  C1     N1     C2     O1       negativ
 UNK      chir_03  C2     C6     C3     C1       negativ
 UNK      chir_04  C3     C2     C5     H7       negativ
 UNK      chir_05  C4     C7     C5     O2       negativ
 UNK      chir_06  C5     C4     C3     H6       negativ
 UNK      chir_07  C6     C7     C2     H3       negativ
 UNK      chir_08  C7     C4     C6     H4       negativ
loop_
_chem_comp_plane_atom.comp_id
_chem_comp_plane_atom.plane_id
_chem_comp_plane_atom.atom_id
_chem_comp_plane_atom.dist_esd
 UNK      plan1     N1        0.020
 UNK      plan1     C1        0.020
 UNK      plan1     H1        0.020
 UNK      plan1     H2        0.020
 UNK      plan2     C1        0.020
 UNK      plan2     N1        0.020
 UNK      plan2     C2        0.020
 UNK      plan2     O1        0.020
 UNK      plan3     C2        0.020
 UNK      plan3     C6        0.020
 UNK      plan3     C3        0.020
 UNK      plan3     C1        0.020
 UNK      plan4     C3        0.020
 UNK      plan4     C2        0.020
 UNK      plan4     C5        0.020
 UNK      plan4     H7        0.020
 UNK      plan5     C4        0.020
 UNK      plan5     C7        0.020
 UNK      plan5     C5        0.020
 UNK      plan5     O2        0.020
 UNK      plan6     C5        0.020
 UNK      plan6     C4        0.020
```

```
 UNK      plan6     C3        0.020
 UNK      plan6     H6        0.020
 UNK      plan7     C6        0.020
 UNK      plan7     C7        0.020
 UNK      plan7     C2        0.020
 UNK      plan7     H3        0.020
 UNK      plan8     C7        0.020
 UNK      plan8     C4        0.020
 UNK      plan8     C6        0.020
 UNK      plan8     H4        0.020
# -------------------------------------------------------
# -------------------------------------------------------
# -------------------------------------------------------
```

List of Publications

Andrejašič M., Pražnikar J, and Turk D. (2008): PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. *Acta Cryst* D**64,**1093-1109.

Plantan I., Selič L., Mesar T., Anderluh P.S., Oblak M., Preželj A., Hesse L., Andrejašič M., Vilar M., Turk D., Kocijan A., Prevec T., Vilfan G., Kocjan D., Čopar A., Urleb U., Šolmajer T. (2007): 4-Substituted trinems as broad spectrum beta-lactamase inhibitors: structure-based design, synthesis, and biological activity. *J Med Chem.* 2007 **50**, 4113-4121.

Štefanić Z., Vujaklija D., Andrišić L., Mikleušivić G. Andrejašič M., Turk D. and Luić M. (2007): Preliminary crystallographic study of Streptomyces coelicolor single-stranded DNA-binding protein. *Croat. chem. acta*, 2007, **80**, str. 35-39.