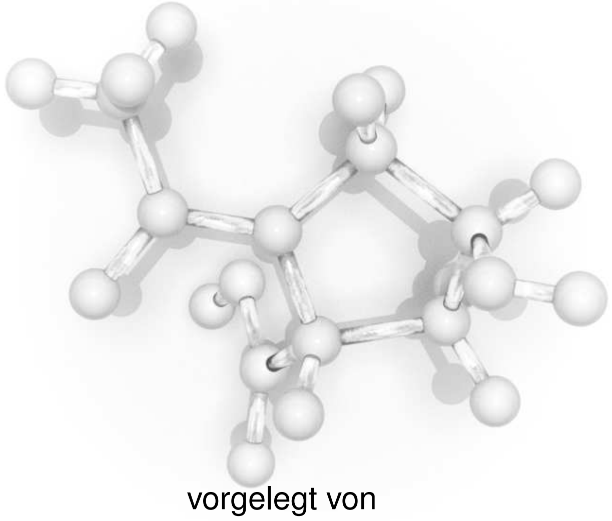
The Role of Hydrogen Atoms and Thermal Displacement Parameters in Crystal Structure Refinement

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm Chemie
der Georg-August-University School of Science (GAUSS)



vorgelegt von

Jens Lübben

aus Oldenburg

Göttingen, 2017

Betreuungsausschuss

Prof. George Sheldrick
Institut für Anorganische Chemie, Georg-August-Universität Göttingen

PD Dr. Birger Dittrich
Institut für Anorganische Chemie und Strukturchemie, Heinrich-Heine-Universität Düsseldorf

Mitglieder der Prüfungskommission

Prof. George Sheldrick
Institut für Anorganische Chemie, Georg-August-Universität Göttingen

PD Dr. Birger Dittrich
Institut für Anorganische Chemie und Strukturchemie, Heinrich-Heine-Universität Düsseldorf

Weitere Mitglieder der Prüfungskommission

Prof. Dr. Ricardo Mata
Institut für Physikalische Chemie, Georg-August-Universität Göttingen

PD Dr. Heidrun Sowa
Geowissenschaftliches Zentrum, Georg-August-Universität Göttingen

Prof. Dr. Dietmar Stalke
Institut für Anorganische Chemie, Georg-August-Universität Göttingen

Prof. Dr. Hartmut Laatsch
Institut für Organische und Biomolekulare Chemie, Georg-August-Universität Göttingen

“With magic, you can turn a frog into a prince. With science, you can turn a frog into a Ph.D and you still have the frog you started with.”

– Terry Pratchett

Contents

List of Figures

Acronyms

ADP atomic displacement parameters

APD anisotropic proton displacement

ARG attached rigid group

BEEF bond enhanced evaluation factor

HAR Hirshfeld atom refinement

HDD hydrogen density distribution

IAM independent atom model

SHADE simple hydrogen ADP estimator

TLS translation-libration-screw

XRD X-ray diffraction

Part I.

Introduction

Introduction

Single crystal X-ray diffraction (XRD) is a powerful and well established technique for analyzing a chemical compound's make-up by determining the positions of atoms in space based on a recorded diffraction pattern. Since the first application of this technique more than a hundred years ago[1], experimental and data processing techniques have improved steadily, making XRD a quick and easy to use tool today. XRD yields valuable results for physicists, biologists and chemists alike and led to great discoveries in many scientific disciplines.[2]

This thesis aims to continue improving the method in order to extend its application to cases that were not easily treatable before. It does so by combining experimental techniques with theoretical computations. For instance hydrogen atoms have characteristics that make modeling by established techniques challenging. Most commonly, the modeling challenges are worked around by estimating the hydrogen atoms' parameters based on the direct chemical bonding partners. The first focus of this thesis is to assert the validity of this approach, expand on it to allow a more elaborate parameterization to increase the estimation accuracy for more extreme experimental conditions, and investigate its implication on the overal model quality.

The downside of a technique as readily available as XRD is that inexperienced users can misjudge experimental data and misinterpret the structural information obtained. To make information obtained by XRD as reliable as possible, validation protocols are required to procedurally check structure models for consistency and correctness. Even though powerful validation protocols already exist, they are not routinely applicable in all fields of crystallography.

The second part of this thesis explores methods to validate results obtained by structure-model optimization against experimental diffraction data. Experimentally ob-

[1] Max von Laue was awarded the Nobel prize for *Diffraction of X-rays by crystals* in 1914.

[2] So far 29 Nobel prizes were awarded to researchers involved in the development of crystallographic methods or their application including F. Crick, J. Watson and M. Wilins for *The helical structure of DNA* (1962), D. Hodgkin for *Structure of many biochemical substances including Vitamin B12* (1964), H. Hauptman and J. Karle for *Development of direct methods for the determination of crystal structures* (1985) and V. Ramakrishnan, T. A. Steitz and A. E. Yonath for *Studies of the structure and function of the ribosome* (2009).

tained data is affected by systematic and random errors that can influence the structure optimization procedure. This thesis aims to improve existing validation protocols – especially in the context of the proposed modeling techniques – to make them more sensitive to potential modeling errors and to simplify their application in the most common scenarios. The modeling of hydrogen atoms is particularly prone to errors due to overfitting of the imperfect experimental data which ties this part closely to the first part of the thesis. The proposed methods aim to aid inexperienced researchers in their interpretation of recorded data while providing a toolbox for experienced researchers to quickly detect critical parameters during the structure refinement procedure.

Both major parts are organized into chapters, each discussing an individual research project. Some of the later chapters will reference methods presented in previous chapters. Each chapter is split into two main sections. The *Methods* sections discuss experimental or analytical methods applied in that chapter. The *Results* sections present and discuss results obtained with the previously introduced methods. The organizational structure of a *Methods* section and the corresponding *Results* section are similar. The appendices contain detailed information on how to reproduce results presented in this thesis.

1. Experimental Techniques

The studies discussed in this thesis heavily rely on experimental data. The experimental techniques employed are described and discussed in this section.

1.0.1. Single Crystal X-ray Diffraction

Single crystal XRD is the central technique used in all studies discussed in this thesis. Single crystal XRD is an experimental technique were single crystals are irradiated by an X-ray beam which is scattered by the periodic lattice of the single crystal. The scattered beam is detected and then used to reconstruct the composition of the crystal. The basics of XRD will not be further discussed in this thesis and are assumed to be familiar to the reader.[1] Instead, this section will focus on the application of XRD in the context of the performed studies with a focus on limitations of the technique and how to overcome them.

Scattering by Hydrogen Atoms

The first focus of this thesis is the modeling of hydrogen atoms in XRD studies. X-rays are scattered by the electrons in the crystal lattice. The electron density in the vicinity of hydrogen atoms is comparably low due to the fact that hydrogen atoms only contain one single electron. This means that accurate data on hydrogen atoms is difficult to obtain by XRD and requires highly redundant diffraction data of high resolution. The need for high resolution seems counter intuitive since high-resolution data contains no information about hydrogen atoms. However, if the positions of heavier atoms are well defined by the high resolution data, the low resolution data can be used to refine the hydrogen atom parameters almost exclusively. The lack of core electrons of hydrogen leads to another challenge. The independent atom model (IAM) assumes that the electron density of an atom is spherical with its centroid at the position of the atomic

[1] The text books by Massa (1996), Giacovazzo *et al.* (1992), Luger (1980), Dunitz (1979) and Rupp (2009) provide excellent introductions to the field of crystallography. The text book by Müller *et al.* (2006) gives a more practical introduction to crystal structure analysis.

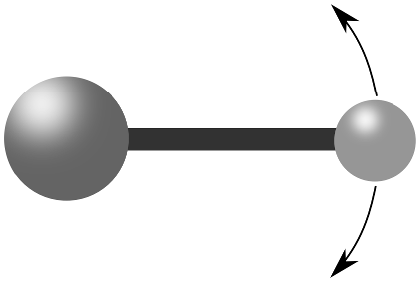


Figure 1.1.: Illustration of librational motion of terminally bonded atoms.

nucleus. However, hydrogen atoms only have a single electron that interacts with the hydrogen atom's bonding partner. This interaction moves the electron density away from the atomic nucleus which implies that application of the IAM does not yield the correct atomic position parameters. The impact of this effect on the structure model can be mitigated by using modified scattering factors (Stewart *et al.*, 1965). This can be avoided by optimizing a single bond oriented dipole which on the other hand adds a significant amount of parameters to a section of the model that is already not well defined. Also, it is not compatible with the most commonly applied scattering factor model – the IAM.

An additional problem associated with the determination of hydrogen atom positions by refinement against XRD data is the vibrational behavior of terminally bonded atoms. A significantly populated vibrational mode of terminally bonded atoms is the librational motion of the terminal atom relative to its bonding partner (Figure 1.1). However, commonly applied displacement models describe atomic motion in an orthogonal basis that is not able to parametrize librational motion. This implies that experimentally determined X–H bond distances are artificially shortened by a small amount (Cruickshank, 1956*b*).

1.0.2. Single Crystal Neutron Diffraction

The second experimental method employed to obtain results discussed in this thesis is single crystal Neutron diffraction. In contrast to XRD a beam of coherent Neutrons is scattered by the nuclei of the atoms in the crystal lattice. Atomic nuclei in a molecule do not interact with each other in any significant way. All interactions between atoms are mediated by electrons which do not contribute to the scattering of the Neutron beam. As a result, the independent atom approximation is much better fulfilled for scattered Neutrons than for scattered photons implying that the measured nuclear density directly correlates with the actual density and is not biased by chemical bonding and similar density deforming effects. Atomic nuclei are orders of magnitude smaller than the electron cloud scattering the photon beam. This means that the scattered amplitudes are virtually independent of the scattering angle.[2] Overall, this implies that positional and vibrational parameters determined via Neutron diffraction are more reliable than equivalent parameters optimized against XRD data. The biggest advantage of Neutron diffraction over XRD in the context of this work is the scattering length of hydrogen atoms. While hydrogen atoms are almost invisible to X-rays due to their limited number of electrons, their Neutron scattering length is significant and can even be improved by substituting hydrogen with Deuterium.

Neutron diffraction has disadvantages as well. Neutron sources with sufficiently high flux to facilitate diffraction experiments are expensive to build and to operate. While X-ray sources suitable for XRD experiments fit in a normal laboratory and take only a few square meters of space and a single person to operate, suitable Neutron sources require nuclear reactors or spallation facilities which require *whole* organizations to build, maintain and operate. Also, even modern high-flux Neutron sources require significantly bigger crystals to perform diffraction experiments in reasonable amounts of time due to the low interaction probability of Neutrons with the atomic nuclei. This makes Neutron diffraction experiments expensive and difficult to schedule and in some cases even impossible due to the required crystal sizes that are not always possible to obtain.

Considering the limitations of Neutron diffraction it has been proven useful to perform Neutron diffraction experiments for a small, carefully selected set of structures and use those structures as references to optimize methods to overcome the limitations of XRD while still maintaining its advantages over Neutron diffraction.

Quasi Laue Diffraction

The most commonly applied experimental setups for single crystal diffraction experiments utilize a monochromatic primary beam. This provides the significant advantage that every measured intensity is associated with one discrete and known wavelength value. It is also possible to perform the experiment with multiple wavelengths. This technique is called *Laue diffraction* if the whole spectrum is used or *Quasi Laue diffraction* if a wavelengths distribution between λ_{min} and λ_{max} is used (Wilkinson and Lehmann,

[2] The scattering angle independence of the diffraction angle is only true for the scatterer at rest. During the experiment atomic nuclei are displaced by thermal motion thereby creating an effective dependence of the scattering amplitude on the scattering angle. In practice this means that a dependence on the scattering angle is observed but is less pronounced than it is for XRD.

1991).[3] A major advantage of this technique is a dramatically reduced data acquisition time. Instead of collecting adjacent fine slices of reciprocal space while slowly rotating the sample as is needed for monochromatic diffraction, each recorded image contains a potentially large part of reciprocal space.[4] This implies that fewer images need to be recorded. A quasi complete data set collected with a monochromatic beam can contain thousands of images. A comparable Laue diffraction data set can consist of less than twenty images. Assuming constant time for recording an image, the reduced data acquisition time can be significant. This is especially critical for Neutron diffraction experiments where the acquisition time for a single image can be days instead of fractions of a second as for XRD experiments.

Laue diffraction has disadvantages over monochromatic data collection as well. Laue diffraction images contain diffracted intensities corresponding to different wavelengths. The position of a reflection in reciprocal space depends on the cell parameters and the energy of the diffracted photons. Performing the deprojection[5] of the diffracted intensities from the detector plane to reciprocal space requires knowledge of both the cell parameters and a reflection's corresponding energy. However, the energy is generally not known which complicates the deprojection process.[6] In practice, this means that cell parameters must be known beforehand and cannot be determined *ab initio* when performing the deprojection. This implies that preliminary experiments must be performed to process the diffraction data.

An additional problem connected to the fact that three dimensional reciprocal space is projected onto a two dimensional detector plane is that different points in reciprocal space end up at the same position on the projection plane thereby making them indistinguishable and effectively making them unusable for further processing steps. In the context of diffraction experiments reflection A with λ_A and B with λ_B are projected onto the same point if the Miller indices of B are multiples of the Miller indices of A and λ_A is

[3] (Quasi) Laue diffraction is an experimental method that is not specific to Neutron diffraction. However, in the context of this thesis the method is exclusively applied for Neutron diffraction experiments and is therefore briefly discussed in this section.

[4] The size of reciprocal space recorded at once depends on the bandwidth of the primary beam's spectrum.

[5] This process is commonly called *indexing* which is effectively the deprojection of the recorded two dimensional image to three dimensional reciprocal space. The term *deprojection* is used here to illustrate what steps *indexing* effectively involves.

[6] Experimental techniques to record the energy in tandem with the intensity exist in the form of Time of Flight Laue diffraction. This technique can solve most of the problems currently associated with Laue diffraction but the technique is not readily available yet. The additional complexity of data collection also introduces additional sources of error. Effectively, the data quality of Time of Flight data is comparable to wavelength indiscriminate recording techniques today.

the same multiple of λ_B. For example reflection $A = (1, 0, 0)$ and $B = (2, 0, 0)$ are indistinguishable if $\lambda_A = 2\lambda_B$. This implies that significant parts of reciprocal space are not accessible by this experimental method resulting in low values for data completeness.

Data processing is further complicated by varying flux across the Neutrons' energy spectrum. Neutron sources do not generate the same amount of Neutrons of each energy. This results in an additional energy-dependent scaling factor that needs to be determined. In practice, the energy resolved flux of the primary beam can be monitored during data collection and applied to the integrated data after each reflection is associated with one energy.

2. Applied Structure Modeling Techniques

This chapter describes the scattering-factor models applied in the studies presented in this thesis. Only brief overviews for each model are given mostly focusing on differences between them and the prerequisites necessary for successful application. The presented scattering factors are descriptions of an atomic scatterer at rest. Models for treating atomic motion are basically independent of the scattering-factor expression and are discussed in the last section of this chapter.

2.1. Independent Atom Model

The IAM is the most simple and most commonly applied scattering-factor model in crystallography. In this context *simple* does not imply that only few parameters are used to describe the model. Macromolecular crystallographers often employ a rigid group model that requires fewer parameters to describe the whole model but effectively uses composite scattering factors consisting of multiple IAM scattering factors. This is usually realized by constraining relative atomic positions of a molecular building block and only optimizing one set of positional parameters plus one set of orientational parameters for the whole group.[1] Rather, *simple* means that each chemical element (plus its ionization states) is represented by one scattering factor.

In the IAM it is assumed that atoms in the crystal lattice are independent of each other and do not interact. Even though the model does not consider atomic interaction, some information about these interactions can still be derived from the model based on inter-atomic distances.

The IAM usually describes the scattering contribution of an atom as a superposition of four Gaussian functions plus one constant factor (Rupp, 2009).[2]

$$f_s^0 = \sum_{i=1}^{4} a_i \cdot exp\left(-b_i \left(\frac{sin\Theta}{\lambda}\right)^2\right) + c \tag{2.1}$$

[1] In practice, this can be realized with restraints as well, providing a more flexible model.

[2] Other IAM implementations use a different sum (Rez *et al.*, 1994). Implementations relevant to this work follow the described approach.

The atom specific parameters a_i, b_i and c are optimized against Hartree-Fock (Jensen, 1994) wave functions and stored in a look-up table. Θ is the scattering angle and λ the wavelength of the diffracted beam. The overall crystal structure is then approximated by placing the appropriate atomic scattering factor at the correct position in the crystal's coordinate system. Since f_s^0 is independent of the orientation of the atom, the model has no means of describing inter-atomic interaction other than analyzing the spacial overlap of spherically symmetrical scatterers.

The main advantages of the IAM are that it is straight-forward to implement and requires the optimization of only three positional parameters for each atom. The main disadvantage is that it provides only rough approximations of structural properties.

2.2. Multipole Model

The multipole model is a modification to the IAM to take spacial anisotropy of an atom's electron density into account. Anisotropy is parametrized via spherical harmonics that depend on the angle to an appropriately chosen reference orientation (Stewart, 1969, 1976). This is commonly implemented as suggested by Hansen and Coppens (1978) by splitting an atom's IAM scattering factor into two parts: the core electrons, treated as non-interacting density, and the valence shell. The valence shell density has a variable amplitude and is deformed by a series of spherical harmonics. The atom's full electron density is then the core density plus the deformed valence density yielding

$$\rho\left(r\right) = P_{core}\rho_{core}\left(r\right) + P_{val}\kappa^3\rho_{val}\left(\kappa r\right) + \sum_{l=0}^{l_{max}} \kappa_l'^3 R_l\left(\kappa_l' r\right) \sum_{m=0}^{l} P_{lm\pm} Y_{lm\pm}\left(\Omega\right) \tag{2.2}$$

with the occupancies P_i and the expansion/contraction parameters κ^j as refinable parameters.[3] The absolute number of parameters used to describe one atom's is now dependent on l_{max}. Even with $l_{max} = 1$ five more parameters than used in the IAM must be optimized. Modern implementations of the multipole model include spherical harmonics up to $l_{max} = 4$. This implies that multipole refinement can only be applied successfully when very precise and accurate diffraction data is available. And even then, appropriate constraints must be chosen carefully to reduce the number of parameters to a manageable degree. The very flexible parametrization of the multipole scattering factors can also lead to strong correlation between parameters, thereby further complicating the optimization procedure. In practice this often means that some form

[3] Similar to the IAM, several multipole implementations with subtle differences exist.

of tailor-made block refinement and tailor-made parametrization models are designed specifically for a particular crystal structure.

However, if refinement yields reliable results, the multipole model provides significant advantages over a comparable IAM model. The modeling of lone-pair populations and bonding electron density provides insight into the electronic structure of molecules in the crystal lattice, thereby allowing evaluation of bonding situations and generally much more precise parameter estimation (Kratzert *et al.*, 2013). Since aspherical density is taken into account, the model will also provide better estimates for bond distances and vibrational characteristics.

2.3. Invariom Model

The invariom model is not a scattering-factor model by itself (Dittrich *et al.*, 2013). Although it is commonly implemented with the multipole model, it can theoretically be applied to any form of scattering-factor model. Strictly speaking, the invariom model defines a method for parametrizing the local chemical environment of an atom and provides means to transfer scattering factors to chemically equivalent environments. This implies that the scattering factor of an atom – independent of the scattering factor itself – can be transferred to atoms in equivalent chemical environments. This provides the advantage that the scattering factor of an arbitrary atom can be determined under the most ideal circumstances and then be transferred to a system that would not allow for the determination of the scattering factor in itself.

In practice, the invariom model is usually combined with the invariom database – a collection of idealized chemical environments that facilitate the determination of scattering factors. The invariom database contains quantum chemically optimized structural models of small molecules. These models are used to generate electron density maps which are subsequently Fourier transformed to obtain artificial diffraction data that is free of experimental errors.[4]

In conclusion, the invariom model benefits from many of the advantages of an aspherical scattering-factor model, like the multipole model, without the need to introduce and optimize additional parameters. On the other hand, the invariom model has its

[4] Note that this does not mean that the data is free of errors altogether. The methods used to compute the data make their own approximations and the scattering-factor model optimized against the generated data is not free of inaccuracies itself. For example the overall error includes basis set errors, approximations in the Hartree-Fock method, approximations in the applied density functional theory, Fourier truncation error when converting the density to frequency space and a finite multipole expansion when optimizing scattering factors against the frequency data.

weaknesses too. As with all constraints – which invariom scattering factors effectively are – the resulting structural model must be interpreted carefully. Information that was put into the model via constraints must not be interpreted freely. If the electron density near an atom is of interest to the researcher, it is advisable to chose a hybrid approach that uses an invariom model for the bulk of the parameters and freely optimizes the parameters relevant for answering the researchers questions. Another shortcoming is the invariom database itself. Nature finds ways to combine chemical elements very creatively thereby generating figuratively infinite numbers of chemical environments. Tabulating all of them in a database is not feasible and, depending on the chemical elements involved, not possible in a consistent way with today's quantum chemical toolbox.[5]

2.4. Hirshfeld Atom Refinement

Hirshfeld atom refinement (HAR) uses a similar approach to scattering factor determination to the invariom model. Instead of pre-computing approximate scattering factors and transferring them from a database, HAR generates scattering factors on-the-fly via quantum chemical computations and iteratively repeats computations and structure refinement to self-consistency (Capelli *et al.*, 2014). HAR performs the following steps:

1. Generate a starting model. Usually the result of an IAM refinement.
2. Generate the electron density corresponding to the model geometry via quantum-chemical methods.
3. Partition the electron density in a way that assigns each voxel partially to atoms contributing to the density at that voxel (Hirshfeld, 1977).
4. Convert the partitioned density into atom-specific scattering factors.
5. Refine model parameters against the measured data using these tailor-made scattering factors.
6. Repeat steps 2–5 to self-consistency.

This approach avoids the challenge of tabulating enormous amounts of chemical environments because the scattering factors are generated specifically for the molecule provided as input.

[5] The basis set currently used is not available for all chemical elements. Suitable basis sets must not use the frozen-core approximation. This makes treatment of heavier elements challenging.

A significant shortcoming of HAR is the modeling of disorder in structures. Disorder is commonly modeled by using partial occupancies for scattering factors effectively multiplying a given scattering factor by a positive number smaller than one. HAR is based on quantum chemical methods to obtain electron densities. This does not allow partial nuclei or partial electrons, thus limiting the possibilities of the method. However in practice, disordered structures are rarely modeled with aspherical scattering models anyway. Another challenge for HAR is the optimization of large structures. The computation of electron densities based on quantum chemical methods scales very unfavorably with the size of the system, resulting in overall long computation times for larger molecules compared to other modeling techniques.

2.5. Modeling of Thermal Vibrations

The scattering-factor models discussed in the previous section describe the scattering contribution of an atom at rest. In a real structure however, atoms get displaced from their equilibrium position. Even at temperatures close to $0\ \mathrm{K}$ zero point vibrations will still affect an atom's position over time. Accounting for that atomic motion is crucial for modeling crystal structures because the displacement reduces the crystals periodicity which in turn affects the scattered beam intensities. Several different models to parametrize atomic vibration in crystal structures exist. The most common ones are discussed in this section.

2.5.1. Isotropic Displacement

An isotropic displacement model is the most simple model for parametrizing atomic vibrations (Grosse-Kunstleve and Adams, 2002). It is based on the approximation that an atom in a crystal structure behaves like a harmonic oscillator with equal force constants for all spacial dimensions, hence *isotropic*. While this is clearly a very rough approximation – atoms don't behave like harmonic oscillators nor is it reasonable to assume that the force constant is independent of its surroundings – the model has the critical advantage of requiring only one parameter to be optimized for each atom: the displacement amplitude. In cases where the data to parameter ratio is low, the data is noisy, parts of a structure are disordered or simply very unpronounced electron density regions are modeled, it is crucial to use as few parameters as possible to keep the refinement stable and to avoid overfitting.

The effect of isotropic displacement on the IAM scattering factor F_S^0 can be described with

$$f_S^B = f_S^0 \cdot e^{-B_{iso}(sin\Theta/\lambda)^2} \tag{2.3}$$

where

$$B_{iso} = 8\pi^2 \left\langle u_{iso}^2 \right\rangle . \tag{2.4}$$

Θ is the scattering angle, λ the wavelength and u_{iso} the amplitude of the harmonic oscillator. As equation 2.3 shows, the effect of thermal motion on the scattering amplitude depends on the scattering angle which makes the inclusion of a displacement term necessary for modeling crystal structures.

2.5.2. Anisotropic Displacement

The anisotropic displacement model (Cruickshank, 1956*a*) is a more detailed and more flexible model of atomic vibration. As the name suggests, the model introduces anisotropy to the harmonic oscillator used to describe atomic motion. This means that instead of one force constant that describes the force needed to displace an atom from its equilibrium position, three force constants are introduced where the direction of displacement determines which force constant is relevant.[6] Common visualization techniques describe atomic displacement parameters (ADP) as an ellipsoid where the lengths of the principle axes correspond to the force constants and the directions of an axis encode their orientation.[7] The anisotropic model adds six parameters to the resting scattering-factor model. The parameters can be interpreted as follows:

- Three parameters encode the direction of the first principle component and its lengths encodes the corresponding force constant.
- Assuming a right-handed orthogonal coordinate system, the second direction is constrained to the plane perpendicular to the first direction. This implies that only two parameters are required to encode the second direction. Again, the lengths of this two dimensional vector encodes the force constant.
- Using a right-handed coordinate system, the third directional vector is the cross

[6] The anisotropic discplacement model does not use force constants as parameters directly. Instead, the force constants are encoded as mean-squared discplacement amplitudes. The term *force constant* is used here for illustration purposes.

[7] Note that the absolute size of a displacement ellipsoid depends on an arbitrarily chosen probability value determining how likely it is that the atomic nucleus can be found within the ellipsoid at a given point in time.

product of the first two normalized directional vectors. This implies that no additional parameters are required to encode its direction and only one parameter – the third force constant – is needed.

- The six parameters are then reorganized into a symmetric 3×3 matrix U^{ij} where the diagonal elements encode the displacement amplitude and the off-diagonal encode the orientation of the displacement axes.

Applying this displacement model yields the following expression for the scattering factor:

$$f_S^A = f_S^0 \cdot e^{-2\pi^2\left(U^{11}h^2a^{*2}+U^{22}k^2b^{*2}+U^{33}l^2c^{*2}+2U^{23}klb^*c^*+2U^{13}hla^*c^*+2U^{12}hka^*b^*\right)} \tag{2.5}$$

U^{ij} are the ADP. h, k and l are the Miller indices. a^*, b^* and c^* are the reciprocal cell vectors. This anisotropic displacement model is still based on the harmonic approximation and adds six optimizable parameters to the three positional parameters required to model one atom. This is more than twice as many parameters as an isotropic displacement model requires. Therefore its application is limited to structures with reasonably high data to parameter ratios and resolutions better than 1.2 Å.

2.5.3. Anharmonic Displacement

In the context of charge density analysis it is often desirable to have an anharmonic description of atomic displacement (Sørensen *et al.*, 2003, Zhurov *et al.*, 2011). The most popular modeling technique is the Gram-Charlier model e.g. (Johnson, 1969). This further complicates the atomic scattering factor and adds a significant amount of parameters to the model for each atom. Anharmonic modeling of atomic vibrations is not compatible with other methods that are essential for the work discussed in this thesis. Hence a more detailed introduction to anharmonic modeling of vibration is omitted. Possible modifications to include anharmonic motion in the presented methods are discussed in the corresponding sections.

2.5.4. Rigid Group Displacement

Another possibility to parametrize atomic displacement is to segment the molecular framework into rigid groups. A rigid group is a collection of atoms that have different atomic positional parameters but share the same vibrational parameters.[8] As a result, fewer parameters are needed to describe the overall atomic displacement in the structure. This can be advantageous for large structures refined against comparably poor data where the refinement of additional parameters is not feasible. Especially protein crystal data sets that do not reach atomic resolution[9] often rely on this displacement model (Merritt, 1999). The most commonly used implementation of a rigid group displacement model is the translation-libration-screw (TLS) model (Schomaker and Trueblood, 1968). The TLS model describes the vibrational movement of a group of atoms as two separate parts: translational movement T and librational movement L. S encodes the coupling between the two parts. Translational movement is considered to be a movement where all atoms of a rigid group move in the same direction – describable by a translation vector with a length corresponding to the displacement amplitude and a direction corresponding to the movement direction. The overall translational motion is encoded in the 3×3 matrix T. Librational movement is considered vibrational movement where the whole rigid group is rotated around an arbitrary axis describable by a rotation axis with a length corresponding to the libration amplitude. The overall librational motion is encoded in the 3×3 matrix L. The coupling is encoded in the 4×4 matrix S resulting in 20 parameters for each rigid group overall (Schomaker and Trueblood, 1968).

Assuming the 20 parameters are known via optimization against experimental data, the ADP of atom k can be computed with

$$\begin{aligned}
U_{11}^{TLS} =& L_{22}z^2 + L_{33}y^2 - 2L_{23}yz + 2S_{21}z - 2S_{31}y + T_{11} \\
U_{22}^{TLS} =& L_{11}z^2 + L_{33}x^2 - 2L_{13}xz - 2S_{12}z + 2S_{32}x + T_{22} \\
U_{33}^{TLS} =& L_{11}z^2 + L_{33}x^2 - 2L_{12}xy - 2S_{23}x + 2S_{13}y + T_{33} \\
U_{12}^{TLS} =& - L_{33}xy + L_{23}xz + L_{13}yz - L_{12}z^2 \\
& + (S_{22} - S_{11})z + S_{31}x - S_{32}y + T_{12} \\
U_{13}^{TLS} =& - L_{22}xz + L_{23}xy - L_{13}y^2 + L_{12}yz \\
& + (S_{11} - S_{33})y + S_{23}z - S_{21}x + T_{13} \\
U_{23}^{TLS} =& - L_{11}yz - L_{23}x^2 + L_{31}xy + L_{12}xz \\
& + (S_{33} - S_{22})x + S_{12}y - S_{13}z + T_{23}
\end{aligned} \quad . \tag{2.6}$$

[8] The displacement description of an atom in a rigid group generally depends on the atom's positional parameters and the rigid group's displacement parameters. This implies that two different atoms do not necessarily share the same displacement ellipsoids. Instead, the parameters used to generate the displacement ellipsoid are shared among atom's within the same rigid group.

[9] Atomic resolution means that complete diffration data up to a resolution of about $d = 1.2$ Å is available.

where (x, y, z) is the positional vector of atom k in Cartesian space.

Compared to an anisotropic displacement model without rigid group constraints, the TLS model reduces the overall number of parameters if a rigid group contains more than four atoms on average. This is particular efficient if parts of a structure are known to be comparably rigid and significant relative motion occurs mostly between these rigid groups. Protein structures are a good example for this: covalently bonded atoms within an amino acid group will most likely follow the rigid group approximation well. The flexibility of the protein structure is modeled by allowing amino-acid groups to move relative to each other.[10]

In the context of this thesis the TLS model is used differently. Instead of optimizing TLS parameters against experimental diffraction data, the parameters are optimized against anisotropically modeled ADPs that were refined against experimental data. This protocol allows to use the information encoded in some well defined parameters in a rigid group to make predictions about other atoms in the same group that are less well defined by experimental data. For example the vibrational behavior of a hydrogen atom can be extrapolated from the vibrational behavior of the carbon atom it is bonded to. Within this thesis, the refinement of TLS parameters against experimental data will be denoted *TLS-Refinement* and the optimization of TLS parameters against anisotropic ADPs will be denoted *TLS-Analysis* to avoid confusion.

2.5.5. Segmented Rigid Body Displacement

The TLS model described in the previous section assumes that a molecule can be described by independent rigid bodies that move relative to each other. This can be an appropriate model especially if several isolated molecules are present in the asymmetric unit. However, larger molecules – like proteins – do not consist of isolated rigid bodies, meaning that the motion of one amino acid (assuming that the amino acid itself is completely rigid) depends on the movement of the next amino acid. Therefore a whole molecule can be considered a series of interconnected attached rigid groups (ARGs) where the motion of each group is constrained by the motion of the groups it is attached to (Dunitz and White, 1973, Schomaker and Trueblood, 1998).

This is taken into account by an extension to the TLS model. The TLS+ARG model. The extension adds seven parameters A for each rigid group to the 20 parameters from the TLS model resulting in the following expression for an atom's ADP:

[10] Modeling a amico-acid chain this way can easily result in implausable displacement models. A modified model constraining the ridig body motion is discussed in the next section.

$$\begin{aligned}
U_{11}^{TLS+ARG} =& U_{11}^{TLS} + V_1^2 A_1 + 2zV_1A_3 - 2yV_1A_4 + 2V_1A_5 \\
U_{22}^{TLS+ARG} =& U_{22}^{TLS} + V_2^2 A_1 - 2zV_2A_2 + 2xV_2A_4 + 2V_2A_6 \\
U_{33}^{TLS+ARG} =& U_{33}^{TLS} + V_3^2 A_1 + 2yV_3A_2 - 2xV_3A_3 + 2V_3A_7 \\
U_{12}^{TLS+ARG} =& U_{12}^{TLS} + V_1V_2A_1 - zV_1A_2 + zV_2A_3 \\
& + (xV_1 - yV_2)A_4 + V_2A_5 + V_1A_6 \\
U_{13}^{TLS+ARG} =& U_{13}^{TLS} + V_1V_3A_1 + yV_1A_2 + (zV_3 - xV_1)A_3 \\
& - yV_3A_4 + V_3A_5 + V_1A_7 \\
U_{23}^{TLS+ARG} =& U_{23}^{TLS} + V_2V_3A_1 + (yV_2 - zV_3)A_2 - xV_2A_3 \\
& + xV_3A_4 + V_3A_6 + V_2A_7
\end{aligned} \tag{2.7}$$

with

$$R = v - P \tag{2.8}$$

and

$$V = t \times R = (V_1, V_2, V_3) \tag{2.9}$$

where P is the shortest distance between t and the Cartesian origin. The same way as the TLS model can be optimized against experimental diffraction data or against already optimized ADPs, the TLS+ARG model can be used for both applications. The latter one is relevant to this work.

Part II.

Hydrogen Atoms

3. Riding Hydrogen Atom Model

The most commonly applied model for parameterizing hydrogen atoms in crystal structures is the *riding atom model* (Sheldrick, 2008). The model defines the position of an atom as a set of fixed distances to reference atoms. The angle between two bond vectors can also be considered a distance – namely the distance between the not bonded entities. In the most common case the riding atom model is used to place hydrogen atoms on idealized positions based on the geometry of the more well defined framework of heavy atoms.[1] The model allows the free refinement of the framework's atomic positions while automatically updating the positions of the riding atoms. This is a significant advantage over constraining the positions of hydrogen atoms to absolute coordinates, which would require manual updating after each refinement cycle.

The thermal displacement parameters of riding atoms can be treated in a similar fashion based on the following consideration: if the atomic position of the riding atom is constrained to the heavy atom, the riding atom must follow a similar vibrational motion plus the vibration relative to the heavy atom. In practice this means that the principal components of the heavy atom's ADP, which correspond to the displacement amplitudes, are averaged and used to estimate the amplitude of an isotropic displacement model for the riding atom. Since the riding atom, typically hydrogen, is lighter than the heavy atom, the isotropic ADP is then multiplied by an empirical factor to take the mass difference into account. In the *SHELXL* program which is used as a reference here this empirical factor is 1.5 for hydrogen atoms riding on a sp^2 hybridized carbon atom and 1.2 in all other cases.

The validity of those empirical factors was investigated in the context of their temperature dependence (Bürgi and Capelli, 2000, Busing and Levy, 1964). A series of Neutron diffraction data sets, which facilitate the determination of hydrogen ADPs empirically, was analyzed and compared to structure models carefully refined against high resolution XRD data. Two scattering factor models (invariom model and HAR) were tested. Hydrogen ADPs computed from ONIOM computation results provide a second reference data set to minimize the influence of systematic errors on the analysis (Lübben

[1] In this context all atoms heavier than hydrogen are considered heavy atoms.

et al., 2014). It was investigated whether the temperatur dependence of hydrogen ADPs relative to its bonding partner's ADP shows the same temperature dependence across all investigated structure models and whether that dependence is accounted for by the riding atom model.

3.1. Methods

N-acetyl-L-hydroxyproline monohydrate was used as a test case for investigating the temperature dependence of hydrogen ADPs in the riding atom model. Neutron diffraction data sets collected at 9, 150, 200 and $250\ \mathrm{K}$ were used as benchmarks and were compared to high resolution X-ray diffraction data sets collected at 9, 30, 50, 75, 100, 150, 200 and $250\ \mathrm{K}$ (Lübben *et al.*, 2014).

The temperature dependence is visualized by plotting the displacement amplitude of a hydrogen atom relative to the mean displacement amplitude of the bonded atom against the measurement temperature. Plots were generated for the Neutron diffraction data sets, two models optimized against XRD data and the theory derived models. The general shape of these plots is compared to verify that the temperature can be determined reliably. Finally, the plots are compared to the riding atom model.

3.1.1. Experimental Details[2]

Single crystals of the compound N-acetyl-L-hydroxyproline monohydrate (NAC·H_2O) were grown by slow evaporation of a saturated solutions prepared in hot acetone. Crystals grow to sizes suitable for neutron diffraction. A series of multi-temperature X-ray diffraction data collections at 9, 30, 50 and $75\ \mathrm{K}$[3] on the same specimen with dimensions of $0.34 \times 0.28 \times 0.28\ \mathrm{mm}$ ($0.5\ \mathrm{mm}$ pinhole) was collected at the DORIS beamline D3 at the HASYLAB/DESY synchrotron in Hamburg. The experimental setup consisted of an Oxford Diffraction open-flow Helium gas-stream cooling device, a Huber type 512 four-circle diffractometer and a $165\ \mathrm{mm}$ MAR CCD area detector. A wavelength of 0.5166 Å and a detector distance of $40.3\ \mathrm{mm}$ were chosen, allowing a high resolution of $d = 0.50$ Å or $\sin\theta/\lambda$ of 1.0 Å^{-1} to be reached with a single detector setting.

[2] This section contains excerpt from (Lübben *et al.*, 2014).

[3] Post analysis of the temperature and volume dependence of unit-cell parameters (see Figure 3.1) showed that the data point at 67 K (as indicated on the low-T device) was an outlier, probably due to inaccuracies caused by heating the cold stream of helium gas to higher temperatures. We have corrected this temperature to 75 K, as derived from a plot of the increase of the unit-cell volume with temperature. Another reason for the deviating behavior might be rotational disorder and this is discussed below.

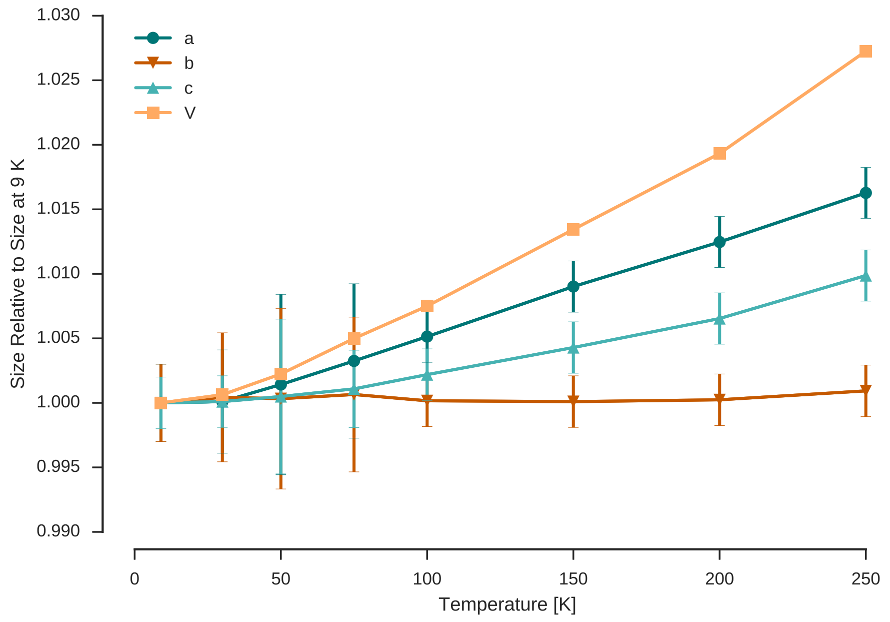


Figure 3.1.: Temperature dependence of the lattice constants of the X-ray data of *N*-Acetyl-L-Hydroxyproline monohydrate. Unit-cell parameters and volume are normalized to the lowest data point at $9\ \mathrm{K}$. Estimated standard deviations are also plotted. Connecting lines are guides to the eye.

The XDS program (Kabsch, 2010) was used for data integration and scaling. Standard deviations of the unit-cell parameters were obtained by calculating the variance of intermediate cells during integration.

A detector correction (Johnas *et al.*, 2006) was applied to properly correct for the effect of oblique incidence (Wu *et al.*, 2002) on the measured intensities. An empirical absorption correction was not performed at this short wavelength; Friedel opposites were merged. The structural model, cell settings but not the atom notation of the original structure determination by Hospital *et al.* (1979) as given in the cif file of the Cambridge Structural Database refcode NAHYPL were used as input. Preliminary least-squares refinements were initialized with this model and performed with the program *SHELXL* (Sheldrick, 2008).

Data sets at 100, 150 and 200 and $250\ \mathrm{K}$ were collected on an Xcalibur S diffractometer equipped with a Mo $K\alpha$ sealed tube. Here an analytical absorption correction was performed following the method by Clark and Reid (1995) as implemented in the program CRYSALIS RED (Oxford-Diffraction-Ltd., 2006) employed for data reduction;

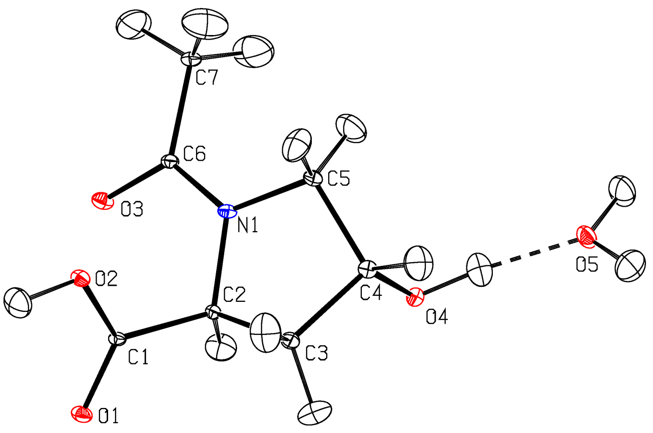


Figure 3.2.: ADPs of *N*-Acetyl-L-Hydroxyproline monohydrate from neutron diffraction at $T = 9\ \mathrm{K}$. Ellipsoids at 50 % probability (Burnett and Johnson, 1996).

Friedel mates were not merged. A second specimen was used for these four higher temperatures. High-resolution data ($\sin\theta/\lambda \geq 1$) were again measured with the exception of the data set at $250\ \mathrm{K}$.

Neutron diffraction data was collected at the OPAL reactor on the Koala beamline at ANSTO, the Australian Nuclear Science and Technology Organization in Lucas Heights, Australia. Data was collected at temperatures of 9, 150, 200 and $250\ \mathrm{K}$ and processed with LAUEG (Campbell, 1995) using the same specimen with a size of $1.8 \times 1.4 \times 0.5\ \mathrm{mm}$ and the Laue time of flight method. 16, 12, 12 and 10 images with exposure times of 42 minutes was collected for each data set. Unit-cell parameters from X-ray diffraction data collections at the respective temperature were used for indexing and data integration. The CRYSTALS program (Betteridge *et al.*, 2003) was used for the refinement of positions and ADPs for all atoms. An isotropic extinction parameter was required due to good crystal quality and comparably large specimen size for the neutron data. CCDC 977814-977817 contains the supplementary crystallographic information for the neutron data. These files can be obtained free of charge from the Cambridge Crystallographic Data Centre via www.ccdc.cam.ac.uk/data_request/cif. A depiction of the molecule with its atomic numbering scheme and anisotropic ADPs at $9\ \mathrm{K}$ from neutron diffraction is shown in Figure 3.2.

3.1.2. Compared Values

The relative amplitude of hydrogen and heavy atom displacement parameters must be quantified in order to investigate the temperature dependence.[4] This is done by computing U_{rel} for every hydrogen atom[5] which is defined as

$$U_{rel} = \frac{U_{iso}}{U_{eq}} \tag{3.1}$$

with

$$U_{eq} = \left(U_{11} + U_{22} + U_{33}\right)/3 \tag{3.2}$$

for

$$U_{ij} = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{pmatrix}. \tag{3.3}$$

Values for U_{eq} were grouped by equivalent chemical environment (following the invariom partitioning scheme described in section 2.3) to obtain more accurate values through averaging. Table 3.1 lists all atoms and their corresponding invariom names.[6] Figures 3.3 and 3.4 show values of U_{rel} grouped by invariom name plotted against the diffraction temperature.

3.1.3. Benchmark Values from Neutron Diffraction

Neutron single crystal diffraction yields very reliable information about the thermal motion of atoms and is the *de facto* gold standard for obtaining accurate displacement description in crystallography. The good reliability of of Neutron diffraction data in this context is due to the fact that hydrogen nuclei have a comparably large scattering length when irradiated by Neutrons (see section 1.0.2). In contrast to XRD, which does not permit the free optimization of hydrogen atom parameters, the optimization of an isotropic displacement model for hydrogen atoms against Neutron diffraction data is feasible.

[4] All computations are performed in Cartesian space.

[5] Steward (1972) demonstrated that the correct value of U_{eq} (Fischer and Tillmanns, 1988) is between the arithmetic and the geometric mean. Considering the overall accuracy of the obtained values, this difference is negligible.

[6] Note that none of the figures show standard deviations for any of the displayed values. The most often occurring chemical environment H1c[1c1h1h] shows a variance between 0.003 and 0.3 across all temperatures and experiments. Assuming the worst case that the standard deviation is the highest observed deviation of 0.3 in all cases, it is still justifiable to extract a general trend from the obtained results.

Atom Name	Invariom Name	Model Compound
O(1)	O2c	formaldehyde
O(2)	O1c1h	methanol
O(3)	O1.5c[1.5n1c]	acetamide
O(4)	O1c1h	methanol
O(5)	O1h1h	water
N(1)	N1.5c[1.5o1c]1c1c	*N,N*-dimethylacetamide
C(1)	C2o1o1c	acetic acid
C(2)	C1n1c1c1h	2-aminopropane
C(3)	C1c1c1h1h	propane
C(4)	C1o1c1c1h	2-propanol
C(5)	C1n1c1h1h	ethylamine
C(6)	C1.5o1.5n[1c1c]1c	*N,N*-dimethylacetamide
C(7)	C1c1h1h1h	ethane
H(1,2)	H1o[1c]	methanol
H(3)	H1c[1n1c1c]	2-aminopropane
H(4,5)	H1c[1c1c1h]	propane
H(6)	H1c[1o1c1c]	2-propanol
H(7,8)	H1c[1n1c1h]	ethylamine
H(9,10,11)	H1c[1c1h1h]	ethane
H(12,13)	H1o[1h]	water

Table 3.1.: Atom labels and their corresponding invariom names as well as the model compound defining the idealized chemical environment. Figure 3.2 shows the atom labeling.

Figure 3.3 shows the relative size of thermal ellipsoids grouped by similar chemical environment.

3.1.4. Benchmark Values from Theory

To complement the experimental benchmark values with values derived from theory, a two-layer ONIOM computation (Svensson *et al.*, 1996) was performed and is combined with the rigid body description of the structure models derived via TLS fit (Whitten and Spackman, 2006). The ONIOM computations were initiated with the atomic coordinates taken from the invariom model refined against XRD data collected at $9\,\mathrm{K}$. The positions of hydrogen atoms were set to idealized positions as defined by the appropriate AFIX commands available in *SHELXL*. The X–H bond lengths were elongated to approximate those obtained from Neutron diffraction experiments. The program *BAERLAUCH* (Dittrich, 2009, Dittrich *et al.*, 2012) was used to generate a cluster of 17 asymmetric units. The theory level for optimizing the geometry of the cluster was B3LYP/cc-pVTZ (Becke, 1988). Vibrational frequencies for the geometry optimized central unit were obtained with fixed positions of the 16 outer units at the theory level B3LYP/cc-pVTZ:B3LYP/3-21G. This procedure has proven to yield reliable estimation of vibrational modes in previous studies (Whitten and Spackman, 2006). Computed frequencies and amplitudes were converted to Cartesian coordinate space and used to compute ADPs as described in section 3.1.5 and were used as internal ADPs for the following study.[7]

For each of the XRD data sets the computed internal ADPs were subtracted from the refined ADPs of all freely optimized atoms. A TLS fit was then performed against the difference ADPs yielding the 20 parameters T_{11} to S_{23}.

In a final step the estimated external ADPs U_{ij}^{TLS} of each hydrogen atom were added to the internal ADPs U_{ij}^{ONIOM} of the corresponding atom from the ONIOM computation. Results for each temperature, grouped by similar chemical environment, are shown in figure 3.3.

3.1.5. Converting Vibrational Modes into ADPs

The studies discussed in this thesis involve the combination of experimentally derived ADPs with ADPs estimated by quantum chemical computations. In crystallography ADPs encode the averaged and squared displacement of an atom in the direction of three perpendicular axes. The program GAUSSIAN, which was used for predicting

[7] A more detailed description of the concept of internal and external ADP is given in chapter 4.

equivalent information from theory, describes displacement in the form of vibrational normal mode frequencies and corresponding displacement amplitudes for each atom. In order to combine theory and experiment in this context, it is necessary to convert the normal mode representation of thermal motion into the ADP representation. This section describes how to facilitate the conversion.

The output of the GAUSSIAN program after frequency computation includes a table listing every vibrational frequency $\overline{\nu}_j$. For each frequency a column vector $\overline{d}_j$ of displacement amplitudes is provided. The vector lengths is $3 \cdot i$ for a molecule with i atoms. Equation 3.4 shows the composition of $\overline{d}_j$ for a molecule with two atoms.

$$\overline{d}^T = (x_a, y_a, z_a, x_b, y_b, z_b) \tag{3.4}$$

D is a matrix consisting of all column vectors $\overline{d}_j$. First, the excitation δ_j of each mode is computed with

$$\delta_j = \frac{1}{tanh\left(\frac{h \cdot k \cdot \overline{v}_j}{T}\right)} \cdot \frac{h \cdot c}{\overline{v}_j \cdot \mu_i} \tag{3.5}$$

where h is the Planck constant, k is the Boltzmann constant, T is the temperature, c is the speed of light and μ_j is the reduced mass associated with the vibrational mode j. δ_j is then multiplied by the Kronecker symbol ξ_i to yield Δ.

$$\Delta = \xi_i \cdot \delta_j \tag{3.6}$$

The mean squared displacement matrix U is then computed with

$$U = D^T \cdot \Delta \cdot D \tag{3.7}$$

where the i^{th} 3×3 sub-matrices along the diagonal corresponds to the ADP of the i^{th}

atom:

$$\left(\begin{array}{ccccccccc}
\cline{1-3}
\multicolumn{1}{|c}{U_{11}} & U_{12} & \multicolumn{1}{c|}{U_{13}} & U_{14} & U_{15} & U_{16} & U_{17} & U_{18} & U_{19} \\
\multicolumn{1}{|c}{U_{21}} & U_{22} & \multicolumn{1}{c|}{U_{23}} & U_{24} & U_{25} & U_{26} & U_{27} & U_{28} & U_{29} \\
\multicolumn{1}{|c}{U_{31}} & U_{32} & \multicolumn{1}{c|}{U_{33}} & U_{34} & U_{35} & U_{36} & U_{37} & U_{38} & U_{39} \\
\cline{1-6}
U_{41} & U_{42} & U_{43} & \multicolumn{1}{|c}{U_{44}} & U_{45} & \multicolumn{1}{c|}{U_{46}} & U_{47} & U_{48} & U_{49} \\
U_{51} & U_{52} & U_{53} & \multicolumn{1}{|c}{U_{54}} & U_{55} & \multicolumn{1}{c|}{U_{56}} & U_{57} & U_{58} & U_{59} \\
U_{61} & U_{62} & U_{63} & \multicolumn{1}{|c}{U_{64}} & U_{65} & \multicolumn{1}{c|}{U_{66}} & U_{67} & U_{68} & U_{69} \\
\cline{4-9}
U_{71} & U_{72} & U_{73} & U_{74} & U_{75} & U_{76} & \multicolumn{1}{|c}{U_{77}} & U_{78} & \multicolumn{1}{c|}{U_{79}} \\
U_{81} & U_{82} & U_{83} & U_{84} & U_{85} & U_{86} & \multicolumn{1}{|c}{U_{87}} & U_{88} & \multicolumn{1}{c|}{U_{89}} \\
U_{91} & U_{92} & U_{93} & U_{94} & U_{95} & U_{96} & \multicolumn{1}{|c}{U_{97}} & U_{98} & \multicolumn{1}{c|}{U_{99}} \\
\cline{7-9}
\end{array}\right) \tag{3.8}$$

In practice, not all computed modes are meaningful for the description of thermal vibrations. The frequency computation output might contain imaginary vibrational modes which correspond the saddle points on the potential hyper-surface instead of minima. Such modes need to be filtered out before computing ADPs.

Depending on the application of the estimated ADPs it might also be necessary to remove frequencies below a certain threshold from the ADP computation. Low frequency modes usually correspond to distortion of the atom framework and involve the motion of many atoms at once. For certain applications it is desirable to describe only the motion of an atom relative to its immediate neighbor atoms. In this case an appropriate low-frequency cutoff needs to be chosen (Madsen *et al.*, 2013). Another reason for omitting the lowest frequencies is that the accuracy of these modes is lower than for higher frequencies. Considering that these modes have a disproportionally large impact on the overall displacement (see equation 3.5 for small values of $\overline{\nu}_j$) it is usually recommended to ignore these values in the context of this thesis. Leaving out low frequencies leads to underestimation of internal ADPs. However, when the estimated ADPs are combined with a TLS fit, the missing part is absorbed in the TLS parameters leading to no observable errors in all studied cases.

3.1.6. ADP Ratios from X-ray Diffraction

As discussed in previous sections, refining hydrogen atom model parameters against XRD data requires a more sophisticated scattering factor model than the IAM. But even with highly accurate, high resolution data, an appropriate scattering factor model and a carefully selected refining strategy, the refined parameters should not be trusted blindly (Jelsch *et al.*, 1998)(Dittrich *et al.*, 2008). To ensure the best achievable results, two different refinement techniques were employed. The results were cross-referenced to check if the independently obtained results are comparable. The selected refinement techniques were:

Invariom Model with constrained hydrogen-atom positions and a freely refined isotropic displacement parameter for each hydrogen atom.

HAR with freely refined hydrogen-atom positions and a freely refined isotropic displacement parameter for each hydrogen atom.

3.2. Results

3.2.1. Benchmark Values from Neutron Diffraction

Figure 3.3 (top) shows the values of U_{rel} for all temperatures grouped by invariom name. All chemical environments show similar temperature dependence. Environment H1c[1c1h1h] (a hydrogen atom in a methyl group) shows significantly larger U_{rel} values. Considering that the ADP of the carbon atom in the methyl group is not smaller than other heavy atom ADPs, this must imply that the bonded hydrogen atom ADPs are systematically larger than other hydrogen ADPs. This is plausible since methyl groups often show signs of rotational disorder. The optimized structural model did not account for disorder, which can lead to the absorption of density smearing into the ADP of hydrogen atoms.

Overall, it can be seen that U_{rel} is significantly larger at temperatures below $150\ \mathrm{K}$. At higher temperatures U_{rel} appears to remain constant. The temperature at which the ratio stops being constant cannot be extrapolated from this data because data sets between 9 and $150\ \mathrm{K}$ are not available.

3.2.2. Benchmark Values from Theory

Figure 3.3 (bottom) shows the values of U_{rel} for all temperatures grouped by invariom name. The plot shows similar trends than the previous one but all environments are more similar. This supports the hypothesis that disorder causes the enlarged ADPs of methyl group hydrogen atoms. The quantum-chemical computation does not account for multiple conformations. The effect of a disordered methyl group can therefore not be reproduced by the computations, and the effect does not show in the plot. Again, U_{rel} remains almost constant at temperatures above $150\ \mathrm{K}$. Above that temperatures U_{rel}

remains between a value of 1.2 and 1.5 which are the default values for the riding atom model used in the SHELXL program.

3.2.3. ADP Ratios from X-ray Diffraction

Figure 3.4 shows the temperature dependence of U_{rel} in the models optimized against XRD data. Overall, the data shows significantly more random errors. This is to be expected because the scattering contribution of hydrogen atoms in XRD experiments is very low. Therefore hydrogen atoms are more strongly affected by random errors than other model parameters. However, the overall shape of the plots is similar to both benchmark studies discussed before.

Both models show enlarged ADPs for the methyl-group hydrogen atoms, which is consistent with the Neutron diffraction study. This supports the hypothesis that disorder causes the effect because the XRD studies should be affected by disorder the same way the Neutron diffraction study is.

Only few differences between the invariom model and the HAR model are visible. The most significant difference can be observed for the hydroxyl group's hydrogen atom with the invariom name H1o[1c]. The difference is most likely due to how packing effects are treated by the two different scattering factor models. The invariom model does not take crystal packing into account because the invariom database does not facilitate storage and transfer of packing information (see section 2.3 for details). This can cause small errors for atoms that are strongly affected by crystal packing e.g. hydrogen atoms involved in strong hydrogen bonding. HAR generates tailor-made scattering factors that are specific to the studied structure. This implies that packing affects are accounted for to some degree. Moreover, HAR can utilize Hirshfeld partitioned point charges to approximate the crystal field. The hydroxyl group hydrogen atom in the studied structure being part of a hydrogen bond is most likely the reason for the observed differences between both models.

3.2.4. Summary & Conclusion

U_{rel} values from all methods are in good agreement with each other. A dependence of U_{rel} is clearly visible and significant for temperatures below $100\ K$. These results are also in good agreement with the physical principles behind vibrational states.

At a sufficiently high temperature all vibrational states should be excited. A state's excitation level is thereby determined by the associated vibration frequency. The lower the

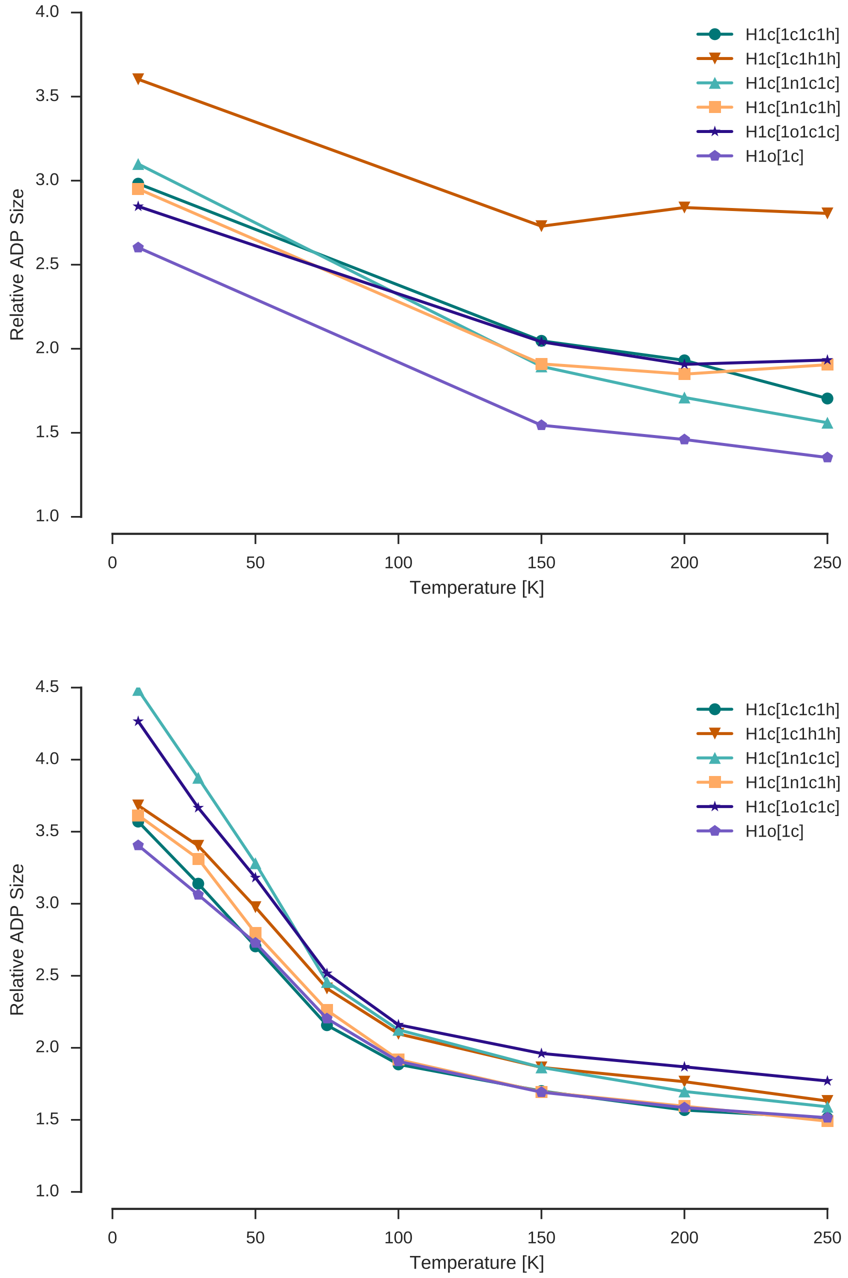
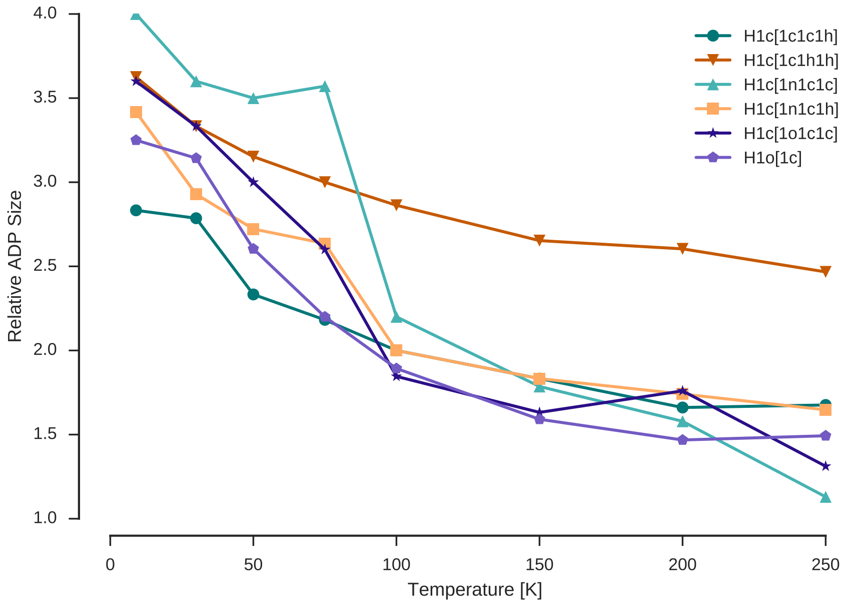


Figure 3.3.: Top: temperature dependence of U_{rel} obtained by Neutron diffraction. Bottom: temperature dependence of U_{rel} obtained by ONIOM computations.



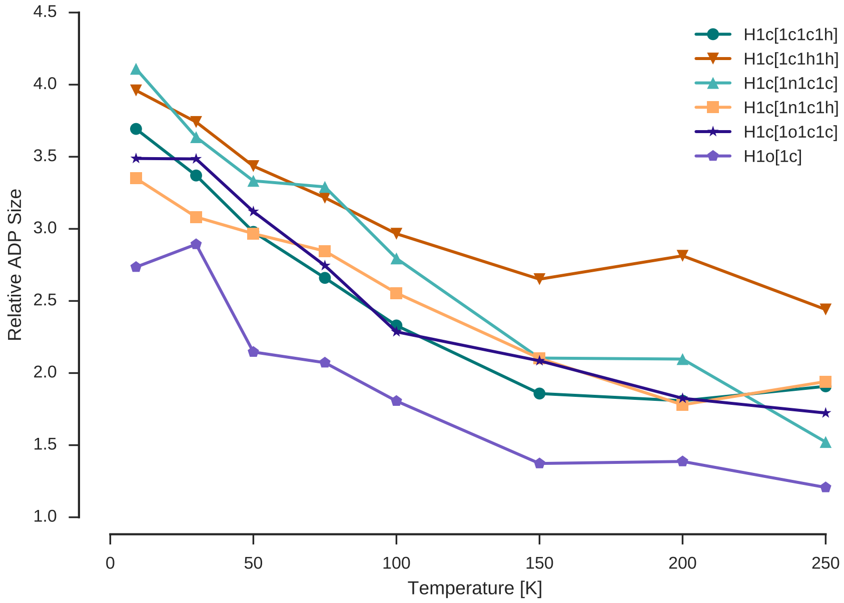


Figure 3.4.: Top: temperature dependence of U_{rel} obtained by invariom refinement against XRD data. Bottom: temperature dependence of U_{rel} obtained by HAR against XRD data.

frequency the higher the excitation. The low frequency modes in a molecular crystal are those that displace the molecule as a whole relative to its lattice neighbor. The atomic displacement caused by these vibrations is – in good approximation – equal for bonded atoms. Including X–H atom pairs. The next higher frequencies belong to those motions that describe the deformation of the molecular framework e.g. stretching of helical structures. Motion of this kind also affects bonded atom pairs almost equally. The highest vibrational frequencies are associated with the displacement of atoms held together by comparably strong forces, namely atoms bonded to each other or those connected by a small number of bonds. Because the interaction energy holding these atoms at their positions is much stronger than long range intra-molecular forces or inter-molecular forces, a lot of energy is required to displace these atoms from their energetically ideal position. Therefore, these displacements correspond to high energy – meaning high frequency – vibrations.

In the context of U_{rel} the important characteristic of these high energy vibrations is the fact that the displacement caused by these vibrations depends on the atomic mass. Considering an approximately harmonic potential, the atomic displacement of two bonded entities caused by these modes should be proportional to an atom's mass. This means that a high energy mode should displace a hydrogen atom approximately six times as much as the bonded carbon atom. For vibrational modes that do not significantly involve the motion of bonded atoms relative to each other, the displacement of each atom is mass independent and therefore equal for X–H atom pairs. This is the reason why the thermal ellipsoids of lighter atoms are usually bigger than those of bonded heavier atoms.

However, if those were the only principles affecting the size of thermal displacement ellipsoids, the temperature dependence of U_{rel} could not be explained. The relative size of ellipsoids of X–H pairs should be constant across all temperatures which is only supported by the collected data for temperatures above $100\ K$. The temperature dependence requires another effect to be considered - zero point vibrations. Vibrations displacing two atoms relative to each other cannot be described accurately by a classical oscillator model. Instead one must consider quantum mechanical effects which also involves the fact that the lowest energy state of a quantum oscillator has a non-zero energy. Therefore it also involves non-zero displacements of the oscillating atoms. This means that no matter how low the temperature during data collection was, the zero point displacement of high energy vibrations will always lead to relative displacement of X–H pairs. On the other hand, lower energy vibrations involve the movement of many atoms at once, which reduces zero point energy effects to a point where they become negligible. This implies that at low temperatures there is basically no displacement caused by these vibrations.

Considering all of these factors one can explain the observed temperature dependence: at low temperatures high energy vibrations are dominant due to zero point energy effects. This leads to a big difference in ellipsoid size for X–H pairs. As the temperature rises, vibrational states are excited. Because the excitation level is frequency dependent, the low energy modes are more strongly excited which means that mass independent displacement (equal for both atoms in X–H pairs) become more dominant. This explains the drop in the relative ADP size between $9\ \mathrm{K}$ and $100\ \mathrm{K}$. At higher temperatures the system contains enough thermal energy to excite all vibrational states, making the relative ellipsoid size more and more independent of the temperature which fits the observed data.

It is therefore recommended to consider the temperature dependence of U_{rel} when estimating ADPs for hydrogen atoms especially at temperatures $150\ \mathrm{K}$. This could be done by fitting a temperature dependent scale factor against the data presented in this section (Madsen and Hoser, 2015) or by including the measurement temperature directly in the estimation procedure as discussed in the following sections.

4. Estimation of Hydrogen Atom Displacements

As discussed in the previous sections, accurate and detailed descriptions of hydrogen atoms are not easily obtainable with experimental techniques. Even model optimization against high resolution XRD data requires a lower level of detail for hydrogen atoms compared to heavier elements. Neutron diffraction experiments do yield the required data but are expensive to perform and are not available for routine work. As a result, hydrogen atom parameters in XRD studies are often not refined at all or a less detailed model is applied. Possible modeling choices include the riding atom model (section 3), refining only the atomic positions of hydrogen atoms, refining only an isotropic displacement description or any combination of those. On the other hand, a detailed and accurate description of hydrogen atoms is necessary to study molecular interactions which are most likely mediated by contacts between hydrogen atoms (Dominiak *et al.*, 2012). Also, studies relying on thermodynamic properties require detailed information about thermal motion of all atoms to reliably estimate entropy contributions (Madsen and Larsen, 2007). A detailed parametrization of hydrogen vibrations also leads to higher precision of the overall model (Brock *et al.*, 1991).

Thermal motion models are also prone to absorb crystal packing deficiencies into the ADPs. The crystallographic method only records a space and time averaged representation of the crystal. Therefore errors in crystal packing or conformational changes over time can manifest indistinguishable from thermal motion in the diffraction data. Systematic errors will be introduced, if ADPs are determined purely by optimizing a model against that data. An estimation method for ADPs can be useful to validate empirically determined ADPs (Bürgi and Capelli, 2000).

The most commonly applied method to estimate ADPs of hydrogen atoms is the simple hydrogen ADP estimator (SHADE) Server (Madsen, 2006). The SHADE Server relies upon a library of structure models refined against Neutron diffraction data. ADPs for hydrogen atoms are taken from this library and are transferred to chemically similar atoms in the studied structure. The parameters are then combined with a TLS+ARG model to take rigid body motion of the molecule into account.

This section investigates an alternative method to obtain estimated values for hydrogen ADPs (Lübben *et al.*, 2015). The method is based on invariom partitioning (Dittrich *et al.*, 2013), the invariom database and a segmented rigid model description (TLS+ARG).

4.1. Methods

The SHADE method (Madsen, 2006), (Madsen and Hoser, 2014) and the method presented here are based on the assumption that thermal motion of an atom in a crystal structure can be separated into two independent contributions: internal ADP and external ADP (Schomaker and Trueblood, 1968). The internal ADP describes how the atoms within the asymmetric unit move relative to each other. The external ADP describes how a rigid asymmetric unit moves relative to other asymmetric units.

This separation works well for small, rigid molecules. However, larger and more flexible molecules cannot be described well as one rigid body. One solution to this problem is to *cut* a more flexible molecule into smaller units. Each unit is chosen such that it satisfies a rigid body approximation. While each of these units – or segments – is rigid in itself, different segments are allowed to move relative to each other. In the TLS+ARG model this is achieved by defining a bond separating two rigid segments as a rotation axis that one segment rotates about, while the other segment does not (Schomaker and Trueblood, 1998).

4.1.1. Rigid Body Segmentation

The segmentation procedure – the selection of bonds between supposedly rigid segments – can be done manually. However, the procedure can be tedious for larger molecules and introduces bias by the researcher. Moreover, most molecular systems do not consist of segments that are obviously rigid to the human eye (Merritt, 1999). An automated rigid body segmentation algorithm is presented that works around that problem. The procedure is based on the analysis of ADPs and the connectivity of the atomic framework. The method requires no user input and will consistently result in the same segmentation model for the same input data. This streamlines the application of the TLS+ARG method significantly and makes it feasible to be applied in routine structure analysis.

The algorithm requires a certain level of detail of the structure model input in order too work correctly. The input model must contain atomic positional data and anisotropically refined displacement parameters for non-hydrogen atoms. It is also recommended to limit the application of the method to data collected at temperatures below $150\,\mathrm{K}$. Above that threshold ADPs become too large and contain too many statistical and systematic errors for the algorithm to produce plausible results e.g. deviation from harmonicity.

1. In a first step the algorithm searches for all single bonds (Blom and Haaland, 1985) in the input molecule. Each single bond is considered to be a potential rotation axis connecting two rigid groups. Next, atoms are grouped into segments that are connected by single bonds. If a system is circular, implying that removing a single bond will result in only one molecule instead of two as is the case when cutting a non-circular bond, the single bond is ignored. Each group created this way must consist of at least 8 atoms. This is required to achieve a stable subsequent TLS+ARG fit and to avoid problems with conic sections.[1] To reduce the number of potential groups that need to be checked in the following steps only single bonds are considered in this step. This is based on the assumption that only single bonds imply rotation barriers low enough to facilitate a low energy vibration. High energy vibrations are not considered in this approach because only those vibrations that have the most significant impact on the overall ADP size are modeled.

2. The second step is performed for each of the previously generated groups. The relative displacement ΔH_{ij} in bond direction of all atom pairs within the group is computed (see equation 4.1 to 4.3).

$$H_{ij} = U_i^T \cdot \overline{v}_{ij} \cdot U_i \tag{4.1}$$

$$H_{ji} = U_j^T \cdot \overline{v}_{ij} \cdot U_j \tag{4.2}$$

$$\Delta H_{ij} = H_{ij} - H_{ji} \tag{4.3}$$

U_i is the ADP of atom i, U_j is the ADP of atom j and $\overline{v}_{ij}$ is the normalized difference vector of the positions of atom i and j. Both atom i and atom j must be part of the same segment. For a segment a consisting of n atoms ξ_a can be computed

[1] If all atoms in a rigid group lie on a conic section, TLS+ARG parameters become linearly dependent and the optimization will fail.

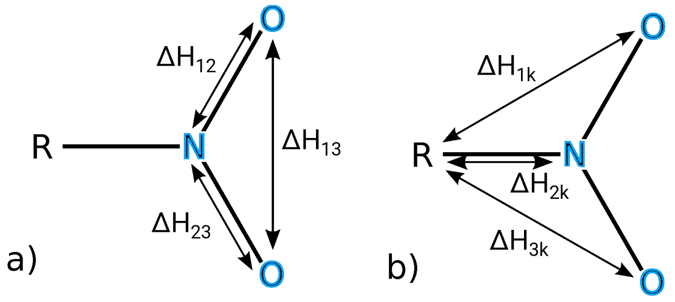


Figure 4.1.: Illustration of the rigidity criterion. Figure a) illustrates how ξ_a is computed. Figure b) illustrates how Ξ_a is computed. The average value of ΔH_{ik} (b) must be twice as big as the average value of ΔH_{ij} (a) for a group to be treated as a rigid group.

as

$$\xi_a = \frac{1}{n}\sum_{i}^{n}\sum_{j\neq i}^{n}\left(\Delta H_{ij}\right). \tag{4.4}$$

The analogous value

$$\Xi_a = \frac{1}{n}\sum_{i}^{n}\sum_{k\neq i}^{n}\left(\Delta H_{ik}\right). \tag{4.5}$$

is computed for all atom pairs where atom i is part one group a and atom k is part of another group. The criterion of

$$\Delta\xi_a < 0 \tag{4.6}$$

with

$$\Delta\xi_a = 2\cdot\xi_a - \Xi_a \tag{4.7}$$

is then used to decide whether a group is considered to be rigid. A value of $\Delta\xi_a$ greater than 0 means that the group is rejected. This criterion is determined empirically. It is based on the assumption that a group must be rigid and, at the same time, show movement relative to the rest of the molecule. If the first criterion is not fulfilled, the atoms are not part of the same rigid group. If the second criterion is not fulfilled, the atoms do belong to the same rigid group but the group should be larger and include other atoms. Figure 4.1 illustrates the meaning of $\Delta\xi$ at the example of a fictitious molecule.

The number of groups is now reduced and only contains those groups that display little relative atomic displacement within the group but significant relative motion to the rest of the molecule. This step is essential to remove the risk of overfitting that could occur when too many groups are allowed to move even though no relative motion was observed in the experiment. Instead of rigid group movement the TLS+ARG fit would then fit errors in the model.

3. Another issue is that at this point, even though no group consists of less than eight atoms, two groups can share all but one atom because two neighboring single bonds were chosen as rotation axes and none of them was rejected in the rigidity test that was performed in the previous step. Accepting both groups would result in three additional groups in total: the whole molecule minus the atoms of the first group, the whole molecule minus the atoms of the second group and the atoms of the first group minus the atoms of the second group. Applying the same criteria as in the first step where all groups consisting of fewer than eight atoms were rejected, all groups need to be cross referenced to make sure no selection of two groups implies a third group of fewer than eight atoms. This is done in an approximate manner to reduce the number of checks. Instead of checking all possible combinations of groups, the groups are sorted by their associated value of $\Delta\xi_a$ starting with the highest value. The group with the highest value of $\Delta\xi_a$ will always be accepted. The group with the second highest value is then compared to the first group by counting the number of additional bonds between the bond defining the first group and the bond defining the second group. If more than six additional bonds are between both bonds, the second group is accepted. Otherwise the group is rejected. When the group with the third highest value of $\Delta\xi_a$ is checked, the check is performed against all already accepted groups (either one or two) until all groups are either accepted or rejected. The set of accepted groups is the segmentation model generated by the algorithm. Figure 4.2 shows a visualization of the algorithm output. The algorithm is implemented in the *APD-Toolkit* software package that was developed to perform all analyses for this project.

4.1.2. Estimation of Internal ADPs

External ADPs for hydrogen atoms can be estimated via TLS fit. Internal ADPs for hydrogen atoms are not accessible from standard XRD measurements. Instead they have to be derived from theory or other experimental techniques. This section describes how to derive the information from a library of theory based, idealized chemical environments – the invariom database.

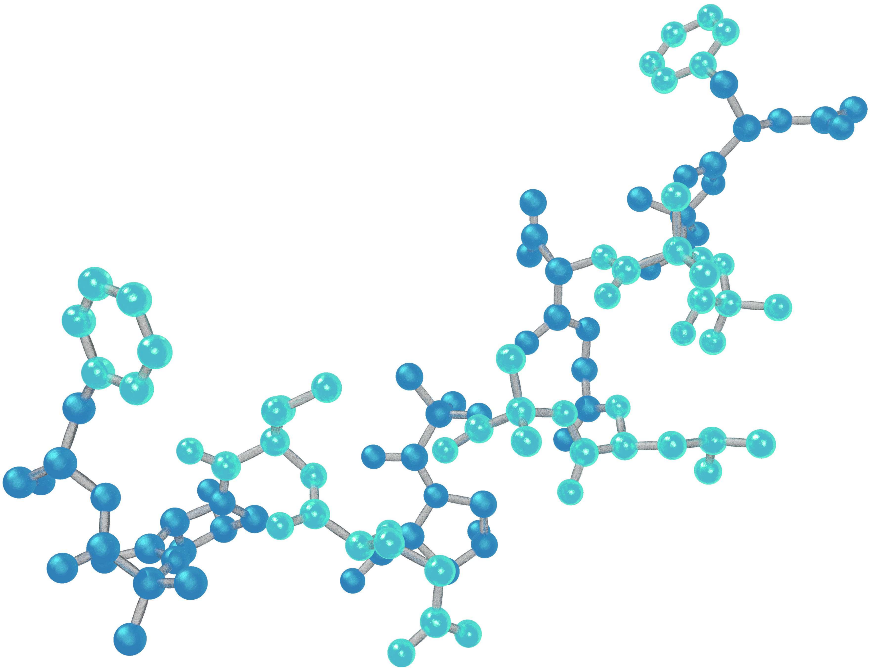


Figure 4.2.: Artistic visualization of the segmentation algorithm output at the example of an oligopeptide (PDB code 4G13). Note that a different rigid group size threshold was chosen for visualization purposes here.

The invariom database (Dittrich *et al.*, 2013) is a library of molecular data – model compounds – obtained via quantum chemical computations with the GAUSSIAN software package (Frisch *et al.*, 2013). Each model compound consists of the optimized molecular geometry, additional information like vibrational frequencies as well as a partitioning and transferability scheme that facilitates the association of arbitrary atoms in arbitrary chemical environments with their corresponding idealized model compound. The presented method applies the invariom partitioning scheme to an experimentally derived structural model to transfer localized vibrational data from the invariom database to each atom. The procedure involves several approximations:

1. Internal atomic vibration is localized.
2. Internal atomic vibration is transferable.
3. Internal and external vibrations are separable.

The first approximation is certainly not strictly true in real systems. Each vibrational mode in the invariom database displaces all atoms in the model compound, not only the one that is about to be transferred. However, the vibrational modes are not transferred directly. Instead the average displacement of the atom of interest relative to its immediate neighbor atoms is transferred. This procedure still ignores the displacement relative to the rest of the molecule. On the other hand, the subsequently applied TLS fit will most likely absorb the errors introduced this way. The second and the third point needed to be verified by applying the method to structures with known vibrational properties, for example structural models also optimized against Neutron diffraction data.

The first step in estimating internal ADP is partitioning of the molecular structure by applying the invariom partitioning scheme. The result is a list of keys that bind each atom in the structure to an atom with equivalent chemical environment in a model compound. Next, the appropriate model compounds are extracted from the invariom database and ADPs are computed based on the frequency information provided by GAUSSIAN (see section 3.1.5 for details). The ADPs are then transferred to the correct coordinate system with respect to local symmetry. This is implemented by looking for characteristic vectors within the chemical environment of an atom and its invariom in the model compound. To successfully transfer an atom, three of these vectors are required: one to specify the position of the atom in space, and two more to specify its orientation. Assuming a right-handed Cartesian coordinate system, which is used for this application, the third orientation vector is implicitly known because it must be perpendicular to the first two. The three characteristic vectors must be known for both the atom and its invariom.

For each atom, the following sequence is performed until three vectors are found:

- The first positional vector is trivial to determine and is simply the position of the atom in space.
- The chemical element types of all neighboring atoms are checked. If the element type occurs only once in the chemical environment, the position of that atom is accepted as a characteristic vector.
- For each next nearest neighbor atom in the environment the chemical element symbols of the direct neighbor atom and the next nearest neighbor atom are concatenated. If the concatenated symbol sequence of an atom is unique, the next nearest atom's positional vector is accepted as a characteristic vector.

If fewer than three vectors are chosen after the sequence terminated, local symmetry must be present. This implies that the missing vectors can be chosen to be arbitrary atomic position vectors as long as they are chosen to be equal in both the environment of the studied structural model and the environment of the model compound. The sequence of symbols in the invariom name are fixed. Therefore checking atoms in the order of their appearance in the invariom name will return consistent results for all environments.

When three characteristic vectors are known, the internal ADPs can be transferred from the invariom database's coordinate system to the crystal's coordinate system as follows[2]: the parameters in the invariom database are stored in an metrical cubic cell[3] with cell lengths of 30 Å. To streamline all transformation operations, the parameters are first transformed to Cartesian coordinate space. If V is the unit cell volume, the matrix M_{fc} is used to transform from fractional space to Cartesian space:

$$M_{fc} = \begin{pmatrix} a & b \cdot cos(\gamma) & c \cdot cos(\beta) \\ 0 & b \cdot sin(\gamma) & c \cdot \frac{cos(\alpha) - cos(\beta) \cdot cos(\gamma)}{sin(\gamma)} \\ 0 & 0 & c \cdot \frac{V}{sin(\gamma)} \end{pmatrix}. \tag{4.8}$$

If $M_{fc,inv}$ is M_{fc} with $a = b = c = 30$ Å and $\alpha = \beta = \gamma = 90°$ and $M_{cf,cryst}$ is M_{cf} with the crystal's cell parameters, the atomic position of an atom in the invariom database v_{inv} in the crystal's coordinate system v_{cryst} can be computed as

$$v_{cart} = (M_{fc,inv} \cdot v_{inv}). \tag{4.9}$$

These equations are used to transfer the characteristic vectors from invariom space to Cartesian space. The matrix representation of an ADP in invariom space

$$U_{ij,inv} = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{pmatrix} \tag{4.10}$$

is transferred to the crystal's coordinates system with

$$U_{ij,cart} = N \cdot U^*_{ij} \cdot (N)^T \tag{4.11}$$

where

$$U^*_{ij} = M_{fc,inv} \cdot U_{ij,cart} \cdot M^T_{fv,inv} \tag{4.12}$$

[2] The following section contains parts taken from (Lübben *et al.*, 2015).

[3] The space group of the artificial structure is $P\overline{1}$. However the lattice parameters form a cubic cell.

and

$$N = \begin{pmatrix} a^{-1} & 0 & 0 \\ 0 & b^{-1} & 0 \\ 0 & 0 & c^{-1} \end{pmatrix}. \tag{4.13}$$

a, b and c are the cell constants of the crystal. Now that the ADP and the characteristic vectors are in the same coordinate space, the rotation operation for mapping equivalent characteristic vectors onto each other can be determined.

The basis of the method used for determining the rotation mapping is the quaternion representation of rotation. Thereby the quaternion, a generalization of a complex number with three independent imaginary parts, encodes the orientation of a rotation axis in three-dimensional space and the angle by which an object is rotated.[4] If $\overline{v}_i^{inv}$ is the i^{th} characteristic vector in the model compound in Cartesian space, $\overline{v}_j^X$ is the j^{th} characteristic vector of the studied structure also in Cartesian space and the i^{th} and j^{th} element are corresponding characteristic vector pairs, matrix S can be determined as

$$S = \sum_{i=j=1}^{n=3} \left(\overline{v}_i^{inv} \otimes \overline{v}_j^X \right) \tag{4.14}$$

where $\overline{v}_i^{inv} \otimes \overline{v}_j^X$ denotes the outer product resulting in a 3×3 matrix. The matrix elements of S are then used to construct the 4×4 matrix N:

$$N = \begin{pmatrix} S_{11}+S_{22}+S_{33} & S_{23}-S_{32} & S_{31}-S_{13} & S_{12}-S_{21} \\ S_{23}-S_{32} & S_{11}-S_{22}-S_{33} & S_{12}+S_{21} & S_{31}+S_{13} \\ S_{31}-S_{13} & S_{12}+S_{21} & S_{22-S_{11}}-S_{33} & S_{23}+S_{32} \\ S_{12}-S_{21} & S_{31}+S_{13} & S_{23}+S_{32} & S_{33}-S_{11}-S_{22} \end{pmatrix} \tag{4.15}$$

If e_i are the eigenvalues of S and $\overline{e}_i$ are the corresponding eigenvectors, the quaternion mapping $\overline{v}_i^{inv}$ onto $\overline{v}_j^X$ is the eigenvector $\overline{e}_{max}$ corresponding to the largest eigenvalue e_{max} (Besl and McKay, 1992).

The quaternion $\overline{e}_{max}$ can be used as is to rotate a point in space. In order to streamline coordinate transformation processes it might however be desirable to convert the quaternion representation into a matrix representation. Converting $\overline{e}_{max}$ to a matrix R yields

$$R = \begin{pmatrix} r_1^2+r_2^2-r_3^2-r_4^2 & 2\cdot(r_2r_3-r_1r_4) & 2\cdot(r_2r_4+r_1r_3) \\ 2\cdot(r_3r_2+r_1r_4) & r_1^2-r_2^2+r_3^2-r_4^2 & 2\cdot(r_3r_4-r_1r_2) \\ 2\cdot(r_4r_2-r_1r_3) & 2\cdot(r_4r_3+r_1r_2) & r_1^2+r_2^2-r_3^2+r_4^2 \end{pmatrix} \tag{4.16}$$

[4] The following equations assume a quaternion is represented by a vector with 4 elements.

with

$$r = \overline{e}_{max} \tag{4.17}$$

and R_i referencing the i^{th} element of the 4D vector representation of the quaternion e_{max}. $U_{ij,cart}$ can then be rotated to match the orientation of the studied structure with

$$U^*_{ij,cart} = R^T \cdot U_{ij,cart} \cdot R. \tag{4.18}$$

It might be interesting to note that this method for $n > 3$ (see equation 4.14) yields the best possible mapping of point set i and point set j instead of the exact mapping. For example this can be useful to compute the best possible superposition of two similar structural motives. While this feature is not taken advantage of for this particular purpose, it is used throughout the implementation of the overall procedure.

4.1.3. Comparison of ADPs

It is crucial for this study to establish a quantitative comparison criterion for ADPs that are expected to be equal or similar. Visual inspection of ADPs is deemed too inaccurate for this purpose. Instead the quantitative comparison method proposed by Whitten and Spackman (2006) is used. The method works by computing the spatial overlap of two ADP ellipsoids and outputs a similarity index S between 0 – perfect overlap and 100 – no overlap. S is computed by first expressing the ADP as a probability density function $p(u)$ with

$$p(u) = \left(\frac{\det U^{-1}}{8\pi^3}\right) exp\left(-\frac{1}{2}u^T U^{-1} u\right). \tag{4.19}$$

The overlap of two probability functions $p_1(x)$ and $p_2(x)$ is given by

$$T = \int (p_1(x) \cdot p_2(x))^{1/2} \, \mathrm{d}^3 \cdot x = \frac{2^{2/3}\left(\det\left(U_1^{-1} \cdot U_2^{-1}\right)\right)^{1/4}}{\left(\det\left(U_1^{-1} + U_2^{-1}\right)\right)^{1/2}}. \tag{4.20}$$

R is then scaled to yield a value in the desired percent scale:

$$S = 100 \cdot (1 - T) \tag{4.21}$$

The proposed ADP estimation method aims to reproduce Neutron diffraction derived ADPs as accurately as possible. It is therefore desired to obtain the smallest possible values of S when comparing estimated ADPs to the ADPs optimized against Neutron diffraction data.

4.1.4. Scaling

Experimentally determined ADPs are usually subject to systematic errors that depend on the exact experimental setup, wavelengths, crystal size, temperature deviations and other factors (Blessing, 1995). Such errors make the comparison of structural models refined against data collected with different experimental setups challenging. A case where this can be especially problematic is the comparison between structures refined against XRD data and those refined against Neutron diffraction data. To reduce effects of such systematic errors on the results of the ADP comparison, ADP were scaled. The scaling procedure is based on the assumption that equivalent parameters – including the ADPs – in two structural models are supposed to be equal and that the only reason they are not is due to different systematic errors in both experiments. If that holds true, a set of scaling parameters can be fitted against pairs of parameters that are expected to be equal in both models. This set of parameters is optimized to make pairs of equivalent ADPs as equal as possible. If the scaling parameters are chosen appropriately, possible systematic errors present in one or both experiments are equalized to some degree.

The parametrization is chosen following the work published by Blessing (1995). The scaling model includes one isotropic scaling factor, adjusting the overall size of each ADP, and one set of anisotropic scaling parameters, adjusting the orientation of each ADP, resulting in seven parameters in total. Assuming the ADP of atom1 should be scaled to the ADP of atom2, the isotropic scaling parameter q_0 can be expressed as

$$U_{ij,2} = q_0 \cdot U_{ij,1}. \tag{4.22}$$

The corresponding anisotropic correction term cas be expressed as

$$U_{ij,2} = \Delta U + U_{ij,1} \tag{4.23}$$

with

$$\Delta U = \begin{pmatrix} q_1 & q_4 & q_5 \\ q_4 & q_2 & q_6 \\ q_5 & q_6 & q_3 \end{pmatrix}. \tag{4.24}$$

The full scaling expression is then

$$U_{ij,2} = q_0 \cdot U_{ij,1} + \Delta U. \tag{4.25}$$

In practice, q_0 to q_6 can be determined via least-squares optimization minimizing

$$min\left(\sum_{k=0}^{k=n}\sum_{k\neq l}^{k=n}(U_{ij,k}-(q_0\cdot U_{ij,l}+\Delta U))\right) \tag{4.26}$$

where n is the number of atoms in the structure.

Results obtained with this scaling method should be checked carefully. The method has been shown to work reasonably well for structures containing atoms with similar atomic mass (Blessing, 1995). However, if a structure has a high variance of atomic masses – which includes studies involving hydrogen and carbon atoms – there is reasonable doubt that the scaling expression is valid. Unfortunately, since hydrogen atom ADPs are not available for XRD structures, it was not possible determine a suitable scaling model.

4.1.5. Validation Against Theoretical Data

First, the ADPs estimated by the presented method (denoted TLS+INV) were compared to those obtained from ONIOM computations (Svensson *et al.*, 1996, Whitten and Spackman, 2006). The ONIOM computations do include the whole geometry data of the structure studied. Hence, if the results obtained by both methods are in good agreement, the transferability and localization assumptions made in section 4.1.2 are shown to be applicable.[5] The overall procedure consisted of the following steps:

1. Performing an ONIOM computation based on the geometry data from XRD.
2. Computing ADPs based on the ONIOM output. (See section 3.1.5 for details).
3. Performing a TLS+ARG analysis using the ONIOM derived ADPs to correct for correlation of internal and external vibrations. (See section 3.1.4 for details)
4. Combining ONIOM ADPs and TLS ADPs to yield values with the designation U_{ij}^{ONIOM}. U_{ij}^{ONIOM} is the sum of $U_{ij,internal}^{ONIOM}$ and $U_{ij,external}^{ONIOM}$

Equivalent values for S were computed for ADPs (designation U_{ij}^{INV}) estimated by the method described in section 4.1.2. The U_{ij}^{ONIOM} set of estimated ADPs is then scaled to the U_{ij}^{INV} set and the spatial overlap S (see equation 4.21) is computed for each pair of equivalent hydrogen atom ADPs. Results are shown in section 4.2.1.

[5] The ONIOM method does not require the localization and transferability approximations made earlier in this chapter since ther is a one-to-one correspondence between the atoms in the structure of interest and the atoms in the ONIOM computation.

4.1.6. Validation Against the SHADE Server

The method most often applied to estimate hydrogen ADPs is the SHADE server (Madsen and Hoser, 2014). Therefore the proposed method is validated against results obtained with the SHADE server. The SHADE server is based on the same assumptions discussed in section 4.1.2 including transferability of vibrational behavior. Instead of estimating displacement amplitudes from theoretical computations, the SHADE server uses a library of small molecule structure models that were refined against Neutron diffraction data. For each structure model a TLS analysis was performed to separate internal and external vibrations. Subsequently, the internal vibrational data was extracted and stored in a database available for transfer to a structure of interest.

Both the proposed method and the SHADE server provide estimates for hydrogen ADPs. The comparison between both methods does not compare the ellipsoids directly but instead compares both estimates to a reference model refined against Neutron diffraction data. The SHADE server has a significant advantage in this comparison study due to the fact that its database is compiled from Neutron diffraction data and therefore shares similar systematic errors as the Neutron diffraction data used as a benchmark.

4.1.7. Influence of Estimated ADPs on Bond Length Accuracy

To assess the impact of the estimation method's result on the overall model accuracy, a bond length accuracy study was performed.

A well known problem of XRD studies is the fact that X–H distances can usually not be determined accurately. Since the centroid of the electron density associated with a hydrogen atom is not at the position of the hydrogen nucleus, the IAM cannot yield accurate hydrogen positions with the standard scattering factor model.[6] One common practice is to place hydrogen atoms at the position of the charge centroid to obtain better figure of merits even though the position is not the correct nuclear position. Other scattering factor models like the multipole model and HAR provide the tools to accurately model hydrogen charge distribution, but the low scattering contribution of hydrogen atoms generally prohibits the optimization of a sufficiently detailed hydrogen model. It was shown that performing HAR to freely optimize hydrogen ADPs and positions improves X–H distance accuracy (Woińska *et al.*, 2016). However, even when high resolution data is available, the optimization often yielded unphysical displacement

[6] Modified scattering factors for hydrogen atoms exist that try to correct for the centroid shift.

ellipsoids.

Therefore it is investigated whether estimated hydrogen ADPs can provide better hydrogen positions and consequently more accurate X–H bond lengths by keeping the ADPs fixed and limiting the refinement to the optimization of the hydrogen positions. This method cuts the number of optimized parameters from nine down to three for each hydrogen atom resulting in a more stable optimization problem. The resulting model uses the same number of parameters as the riding atom model does.

The effect of estimated hydrogen ADPs on the accuracy of bond lengths was investigated by analyzing a series of published test structures. Neutron diffraction data was available for each of the test structures and was used as a benchmark. Complementary ONIOM computations were performed for each structure to serve as an independent, theory derived benchmark.

Each structure was re-refined with three different refinement protocols:

INV Invariom refinement with freely refined hydrogen positions and fixed, estimated hydrogen ADPs.

HAR-Free HAR with freely refined hydrogen positions and freely refined hydrogen ADPs.

HAR HAR with freely refined hydrogen positions and fixed, estimated hydrogen ADPs.

For each model all X–H distances were computed and compared to the corresponding bond distance in the model refined against Neutron diffraction data. The difference for each equivalent atom pair was used to compute

$$wRMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_{ref,i} - d_i}{\sqrt{\sigma_{ref,i}^2 + \sigma_i^2}}\right)^2} \tag{4.27}$$

where d_i denotes a bond distance in a model optimized against X-ray data, σ_i is the corresponding standard deviation and $d_{ref,i}$ and $\sigma_{ref,i}$ are the corresponding values in a reference model (Neutron or ONIOM). The weighted root mean-squared difference ($wRMSD$) is then compared amongst the three refinement protocols for each model.

4.2. Results

4.2.1. Validation Against Theoretical Data

Figures 4.3 to 4.5 show the structural models of the compounds studied in this section.

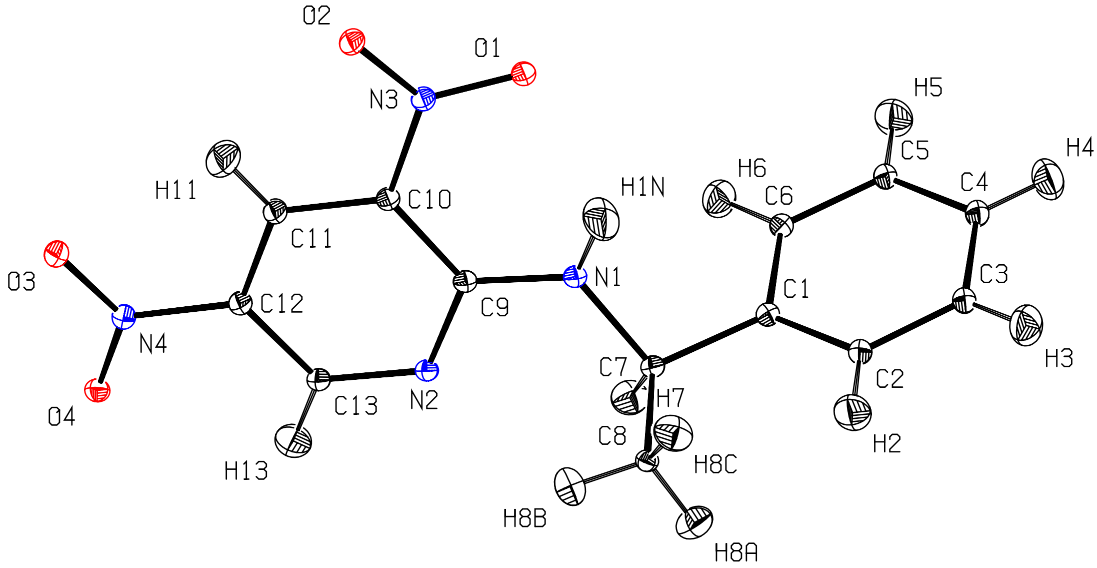


Figure 4.3.: Structural model of Methylbenzylaminodinitropyridine (MBADNP) at 20 K (Cole *et al.*, 2002) with ADPs estimated with the TLS+INV approach.

Tables 4.2.1 to 4.2.1 show the results of the comparison described in section 4.1.5.[7]

Structure 1: MBADNP

The largest discrepancies between the TLS+INV and the TLS+ONIOM method can be observed for hydrogen atom $H1N$. $H1N$ is part of a secondary amine group and is part of an intra-molecular hydrogen bond to oxygen atom $O1$. The ONIOM method

[7] ONIOM computations for Dimethylbiguanidiniumbishydrogensquarate did not converge and were therefore not included in the comparison.

Label	S	Label	S
H11	0.31	H5	0.02
H13	0.07	H6	0.03
H1N	1.56	H7	0.02
H2	0.04	H8A	0.75
H3	0.03	H8B	0.52
H4	0.08	H8C	0.57
$\langle H \rangle$	0.33		

Table 4.1.: Comparison of TLS+INV derived ADPs to TLS+ONIOM derived ADPs of MBADNP.

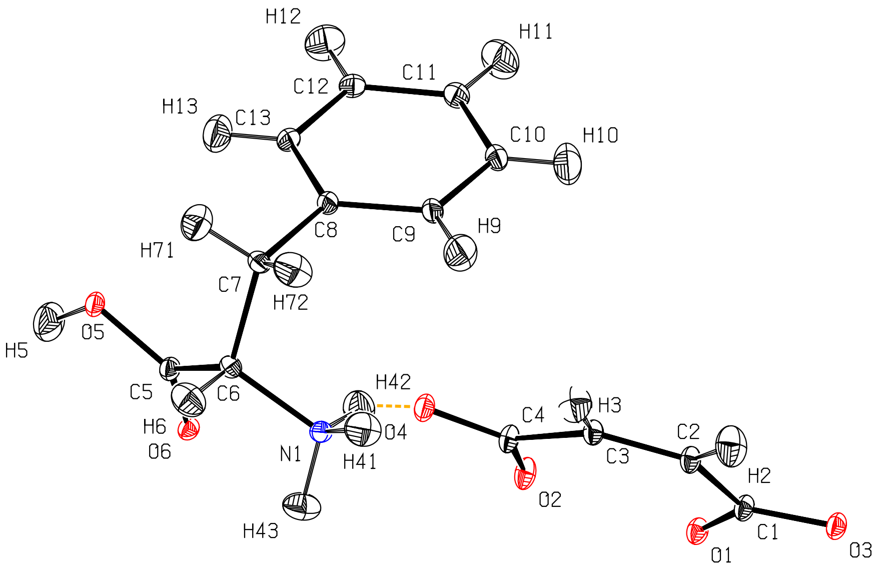


Figure 4.4.: Structural model of L-phenylalaninium hydrogen maleate at 12 K (Grabowsky *et al.*, 2014) with ADPs estimated with the TLS+INV approach.

computes the vibrational behavior of an atom based on the whole molecule cluster at once. This means that that the program performing the computation is *aware* of the hydrogen bonding between $H1N$ and $O1$. The TLS+INV method on the other hand reconstructs the internal vibrational behavior of an atom from small fragments. This implies that the vibrational behavior of atom $H1N$ is estimated in the absence of atom $O1$. This results in a model that does not take the hydrogen bond between both atoms into account which explains the large value of S for atom $H1N$.

The next three largest values of S are observed for the methyl group hydrogen atoms $H8A$, $H8B$ and $H8C$. This is most likely caused by similar effects: atom $H8B$ is part of a weak hydrogen bond to nitrogen atom $N2$. This stabilizing effect can not be modeled by the TLS+INV model. If $H8B$ is stabilized, the same stabilization will effect $H8A$ and $H8C$.

Overall, the agreement between TLS+ONIOM and TLS+INV is good.

Structure 2: L-phenylalaninium Hydrogen Maleate

The agreement of the TLS+ONIOM and the TLS+INV is less good for this structure. The biggest differences are again observed for those hydrogen atoms that are part of

Label	S	Label	S
H10	1.52	H42	1.36
H11	1.41	H43	1.22
H12	1.43	H5	5.13
H13	1.97	H6	1.47
H2	0.64	H71	1.85
H3	1.57	H72	2.54
H41	4.15	H9	1.81
$\langle H \rangle$	2.00		

Table 4.2.: Comparison of TLS+INV derived ADPs to TLS+ONIOM derived ADPs of L-phenylalaninium hydrogen maleate.

Label	S	Label	S
H11	3.51	H1B	0.81
H12	9.58	H2	0.33
H13	3.84	H3	0.20
H14	13.60	H4	0.52
H15	10.52	H5A	0.37
H1A	0.80	H5B	0.57
$\langle H \rangle$	3.74		

Table 4.3.: Comparison of TLS+INV derived ADPs to TLS+ONIOM derived ADPs of Xylitol.

hydrogen bonds: $H5$ and $H41$. The overall worse agreement can be explained by the fact that the asymmetric unit contains two molecules instead of one. The rigid body model used for the TLS-analysis assumes that the whole asymmetric unit forms a rigid body. This is not a good approximation if two or more independent molecules are present. A more flexible rigid body model was tested but did not yield stable results due to the small number of non-hydrogen atoms in the $C_3O_4H_2$ unit.

Structure 3: Xylitol

The comparison study with the structural model of Xylitol confirms the previously observed trend: hydrogen bonding is not taken into account by the TLS+INV method and leads to less good agreement between both models. This effect is particularly prominent for this structure where most hydrogen atoms are involved in hydrogen bonding.

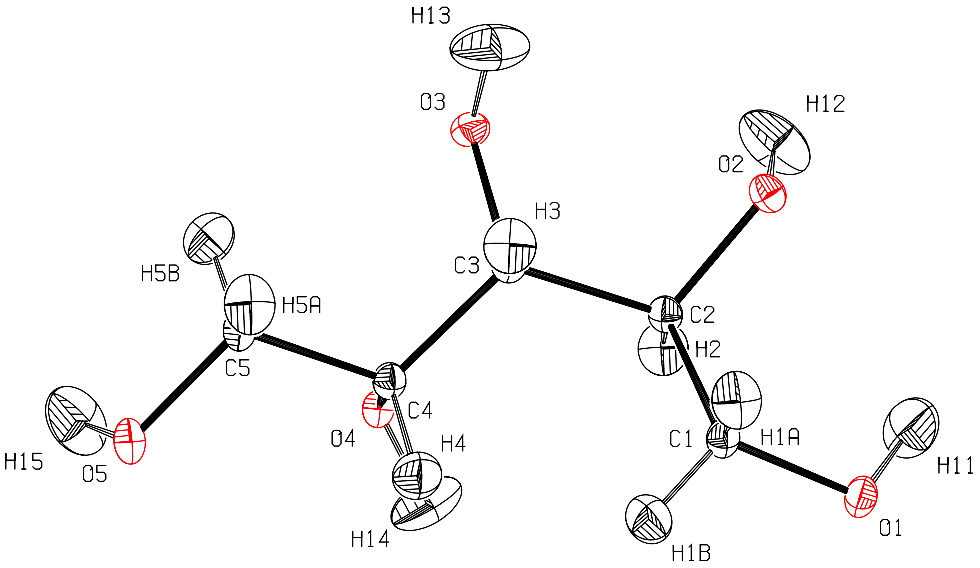


Figure 4.5.: Structural model of Xylitol at 122 K (Madsen *et al.*, 2003) with ADPs estimated with the TLS+INV approach.

Conclusion

The data shows that the agreement between U_{ij}^{ONIOM} and U_{ij}^{INV} depends on whether hydrogen atoms are involved in non-covalent interactions. The TLS+ONIOM approach takes non-covalent intra- and inter-molecular forces into account, which the TLS+INV approach can not. Additional forces acting on an atom dampen its vibrational movement, which explains differences between both approaches for those atoms. Another factor is the overall rigidity of the content of the asymmetric unit. Xylitol and MBADNP are both the only molecules in their respective asymmetric unit. Hydrogen atoms that are not involved in hydrogen bonding are in very good agreement in both structural models. The structure of L-phenylalaninium hydrogen maleate is much more flexible and therefore the localization approximation – assuming that vibrations of atoms in the atomic framework is only dependent on its immediate chemical environment – is not fulfilled as well as it is for a more rigid system. In addition to hydrogen bonding, the influence of non-covalent interactions on the displacement amplitudes can also be seen for the hydrogen atoms in methyl groups. Without taking non-covalent interactions into account properly, methyl groups can rotate more freely than they would in the real crystal lattice where atoms are kept in place by weak interactions with other atoms. This leads to larger ADP of methyl hydrogen atoms estimated by the TLS+INV method. However, overall the ADP estimated by both methods are in acceptable agreement. While the localization and transferability approximations are not fulfilled perfectly, estimated ADPs can still yield reasonable results.

4.2.2. Validation Against SHADE Server

The similarity of hydrogen ADPs between the TLS+INV and the SHADE model, which will be denoted TLS+NEUT here for consistency, is listed in tables 4.4 to 4.7. A summary of the data is shown in figure 4.7.

Structure 1: MBADNP

Table 4.4 lists the comparison result values for structure MBADNP. The TLS+NEUT results are slightly closer to the Neutron diffraction data. Overall, the results are consistent with the results from section 4.2.1.

Structure 2: L-phenylalaninium Hydrogen Maleate

Table 4.5 lists the results of the L-phenylalaninium hydrogen maleate comparison study. The results are again consistent with the results from section 4.2.1: since both the TLS+INV and the TLS+NEUT model to not account for non-covalent interaction, the estimation methods are less accurate overall. The large discrepancy for atom $H71$ is most likely due to ill-determined displacement parameters in the neutron refinement as becomes obvious from visual inspection.

Label	S$_{INV}$	S$_S$	Label	S$_{INV}$	S$_S$
H11	0.44	0.23	H5	0.75	0.28
H13	0.12	0.03	H6	1.17	0.27
H1N	1.35	0.39	H7	0.11	0.14
H2	0.17	0.09	H8A	1.76	1.30
H3	0.92	0.18	H8B	2.38	1.02
H4	0.17	0.14	H8C	2.21	0.90
$\langle H \rangle$	0.96	0.42			

Table 4.4.: Comparison of TLS+INV (S$_{INV}$) derived ADPs to SHADE (S$_S$) ADPs for the example of MBADNP.

Label	S$_{INV}$	S$_S$	Label	S$_{INV}$	S$_S$
H10	3.84	0.52	H42	4.80	0.70
H11	3.19	0.61	H43	3.91	1.08
H12	2.31	0.52	H5	3.82	1.33
H13	4.10	1.49	H6	2.94	0.67
H2	2.05	1.05	H71	13.68	5.71
H3	2.27	0.67	H72	1.90	0.38
H41	4.57	0.73	H9	3.22	0.90
$\langle H \rangle$	3.30	1.17			

Table 4.5.: Comparison of TLS+INV (S$_{INV}$) derived ADPs to SHADE (S$_S$) ADPs for the example of L-phenylalaninium hydrogen maleate.

Label	S$_{INV}$	S$_S$	Label	S$_{INV}$	S$_S$
H11	3.55	0.58	H1B	2.45	0.74
H12	2.85	0.49	H2	0.62	0.55
H13	3.76	0.24	H3	0.07	0.09
H14	1.92	0.91	H4	0.28	0.10
H15	2.47	0.41	H5A	3.41	1.68
H1A	2.46	0.78	H5B	2.97	1.83
$\langle H \rangle$	2.24	0.70			

Table 4.6.: Comparison of TLS+INV (S$_{INV}$) derived ADPs to SHADE (S$_S$) ADPs for the example of Xylitol.

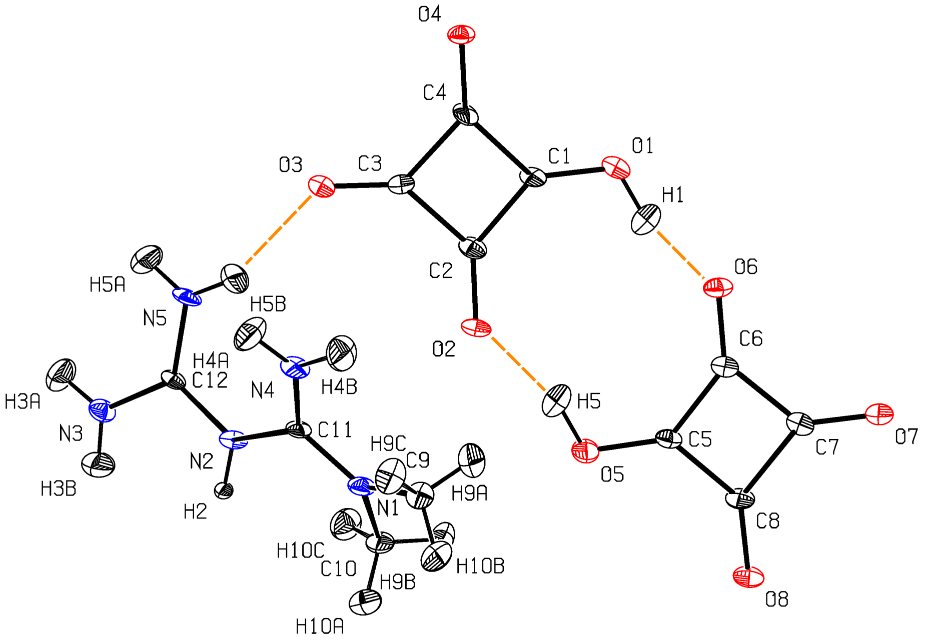


Figure 4.6.: Structural model of Dimethylbiguanidiniumbishydrogensquarate at 130 K (Şerb *et al.*, 2014) with ADPs estimated with the TLS+INV approach.

Structure 3: Xylitol

Table 4.6 lists the results of the Xylitol comparison study. The TLS+NEUT method performs very well for this structure and seems to be affected less severely by hydrogen bonding as the TLS+INV method is. A possible explanation is that the Neutron structure database utilized by the TLS+NEUT method extracted the displacement amplitudes from hydrogen atoms in structures that also showed hydrogen bonding.

Both methods perform very well for the hydrogen atoms not involved in hydrogen bonding: $H2$, $H3$ and $H4$.

Structure 4: Dimethylbiguanidiniumbishydrogensquarate

Table 4.6 lists the results of the Dimethylbiguanidiniumbishydrogensquarate comparison study. The TLS+INV method performs slightly better than the TLS+NEUT method

Label	S$_{INV}$	S$_S$	Label	S$_{INV}$	S$_S$
H1	2.84	0.70	H4B	1.15	0.94
H10A	1.77	2.73	H5	4.04	0.37
H10B	2.15	3.97	H5A	0.73	0.06
H10C	1.60	2.42	H5B	0.70	0.13
H2	1.38	0.98	H9A	1.25	3.51
H3A	1.14	0.63	H9B	1.35	3.17
H3B	0.95	0.97	H9C	0.40	2.00
H4A	1.01	1.15			
$\langle H \rangle$	1.50	1.58			

Table 4.7.: Comparison of TLS+INV (S$_{INV}$) derived ADPs to SHADE (S$_S$) ADPs for the example of Dimethylbiguanidiniumbishydrogensquarate.

in this case. The most likely reason for this is that little Neutron diffraction data containing squarate (or similar) elements that could be used for the SHADE servers Neutron structure database. Here the TLS+INV method profits from the fact that even *exotic* chemical environments are available for estimation purposes.

Overall, both methods perform well considering that three molecules are in the asymmetric unit.

Conclusion

Figure 4.7 shows an overview of the average S values of all compared structures. Considering that the SHADE server is based on transferring parameters refined against Neutron diffraction data, one should note that this method might benefit from error cancellation when comparing the results to those from Neutron diffraction studies. Keeping that in mind the results are very promising. The XRD structures were modeled with the IAM. Using a different scattering factor model e.g. the invariom model yields apparently random shifts in the values displaced in the tables. The random shift has an amplitude of about 0.5 to 0.7. Since the shifts seem to be random without any systematic trends, it is reasonable to assume that the error of each of the displayed values is approximately 0.7. It is also worth noting that the SHADE server was initially calibrated against the structural model Xylitol. Therefore it is not unexpected to see it perform exceptionally well when estimating hydrogen ADP for the Xylitol structure.

Even though the SHADE server seems to outperform the proposed TLS+INV method, the new method still has significant advantages. The results of the TLS+INV method do not depend on experimental data (with the exception of the experimental data of the

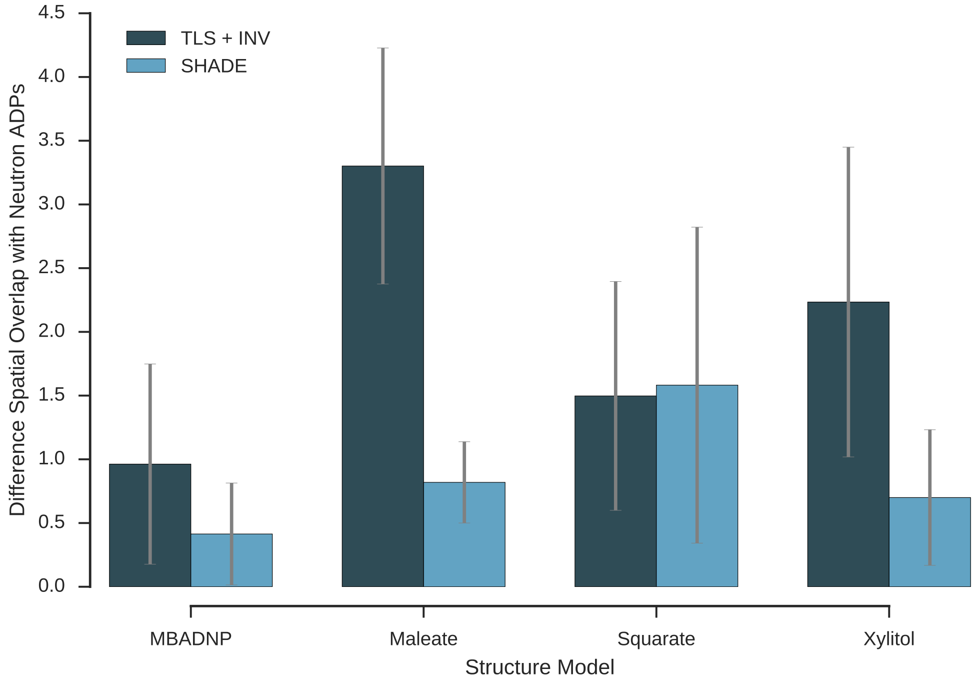


Figure 4.7.: Plot visualizing the fit of estimated ADPs to ADPs refined against Neutron diffraction data. The y-axis shows the mean difference between the estimated ADPs and the reference Neutron model. The error bars show its standard deviation.

structure the method is applied to). Consequently, the method can be applied to a much larger set of structure that are not available to the SHADE server since crystals suitable for Neutron diffraction experiments cannot be grown for all systems. Also, it might not be feasible to systematically generate a complete database of Neutron diffraction based model compounds because the diffraction experiments are time consuming and experiment time is limited. The TLS+INV method also gives access to individual vibrational frequencies that are only at run-time converted to displacement amplitudes compatible with the structural model's parametrization. This facilitates the significant advantage of taking the measurement temperature into account when estimating the displacement. The importance of this feature is demonstrated and thoroughly discussed in chapter 3. Furthermore, the access to individual vibrational modes is beneficial when studying thermodynamic properties of the system. Additionally, the TLS+INV method can leverage the power of the invariom database including all its properties like the scattering factor data base and point charge computation facilities.

Another advantage of the proposed method is the possibility to take anharmonicity of the X–H bond stretching mode into account. A reasonable assumption is that X–H bond stretching modes are not harmonic in nature. The contraction of the bond should require more energy than the elongation of the bond yielding an energy profile close to the Lennard-Jones potential. Computing a fully anharmonic description of an atom from theory is not feasible due to the enormous amount of potential energy surface samples that would need to be computed. However, sampling the energy of differently placed hydrogen atoms along the bond vector could be manageable. These energy samples could then be used to fit a one dimensional anharmonic potential which could be transferred to experimental samples. Assuming that the anharmonicity of the bond stretching mode is the most significant deviation from harmonic motion, this approach could yield a quasi-anharmonic description of hydrogen atoms. That description would be transferable following the invariom approach. Corresponding experiments have not been performed yet but could be useful if highly detailed vibrational descriptions of hydrogen atoms are needed.

4.2.3. Influence of Estimated ADPs on Bond Length Accuracy

The improvement of bond length accuracy upon introduction of estimated hydrogen ADP is investigated. Table 4.8 lists an overview of the analyzed structures. Neutron diffraction data is available for structures ASN, GLN, SER, THR and HYPRO. For these structures a model refined against Neutron diffraction data serves as a reference struc-

Structure	Spacegr.	Temp.	Source type	Reference
D,L-Asparagine · H_2O (ASN)	$P2_12_12_1$	100K	Synchrotron	(Flaig *et al.*, 1999)
		RT	Neutron	(Verbist *et al.*, 1972)
L-Glutamine (GLN)	$P2_12_12_1$	100K	Mo Kα	(Wagner and Luger, 2001)
		RT	Neutron	(Koetzle *et al.*, 1973)
L-Phenylalanine (PHE)	$P2_1$	25K	Mo Kα	(Mebs *et al.*, 2006)
D,L-Proline·H_2O (PRO)	$Pbca$	100K	Synchrotron	(Koritsánszky *et al.*, 1998)
D,L-Serine (SER)	$P2_1/a$	100K	Synchrotron	(Flaig *et al.*, 1999)
		RT	Neutron	(Frey *et al.*, 1973)
L-Threonine (THR)	$P2_12_12_1$	19K	Ag Kα	(Flaig *et al.*, 1999)
		RT.	Neutron	(Ramanadham *et al.*, 1973)
D,L-Valine (VAL)	$P\bar{1}$	100K	Synchrotron	(Flaig *et al.*, 2002)
N-acetyl-L-4-Hydroxy-proline · H_2O (HYPRO)	$P2_12_12_1$	100K	MoKα	(Lübben *et al.*, 2014)
		150K	Neutron	(Lübben *et al.*, 2014)
D,L-Glutamic acid·H_2O (GLU)	$Pbca$	100K	Synchrotron	(Flaig *et al.*, 1999)

Table 4.8.: Overview of the structures studied in the context of improved bond length accuracy.

ture. Generally, the comparison studies demonstrate that the geometry parameters obtained via ONIOM computation are in excellent agreement with the Neutron models (Figure 4.8 (top, label=ONIOM)). This justifies to use ONIOM computation results as the reference structural model in cases where no Neutron diffraction data is available, namely structures GLU, PHE, PRO and VAL.

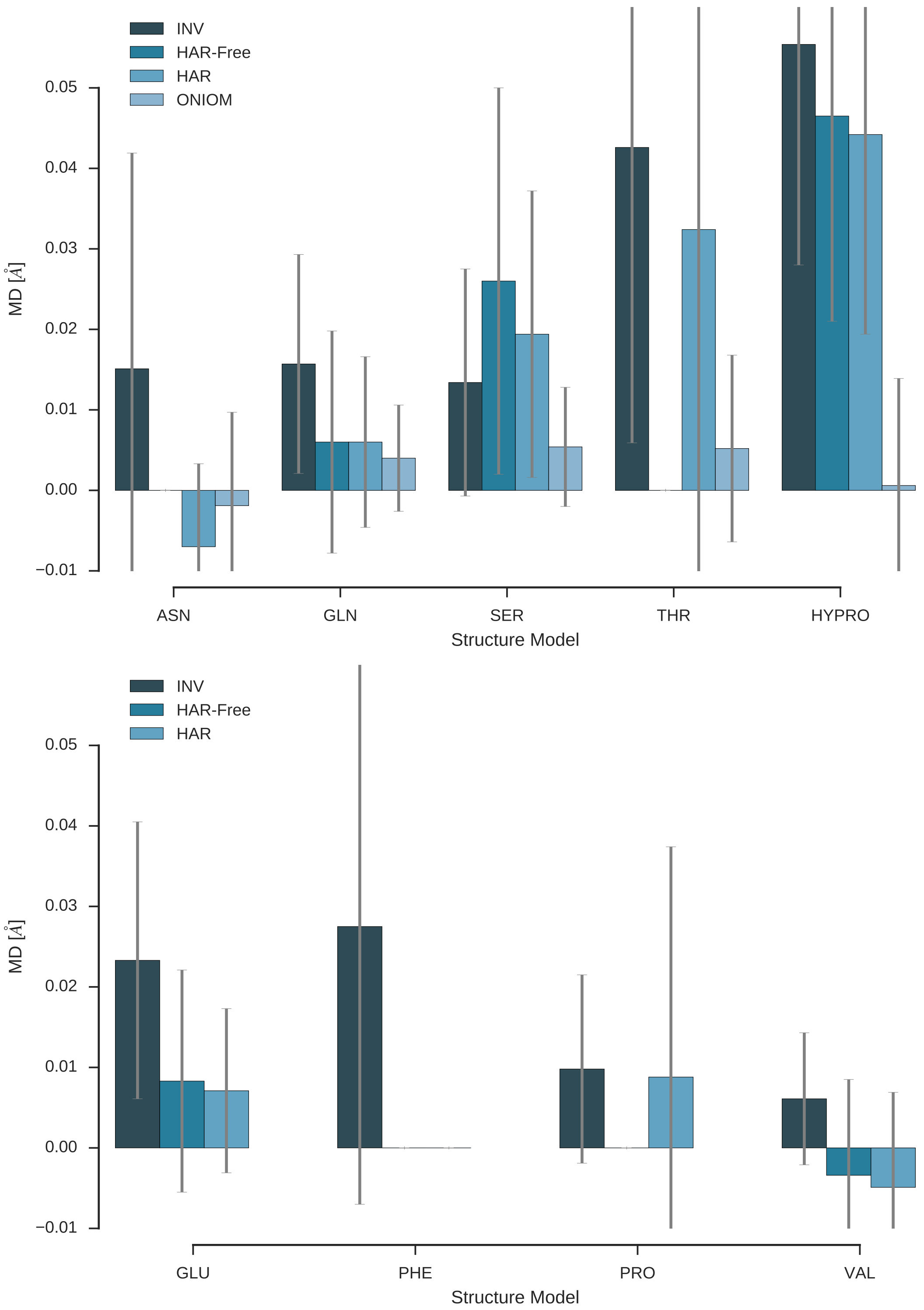


Figure 4.8.: Average difference between X–H bond lengths in the refined models and the reference Neutron model (top) or the ONIOM (bottom) model. Refinements yielding non-positive definite ADPs or fail to converge are omitted.

Figure 4.8 shows the results of the comparison study. The freely refined HAR model fails to converge or yields non-physical displacement parameters for structures ASN, THR, PHE and PRO. This makes routine application of this refinement protocol not recommended since no consistent results can be obtained. In addition to not yielding meaningful models, the results show bigger differences to the reference model than the HAR model with estimated hydrogen ADPs. The fact that the model with fewer parameters yields more accurate results is a clear indication for overfitting in the case of the freely refined HAR model. The optimization of the hydrogen ADPs most likely fits errors in the diffraction data instead of actual vibrational behavior of hydrogen atoms.

The overall less flexible invariom model that does not facilitate the optimization of heavy atom asphericity but rather constrains them to tabulated values yields useful results as well. As expected, the accuracy is not as good as the very flexibly parametrized HAR model but it is the only model that reached quasi convergence and physically plausible results in all studied cases. In the case of structure SER the invariom model yields the most accurate result. Considering that the invariom model uses a less flexible parametrization that should not be able to yield the highest accuracy the most probable reason for this is less accurate or less precise data. This can be considered an important point in favor of the invariom model implying that it is significantly more robust in presence of imperfect data.[8]

This study demonstrates that hydrogen ADPs estimated via the proposed method provide a significant improvement to structural model accuracy without adding parameters to the model. The optimization of hydrogen ADPs, even against high resolution data, cannot be recommended in general.

[8] It is reasonable to assume that the invariom model works particularly well in cases were little intermolecular interaction is involved due to fact that the invariom model does not take these interactions into account.

5. Disorder in *N*-Acetyl-L-Hydroxyproline Crystals

Two crystal structures of *N*-Acetyl-L-Hydroxyproline are investigated. The first structure (*anhydrate*) contains the pure compound. The second structure (*monohydrate*) is the monohydrate of the compound. Both crystalize in space group $P2_12_12_1$ (see Figure 5.1 for information on the crystal packing). The second structure's unit cell is expanded slightly. Except for the additonal H_2O molecule the main difference between both forms is a rotational disorder of the acetyl methyl group. The methyl group in structure 1 is disordered at all investigated temperatures while the hydrate form shows no (or very little) signs of rotational disorder at very low temperatures.[1] The most intuitive explanation – a stabilizing hydrogen bond between the methyl group and the water molecule – can be excluded due to a very long H–O distance in the crystal lattice (Table 5.1). There must hence be another reason for the temperature dependent occurence of disorder in the molecule which will be investigated in this chapter.

Neutron diffraction data, as well as XRD data was collected at different temperatures. The disorder of the methyl group was analyzed in detail and the hydrogen density distribution in the vincinity of the methyl group's carbon atom was investigated.

5.1. Methods

5.1.1. Experimental Details

Crystals of the anhydrate were grown by slowly cooling a saturated solution of *N*-Acetyl-L-Hydroxyproline in hot acetone dryed with CaH_2 to room temperature. Crystals grow to sizes of 0.5 mm. Crystals of the monohydrate are formed by incorporation of water into the crystal lattice at ambient conditions.

Crystals of both the monohydrate and the anhydrate of *N*-Acetyl-L-Hydroxyproline were measured at multiple temperatures to investigate the respective temperature de-

[1] Temperatures from $6\ \mathrm{K}$ to $100\ \mathrm{K}$ were investigated.

	Hydrate		Anhydrate	
Hydrogen	Oxygen	Distance	Oxygen	Distance
Atom	Atom	[Å]	Atom	[Å]
H121	O11	2.558	O11	2.559
	O11^{3}	2.598	O9^{5}	2.931
	O13^{2}	2.962	O11^{4}	3.054
	O1^{3}	2.991	O1^{4}	3.102
H122	O1^{2}	2.525	O1^{2}	2.449
	O11	3.040	O1^{5}	2.859
	O9^{5}	3.102	O11^{5}	2.919
	-	-	O11	2.989
H123	O13^{3}	2.621	O11^{5}	2.584
	O9^{4}	2.962	-	-
	O11	3.187	-	-
	O11^{7}	3.198	-	-

Table 5.1.: Hydrogen – Oxygen distance table. Each table section lists H-O contacts shorter than 3.2 Å for the hydrate and the anhydrate. The superscript numbers correspond to the symmetry operations used to generate the atom from the asymmetric unit. Symmetry operations are listed in table 5.2.

Number	Symmetry Operation
1	$1/2 + x, 1/2 - y, 1 - z$
2	$1 - x, -1/2 + y, 3/2 - z$
3	$1/2 + x, 1/2 - y, 1 - z$
4	$1 - x, -1/2 + y, 1/2 - z$
5	$3/2 - x, 1 - y, -1/2 + z$
6	$1 - x, -1/2 + y, 1/2 - z$
7	$1 - x, -1/2 + y, 3/2 - z$

Table 5.2.: Symmetry operations of both the hydrate and the anhydrate structure.

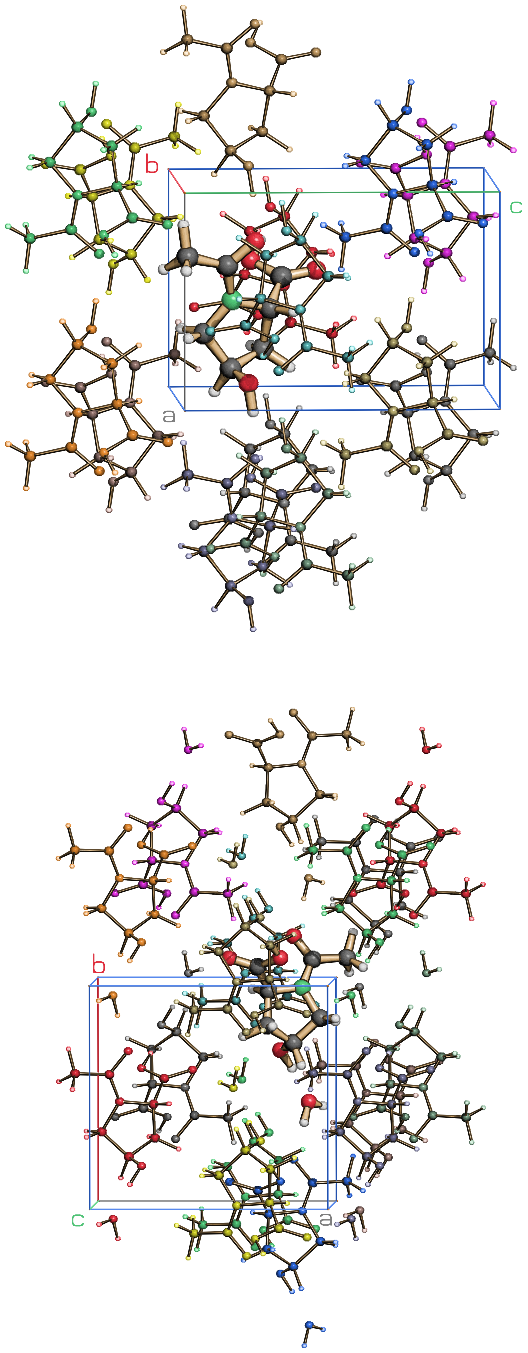


Figure 5.1.: Crystal packing of the anhydrate (top) and the hydrate (bottom).

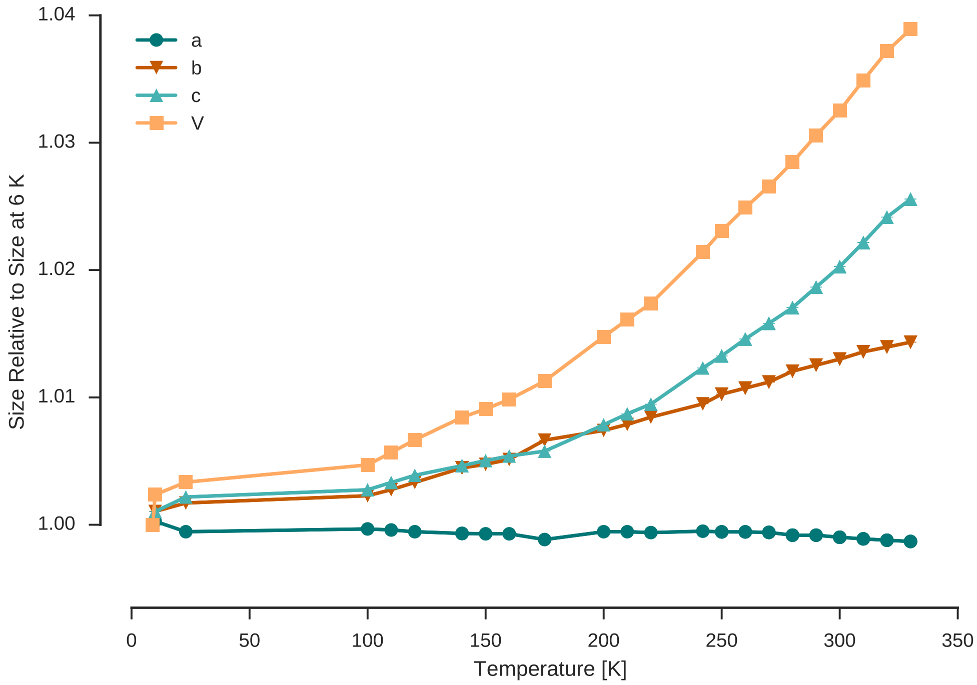


Figure 5.2.: Temperature dependence of the lattice constants of *N*-Acetyl-L-Hydroxyproline anhydrate. Cell constants cannot be determined reliably via quasi-Laue Neutron diffraction.

pendence of rotational disorder. At selected temperatures both an XRD and a Neutron diffraction data set was collected. Measurement temperatures for the monohydrate were 9 K, 150 K, 200 K and 250 K. Due to the more pronounced disorder in the anhydrate, a lower temperature range was selected to get more insight in the onset of disorder. Therefore diffraction data of the anhydrate was measured at 6 K, 23 K, 40 K and 100 K. XRD data of the anhydrate form was collected at several additional temperatures to obtain cell constants with higher accuracy (Figure 5.2).

X-Ray Data Collection

For the multi-temperature experiment of the anhydrate XRD data was collected at the HASYLAB synchrotron facility at beamline P11. Data sets at 9 K, 150 K, 200 K and 250 K were collected at constant temperatures. All additional data sets were collected while slowly raising the temperature of the cryo-stream device during data collection. A short data aquisition time of two minutes makes the data aquisition temperature constant to a good approximation.

The detector distance was set to $137\ \mathrm{mm}$, which was the minimum distance possible. The angle increment was $0.5°$ and a single (or several, visible from the total number of reflections collected below in table 5.3) Φ scan was performed with a single crystal orientation. To minimize radiation damage the crystal was translated by a small amount for each measurement, whereas the overall orientation was left unchanged. The detector used was a DECTRIS PILATUS 6MF pixel array counter. The area detector resolution of this detector is 2463×2527 pixels, with an individual pixel size of $172e^{-6}\ \mathrm{m} \times 172e^{-6}\ \mathrm{m}$. The XDS software (Kabsch, 2010) was used for data integration. The non-active area in between individual counter elements of the detector was masked out and thus not taken into account during integration. Beam divergence and reflecting range were adjusted to values determined by the software after the first pre-integration run. Since for a small molecule there are less reflections per frame than for a macromolecule the XDS software parameter DELPHI was increased to $10°$, which led to more reflections being used in intermediate unit-cell dimension determination during integration. Unit-cell parameters (and their standard uncertainties) were determined from averaging these individual unit-cell determinations (and from computing their variance). During the measurement each frame was irradiated for $0.19\ \mathrm{s}$, which together with a readout period of $0.01\ \mathrm{s}$, gave an exposure time of $0.2\ \mathrm{s}$ per frame. Individual measurements hence took a bit more than two minutes ($144\ \mathrm{s}$) for a $360°$ rotation and the 720 frames collected in each of these single runs. For background determination 144 out of the 720 frames were averaged. The crystal size, with $0.42 \times 0.32 \times 0.31\ \mathrm{mm}$ was a lot bigger than the beam size of $50\ \mu\mathrm{m}$. The program sadabs was used for scaling after conversion of the XDS output file format into a file readable by sadabs with the utiliy xds2sad by G. M. Sheldrick. sadabs was also used to generate an xd.hkl file, where the data were merged according to point group mmm. Systematic absent reflections for space group $P2_12_12_1$ were also eliminated in this processing step.

Data collection of the monohydrate has been described in detail in section 3.1.

Neutron Data Collection

Neutron diffraction data was collected at the *KOALA* beamline of the Australian Nuclear Science and Technology Organization. Data was collected with an Oxford Instruments Image Plate detector and processed with the LAUEG software package (Campbell, 1995).

Temperature [K]	100	110	135	140	150	160	185	200	215
meas. data	9704	5260	5263	5298	5242	5278	5155	5320	5345
unique data	1651	1652	1650	1650	1651	1659	1656	1665	1671
Temperature [K]	230	250	260	270	280	290	300	310	320
meas. data	5372	5341	5336	5309	5232	5355	5374	5289	5314
unique data	1681	1684	1681	1683	1701	1699	1710	1694	1696

Table 5.3.: Number of collected reflections at all measurement temperatures for the structure of anhydrate.

	9 K	150 K	200 K	250 K
Radiation	Neutrons	Neutrons	Neutrons	Neutrons
Unique Reflections	1355	1313	1314	1292
Completeness	65.3	64.2	64.7	63.9
$I\sigma$	51.98	32.00	31.26	24.34
Resolution	0.65	0.65	0.65	0.66
Space Group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
a	9.854(3)	9.9408(2)	9.9748(2)	10.0123(2)
b	9.249(3)	9.2479(2)	9.2492(2)	9.2556(2)
c	10.144(2)	10.1875(2)	10.2103(2)	10.2441(2)
R_1	0.0298	0.0463	0.0420	0.0535

Table 5.4.: Overview of the Neutron diffraction data sets of *N*-Acetyl-L-Hydroxyproline monohydrate.

Structure Solution and Refinement

Structure solution was performed with SHELXT (Sheldrick, 2015*a*) for all XRD data. Structures were refined with SHELXL (Sheldrick, 2015*b*). Subsequent invariom refinement was performed for structures refined against XRD data. Density grids were computed with XD2006 (Volkov *et al.*, 2006). Structural models optimized against Neutron diffraction data were not solved *ab initio*. Instead a model optimized against XRD data collected at the same temperature provided starting values for refinement with SHELXL.

5.1.2. Generating Hydrogen Density Plots

It is useful to analyze the nuclear density distribution of hydrogen atoms in the vicinity of the methyl group's carbon atom to get detailed insight into disorder of the methyl group. This was achieved by analyzing the hydrogen density map of a slightly modified structural model. The following steps were performed for each Neutron diffraction data set:

	6 K	23 K	40 K	100 K
Radiation	Neutrons	Neutrons	Neutrons	Neutrons
Unique Reflections	2108	2142	2152	3964
Completeness	71.9	72.9	73.0	73.3
$I\sigma$	29.23	31.71	31.54	29.5
Resolution	0.55	0.55	0.55	0.55
Space Group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
a	7.320(13)	7.316(13)	7.316(13)	7.318(13)
b	10.533(18)	10.551(18)	10.572(18)	10.557(18)
c	10.558(2)	10.581(2)	10.603(2)	10.587(2)
R_1	0.3987	0.0700	0.0745	0.092

Table 5.5.: Overview of the Neutron diffraction data sets of *N*-Acetyl-*L*-Hydroxyproline anhydrate.

- First, the structure was modeled in as much detail as possible taking overfitting into account. In this case this resulted in a fully anisotropic model with RIGU restraints on all non-hydrogen atoms.[2]
- The model was refined to quasi convergence.
- The methyl group's hydrogen atoms were removed and all parameters were constrained to their current values.
- A single refinement step was performed to obtain the Fourier transform of the structural model but without the contribution of the hydrogen atoms of interest.

This resulting density model is suitable for analyzing the hydrogen density distribution (HDD) of the methyl group with minimal phase errors.

The overall density map of the structural model is available as a three dimensional grid were each grid point samples the density at the corresponding point in the crystal lattice. The grid stores values for discrete points in space and not as a continuous function. Therefore, the density value for a point in space that does not directly correspond to a grid point must be interpolated. Several interpolation methods exist. The most simple one – linear interpolation – determines the value at a position between two grid points assuming a linear function. This means that a position that is half-way between point A and point B has a corresponding value of $(A+B)/2$. While this method is fairly simple, the resulting HDD is not continuously differentiable and will look jagged. This

[2] The effect of modelling slightly disordered parts of the model with multiple conformations was investigated but deemed unnecessary in this context.

can be circumvented by applying quadratic interpolation where the three closest points are determined. A quadratic function is then constructed from the three grid values and used to interpolate values in between. This results in a continuously differentiable interpolated density. However, the density's first derivative will not be. Since the density's derivative might be of interest when analyzing the HDD, cubic interpolation was applied. Cubic interpolation takes a fourth grid point into account by fitting a cubic function to the four closest grid points.

Appropriate points needed to be chosen to sample the HDD. The points of interest were all possible positions the hydrogen atoms of the methyl group can have while rotating about the R–C bond. These points were computed by generating an arbitrary point based on documented 1–2, 1–3 and 1–4 distances of a methyl group's hydrogen atom. That point was then rotated about the R–C axis in steps of 1° while the density was interpolated for each point. The result was plotted against the rotation angle and yields the HDD for that methyl group.

The resulting HDD should obey the three-fold local symmetry of the methyl group. This means that every 120° the HDD should repeat. This side condition was used to estimate the error of the HDD at any given point by computing the mean and the standard deviation based on three supposedly equivalent points. Thus, the final plot is the superposition of the plot with itself, off-set by 120° and 240°.

5.2. Results

5.2.1. Hydrogen Density Distribution

Figures 5.3 and 5.4 show the HDD at the potential positions of methyl-group hydrogen atoms at different temperatures. The local three-fold symmetry is taken into account. Therefore only a 120° section centered at the most likely hydrogen position at the lowest temperature is plotted.

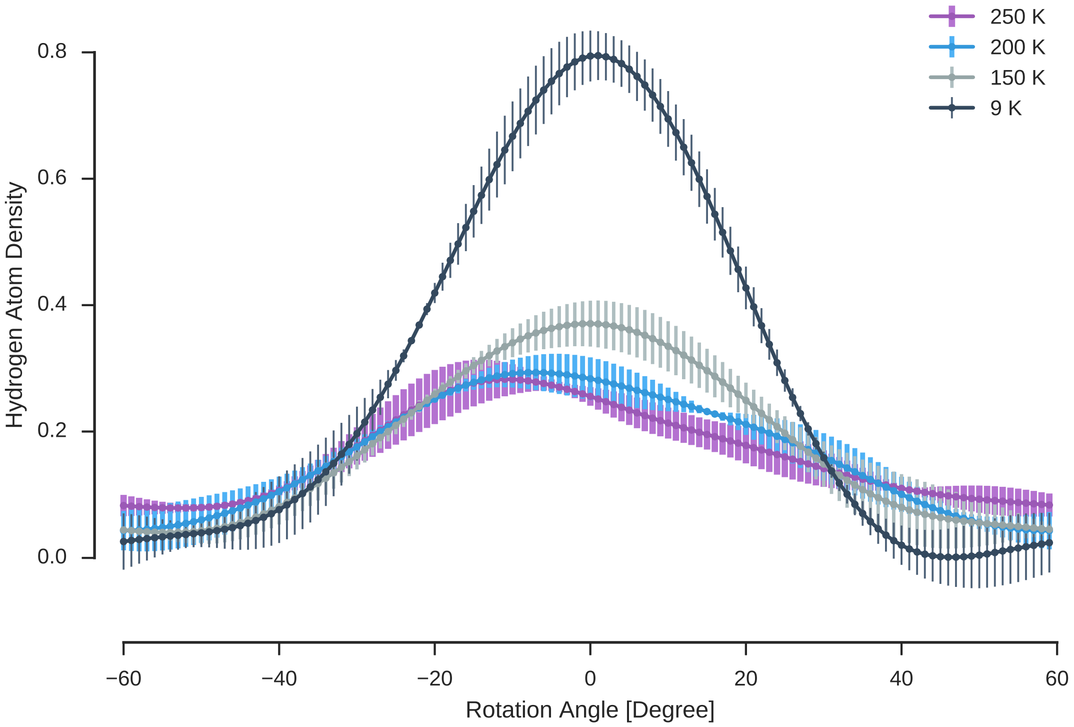


Figure 5.3.: HDD of *N*-Acetyl-L-Hydroxyproline monohydrate.

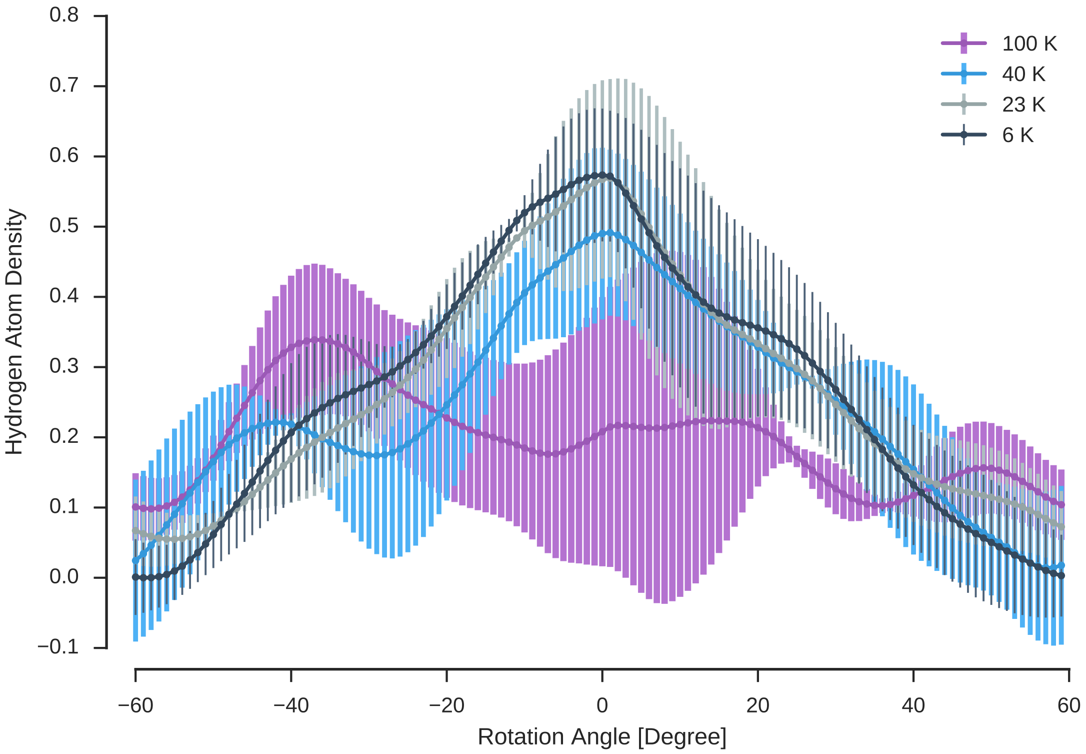


Figure 5.4.: HDD of *N*-Acetyl-L-Hydroxyproline anhydrate.

Figure 5.3 shows that the monohydrate form has well defined hydrogen atom positions even at temperatures above 100 K.[3] The anhydrate shows very different behavior. At a temperature of 6 K the distribution is comparable to the monohydrate form. However, the variance of equivalent density points is significantly higher than the variance in the monohydrate form. At temperatures above 23 K the preferred conformation begins to disappear giving rise to a more disordered structure. The data indicates that at about 40 K a second conformation becomes meta-stable. However, the large estimated error for the density values renders reliable interpretation of the data nearly impossible. At 100 K the second conformations appears to become favored over the low-temperature conformation. The very large error estimates – especially for the data series at 100 K – make it impossible to draw further conclusions.

What can be extracted from the presented data is that the methyl group is stabilized in the monohydrate form. The following explanations are hypothesis based on the limited data available. The effect of hydrogen bonding between methyl group hydrogen atoms and the water molecule's oxygen atom can be excluded as an explanation for the temperature dependent behavior. The structures show no H–O contacts in the range relevant for hydrogen bonding. It is possible that the anhydrate form has multiple local minima in which the methyl group gets *locked-in* during shock freezing. This would explain the shoulders in Figure 5.4. However, if that was the case, the shoulders should become less pronounced upon slowly raising the temperature. Instead, the shoulders become bigger which implies that enough thermal energy is available to cross the rotational barrier. This further implies that the absolute minimum in the rotational potential could be reached from potential local minima of higher energy. This question could potentially be answered with spectroscopic methods that would allow to probe excitations corresponding to a librational vibration about the R–C axis.

Even though no conclusive explanation for the different properties of both structures can be provided at this this point, the data is still a valuable basis for further investigations.

[3] This is represented in the figure by well defined maxima and low estimated standard uncertainties.

Part III.

Validation

6. Validation of Atomic Displacement Parameters

The vast majority of structural models in small molecule crystallography parametrize thermal displacement amplitudes as anisotropic ADPs. In fact, common structure publication procedures require authors of to justify their modeling choices if they chose a different parametrization. Publication procedures also require the structural model to be analyzed in order check for errors which includes validation of the ADPs. Unfortunately, the automated validation procedure (CheckCIF) that is commonly used (Spek, 2009) is not perfect. Especially the method used to analyze ADPs doesn't work well in certain cases.

This section discusses improvements of the automated validation procedure for ADPs to ensure that as many mistakes as possible can be found in structures prior to publishing. The presented method is based on the Hirshfeld test (Hirshfeld, 1976, Rosenfield *et al.*, 1978) which is the *de facto* standard for ADP validation.

6.1. Methods

6.1.1. Hirshfeld Test

The basis for this work is the Hirshfeld test (Hirshfeld, 1976). The Hirshfeld test checks if a pair of bonded atoms has ADPs that are in agreement with fundamental physical properties of atomic vibrations. It does so by computing the displacement amplitudes of atoms in bond direction to their respective neighbor atom. If both atoms have equal atomic masses and the vibrational motion is harmonic in nature, bonded atoms should have the same displacement amplitude in bond direction. The Hirshfeld test value ΔH_{ij} for the bonded atom pair i and j is computed as

$$\Delta H_{ij} = |H_{ij} - H_{ji}| \tag{6.1}$$

with

$$H_{ij} = U_i^T \cdot \overline{v}_{ij} \cdot U_i, \tag{6.2}$$

$$H_{ji} = U_j^T \cdot \overline{v}_{ij} \cdot U_j. \tag{6.3}$$

$\overline{v}_{ij}$ is the normalized vector pointing from atom i to atom j and U_i is the ADP of atom j. In an ideal bonding environment ΔH_{ij} should be zero for atoms with identical atomic mass.

In reality however, atoms do not necessarily have identical atomic masses. The approximation works well for many organic molecular frameworks that consist mainly of carbon, nitrogen and oxygen atoms, but as soon as hydrogen atom ADPs are analyzed or metal atoms are involved, the Hirshfeld test becomes unreliable.

A second limitation of the test is that it is only reliable for atoms involved in at least 3 bonds. Also, the atom and its bonding partners must not be co-planar. The reason behind this is that the Hirshfeld test for one bond only checks whether the ADP is reasonable in the direction of the bond. For an anisotropically refined ADP however, the vibrational description consists of three independent components. If only one of them (or a linear combination of the three components) is checked, the overall displacement can still be unphysical. The test can hence only be conclusive if three linear independent Hirshfeld tests are performed for an atom. This requires obviously at least three bonds to test, which must not be co-planar, because otherwise the bond vectors would be linearly dependent.

6.1.2. Mass-Adjusted Hirshfeld Test

The Hirshfeld test neglects differences in the atomic mass of bonded atoms, rendering it unreliable in those situations. This section presents a novel method to correct for atomic mass related inccuracies in the test. The method scales ADPs based on the corresponding atomic mass.

An ADP is considered scaled if the part of the ADP that is caused by vibrations of the atomic framework[1] itself is multiplied by its atomic mass. The part of the ADP which is caused by rigid body movements of the atomic framework must be equal for two atoms bonded to each other, and therefore must not be scaled. Equations 6.4 and 6.5 show how to obtain the scaled ADP $U_i^\prime$ from the measured ADP U_i^m, the atomic mass m_i and

[1] The overall vibration of an atom is componsed of lattice vibrations – the movement of the asymmetric unit relative to other asymmetric units – and framework vibrations – the motion of an atom relative to its bonding partners.

the part of the ADP caused by framework vibrations U_i^{int}.

$$U_i' = U_i^m \cdot \left(1 - \frac{f_i}{m_i}\right) + U_i^m \cdot f_i \tag{6.4}$$

$$f_i = m_i \cdot \frac{U_i^{int}}{U_i^m} \tag{6.5}$$

U_i^{int} is not explicitly part of the structure model but can be approximated using the following assumptions:

- If a structure is an ideal rigid body where all atoms have the same mass, the average Hirshfeld Test value ($H_{ij} - H_{ji}$) is zero.

- If a structure is an ideal rigid body, but atoms do not have the same mass, the only differences in Hirshfeld Test values must be due to the different masses involved. If that is the case, the correct values of f_i must be those that minimize the average Hirshfeld Test value.

- In conclusion: standard optimization techniques can be used to find f_i and thereby the values of U_i^{int}.

The scaling factor f_i can be determined as follows: The scaled Hirshfeld test value $\Delta H'_{ij}$ of two bonded atoms should be zero.

$$\Delta H'_{ij} = 0 = H'_{ij} - H'_{ji} \tag{6.6}$$

where H'_{xy} is the amplitude of the ADP of atom x in direction of the bond to atom y for a scaled ADP.

The expression for H'_{ij} can be derived directly from equation 6.4.

$$H'_{ij} = H_{ij}^m \cdot \left(1 - \frac{f_i}{m_i}\right) + H_{ij}^m \cdot f_i \tag{6.7}$$

For every bond an atom is involved in, one equation according to 6.6 can be formulated. Each contains two unknowns (f_i and f_j) leading to the following minimization criterion

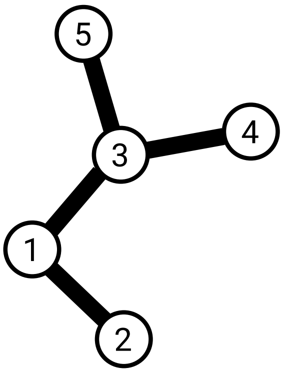


Figure 6.1.: Schematic of an arbitrary molecule used as an example to illustrate the meaning of the presented equations.

where n is the number of atoms:[2]

$$min\left(\sum_{i=0}^{i=n}\sum_{j=0}^{j\neq n}\left(\Delta H'\right)^2\right) \tag{6.8}$$

Since atoms bonded to each other must have similar values of U_i^{ext}, this relationship is used to restrain the values of f_i by using equations 6.9 to 6.12 for all atoms i and j that are bonded. These restraints also work around the problem that terminal atoms only have one bond from which f_i can be derived.

$$U_i^{int} = U_i^m - U_i^{ext} \tag{6.9}$$

$$\frac{f_i}{m_i} = \frac{U_i^{int}}{U_i^m} \tag{6.10}$$

$$U_i^{ext} = U_i^{m_i} - \frac{f_i \cdot U_i^m}{m_i} \tag{6.11}$$

$$U_i^{ext} - U_j^{ext} = 0 \tag{6.12}$$

Equation 6.13 shows the least squares equations for the molecule shown in figure 6.1.

[2] It was tested whether formulating an equivalent expression with lowered weights for 1–3 distances is useful. The test showed no significant improvement. This is probably due to the fact that scaling atom 1 to atom 2 and atom 2 to atom 3 implicitly scales atom 1 to atom 3 with a lower weight.

$$\begin{bmatrix} |H_{21}-H_{12}| \\ |H_{31}-H_{13}| \\ |H_{12}-H_{21}| \\ |H_{13}-H_{31}| \\ |H_{43}-H_{34}| \\ |H_{53}-H_{35}| \\ |H_{34}-H_{43}| \\ |H_{35}-H_{53}| \\ |H_{21}-H_{12}| \\ |H_{31}-H_{13}| \\ |H_{12}-H_{21}| \\ |H_{13}-H_{31}| \\ |H_{43}-H_{34}| \\ |H_{53}-H_{35}| \\ |H_{34}-H_{43}| \\ |H_{35}-H_{53}| \end{bmatrix} = \begin{pmatrix} \frac{H_{12}}{m_1}+H_{12} & -\frac{H_{21}}{m_2}-H_{21} & 0 & 0 & 0 \\ \frac{H_{12}}{m_1}+H_{12} & 0 & -\frac{H_{31}}{m_3}-H_{31} & 0 & 0 \\ -\frac{H_{12}}{m_1}-H_{12} & \frac{H_{21}}{m_2}+H_{21} & 0 & 0 & 0 \\ -\frac{H_{13}}{m_1}-H_{13} & 0 & \frac{H_{31}}{m_3}+H_{31} & 0 & 0 \\ 0 & 0 & \frac{H_{34}}{m_3}+H_{34} & -\frac{H_{43}}{m_4}-H_{43} & 0 \\ 0 & 0 & \frac{H_{35}}{m_3}+H_{35} & 0 & -\frac{H_{53}}{m_5}-H_{53} \\ 0 & 0 & -\frac{H_{34}}{m_3}-H_{34} & \frac{H_{43}}{m_4}+H_{43} & 0 \\ 0 & 0 & -\frac{H_{35}}{m_3}-H_{35} & 0 & \frac{H_{53}}{m_5}+H_{53} \\ -H_{12}m_1 & H_{21}m_2 & 0 & 0 & 0 \\ -H_{13}m_1 & 0 & H_{31}m_3 & 0 & 0 \\ H_{12}m_1 & -H_{21}m_2 & 0 & 0 & 0 \\ H_{13}m_1 & 0 & -H_{31}m_3 & 0 & 0 \\ 0 & 0 & -H_{34}m_3 & H_{43}m_4 & 0 \\ 0 & 0 & -H_{35}m_3 & 0 & H_{53}m_5 \\ 0 & 0 & H_{34}m_3 & -H_{43}m_4 & 0 \\ 0 & 0 & H_{35}m_3 & 0 & -H_{53}m_5 \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix}^T \tag{6.13}$$

Expression 6.13 contains redundant information. In practice, only half of the lines are required, because scaling atom 1 to atom 2 is equivalent to scaling atom 2 to atom 1. With all values of f_i known, all ADPs can be scaled to correct for their different masses. The value of H_{ij} depends on the twelve anisotropic displacement parameters of atoms i and j as well as the norm of the bond vector connecting atom i and atom j, which in return depends on the atomic coordinates of both atoms. In conclusion, nine data points are available in the least-squares fit for each parameter that is optimized. In practice it proved useful to down-weight the U_i^{ext} similarity restraint by multiplying all corresponding matrix rows by an empirically determined factor of 0.15. This is useful because the equations involving the similarity are dealing with numerically larger differences that are about to be minimized. This constant can however be adjusted to enforce a more rigid molecule by choosing a larger weighting factor. A smaller factor implies a more flexible molecule since the equations that enforce similarity of bonded atom's U_i^{ext} get less influence on the overall scaling.

Figure 6.2 shows the effect of ADP scaling. The model is refined against Neutron diffraction data to provide reliable values for the ADPs of hydrogen atoms. Note that hydrogen atoms are not affected by the scaling, because their atomic mass is the refer-

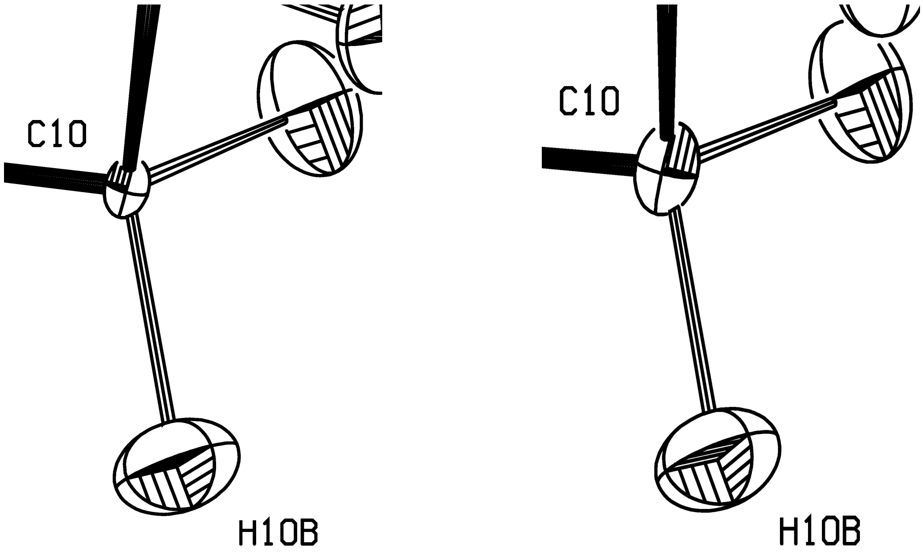


Figure 6.2.: Visualization of the scaling effect at the example of a carbon hydrogen bond. Left: before scaling. Right: after scaling.

ence value to which all other atoms are scaled. Carbon atom *C10* is is 6 times heavier than hydrogen. Therefore the scaling procedure increases the size of the corresponding ADP. Before applying the scaling procedure, the Hirshfeld test value for this bond would have shown a big difference indicating a wrongly determined thermal displacement behavior. After scaling, the Hirshfeld test indicates that the ADPs involved in the carbon-hydrogen bond are in fact perfectly fine.

6.1.3. 3D Hirshfeld Test

The second limitation of the Hirshfeld test that is addressed here is the inaccuracy arising when less than three linearly independent bond vectors are available for an atom. This section introduces two modifications to the Hirshfeld test procedure that have their own strengths and weaknesses. Which of the approaches is most suitable depends on how the testing procedure is implemented, and what the goals of the investigation are.

Distorted Projection

The Hirshfeld test involves the computation of an ellipsoid's size – defined by three orthogonal principle axes – in the direction of an arbitrary normalized vector (see equation 4.1). This is equivalent to projecting each of the principle axes onto the normalized vec-

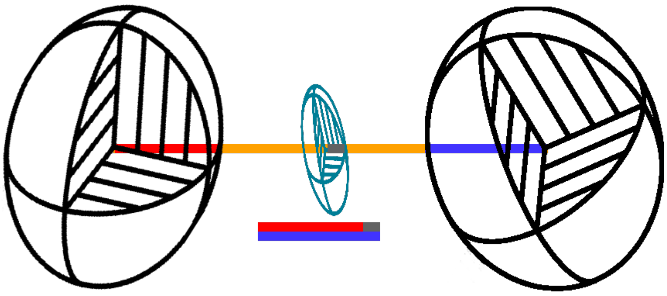


Figure 6.3.: Visualization of a difference ellipsoid.

tor $\overline{v}_{ij}$ and then summing the projected vectors. H_{ij} is then equivalent to the norm of that vector sum. The projection $\overline{x}_p$ of a vector $\overline{x}$ onto the normalized vector $\overline{v}$ can be computed with

$$\overline{x}_p = cos\left(\phi\right) \cdot |\overline{x}| \cdot \overline{v} \tag{6.14}$$

with

$$cos\left(\phi\right) = \frac{\overline{x}}{|\overline{x}|} \cdot \overline{v}. \tag{6.15}$$

ϕ is the angle between $\overline{x}$ and $\overline{v}$. A distorted projection is proposed that substitutes expression 6.14 with

$$\overline{x}_p = \sqrt{cos\left(\phi\right)} \cdot |\overline{x}| \cdot \overline{v}. \tag{6.16}$$

$$H'_{ij} = \left|\overline{x}_p + \overline{y}_p + \overline{z}_p\right| \tag{6.17}$$

This has the effect that the projection sum H'_{ij} – the sum of each principle component's projection onto the bond vector – gets larger the less well aligned the principle axes are to the bond vector. In this context, *aligned* means that one of the principle axes is co-linear to the bond vector. While the traditional Hirshfeld test simply computes the expansion of the ellipsoid in bond direction, the *distorted projection* method includes a penalty function that penalizes deviation from perfect alignment.

Instead of applying this method to each atom's ADP individually, a difference ellipsoid is computed for each bonded atom pair (Figure 6.3). The method proposed above computes one scalar value from the three dimensional displacement representation. Since a method is being discussed to analyze the three dimensional structure of the thermal displacement behavior, differences in three dimensions must be analyzed before the scalar value is computed. This is done by computing the element-wise difference ΔU_{ij} of two bonded atoms' ADPs.

$$\Delta U_{ij} = U_i - U_j \tag{6.18}$$

H'_{ij} can then be computed for each bonded atom pair based on the principle components of the difference ellipsoid ΔU_{ij}. If the atom pair obeys the rigid bond approximation, the difference ellipsoid should have an expansion of zero in the direction of the bond vector, while the eigenvectors perpendicular to the bond vector can have arbitrary lengths. The value thus becomes a direct indication for how well two bonded atoms' ADPs are aligned.

This method works reasonably well for all bonding situations, because it penalizes deviation from the simple assumption that relative displacements of bonded atoms should only occur perpendicular to the connecting bond vector.[3]

An edge case where the method does not work well is when two atoms' ADPs are not well aligned, but are mirror images from each other, with the mirror plane being perpendicular to the bond vector. In this case the ΔU_{ij} would be perfectly aligned with the bond vector even though U_i and U_j are clearly not physically reasonable. In light of the current limitation of the Hirshfeld test this shortcoming can be considered of minor importance.

RIGU Based Testing

The second testing procedure introduced here is based on the *RIGU* restraint available in the *SHELXL* program. (Sheldrick, 2015*b*) The *RIGU* restraint works by rotating an atom's ADP in a way that the Z-axis is aligned with the bond vector to one of its neighboring atoms. This is done once for each bond, yielding one bond-aligned ADP U^k_{ij} for each bond k an atom is part of. Subsequently, U^k_{23} and U^k_{13} are restrained to be zero.[4] The two matrix elements represent the *tilt* of the ellipsoid out of the plane perpendicular to the bond vector, thereby enforcing a displacement model that consists mainly of motion perpendicular to the bond. If a restrained atom is part of a planar local environment, the restraints perpendicular to the bond but within the plane cancel each other, resulting in displacements perpendicular to the plane. An atom in an environment similar to a sp^3 hybridized carbon atom results in all restraints trying to cancel each other out which should yield a mostly spherical ellipsoid where the axis orientation becomes arbitrary.

Similar to the *RIGU* restraint the proposed testing procedure first transfers each ADP in a bond-aligned coordinate system where the z-axis is parallel to the bond vector. Mathematically, this can be done with the procedure described in section 4.1.2. If the the normalized bond vector is used on the left-hand side of the $\otimes$ operator in equation

[3] This assumes that atomic mass differences are taken into account via an appropriate scaling method.

[4] *SHELXL*'s *RIGU* implementation also includes an ellipsoid expansion restraint in addition to the orientation restraint. However, this is not used for the proposed testing procedure directly.

4.14 and the vector $(0, 0, 1)$ is used on the right-hand size, the required rotation matrix is obtained. After the rotation matrix elements U_{23} and U_{13} are extracted and stored in a list for further processing. When this is done for all bonds of a given atom, the arithmetic mean $|r|$ of the list of matrix elements is computed. Additionally, the ADP's elipticity l is computed as the ratio of the ellipsoids longest principle axis divided by the length of its shortest axis.[5] l is used to judge whether an ADP is effectively spherical, which implies that the orientation – encoded in $|r|$ – becomes meaningless. The bond enhanced evaluation factor (BEEF), quality indicator for atom i, can be computed as

$$\text{BEEF} = |r| \cdot (l - 1)\,. \tag{6.19}$$

The term $(l - 1)$ ensures that the BEEF becomes zero for perfectly spherical ellipsoids. In conclusion, a small BEEF can either mean that the displacement ellipsoid of an atom is well aligned to its bond geometry or that the ellipsoid is almost spherical – implying that all displacement directions are equivalent.

This testing procedure only analyzes the orientation of displacement ellipsoids on a per atom basis. This means that – in contrast to the bond centered Hirshfeld test yielding one parameter for each bond – the BEEF procedure yields one parameter for each atom. This also means that the BEEF should always be used in conjunction with the Hirshfeld test to analyze the displacement amplitudes in addition to the displacement directions.

6.2. Results

The proposed modifications to the Hirshfeld test were tested on a set of structures from the literature. The selected models had been refined against Neutron diffraction data. These models have the advantage of including an anisotropic parametrization of atomic displacements of the hydrogen atoms. Since the main advantage of the proposed modifications is the ability to take atomic mass differences into account, the large mass differences between hydrogen atoms and their bonding partners make them ideal test cases. Table 6.1 lists an overview over the selected structural models. Figure 6.8 shows the improvement of the average Hirshfeld test value upon application of the scaling procedure.

[5] The term *elipticity* is used due to its similarity to the elipticity of an ellipses. This should not be confused with the elipticity in the context of topological analysis (Bader, 1990).

Designation	CSD Code	Resolution	Figure	Reference
IRO	208347	0.58 Å	6.4	(Ho *et al.*, 2003)
HYP	977817	0.65 Å	6.5	(Lübben *et al.*, 2014)
GLU	624378	0.55 Å	6.6	(Smrčok *et al.*, 2006)
ANI	166521	0.59 Å	6.7	(Cole *et al.*, 2001)

Table 6.1.: Selected structure models for investigating the Hirshfeld test scaling method.

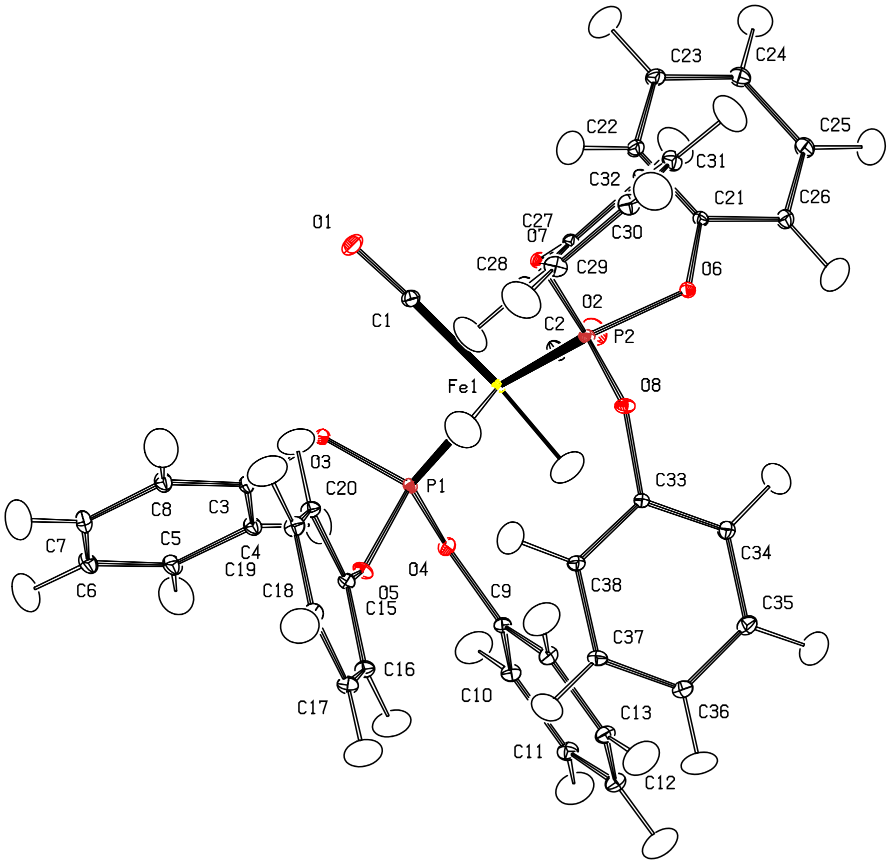


Figure 6.4.: Ortep plot of of structure IRO with atomic numbering scheme.

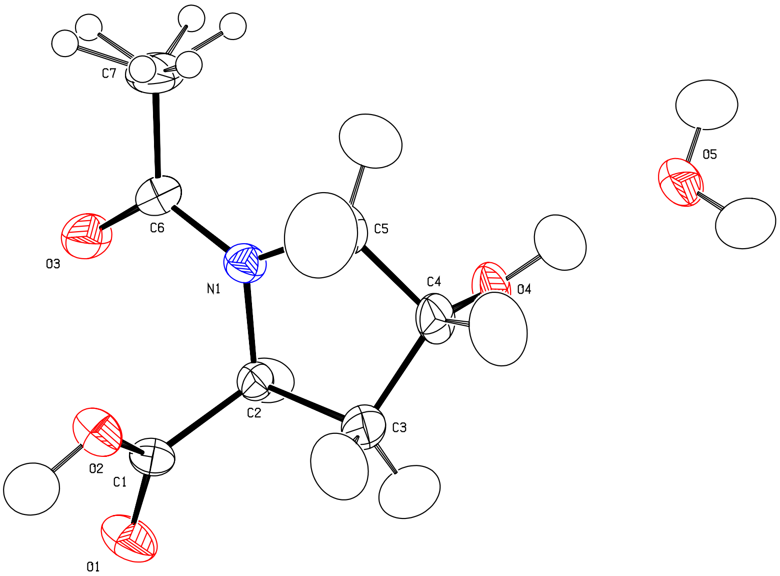


Figure 6.5.: Ortep plot of of structure HYP with atomic numbering scheme.

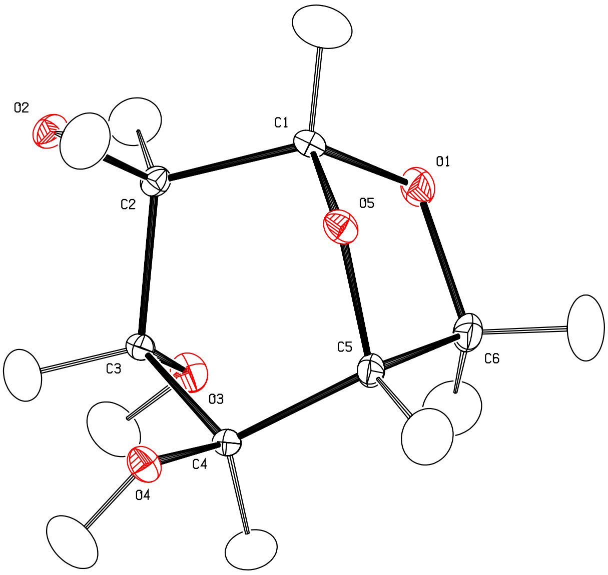


Figure 6.6.: Ortep plot of of structure GLU with atomic numbering scheme.

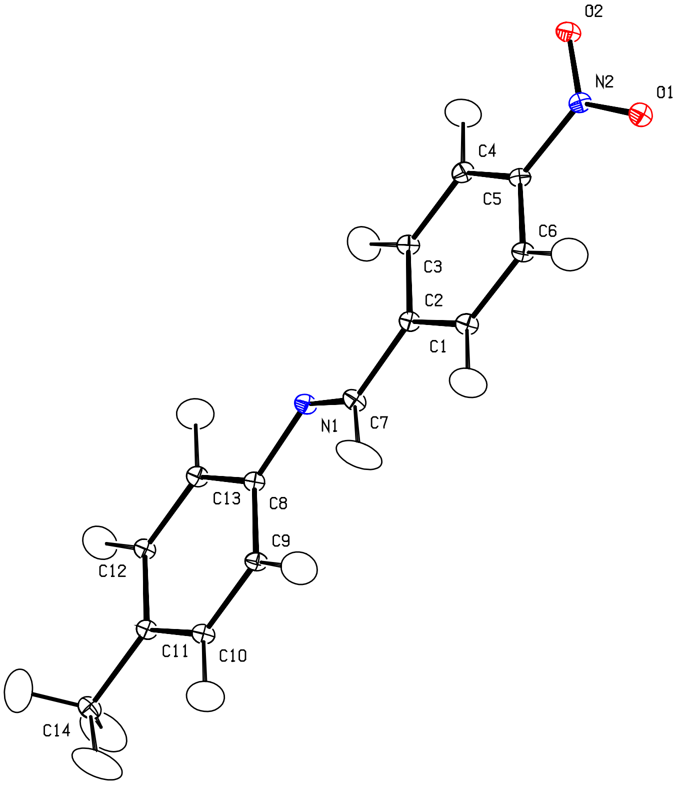


Figure 6.7.: Ortep plot of of structure ANI with atomic numbering scheme.

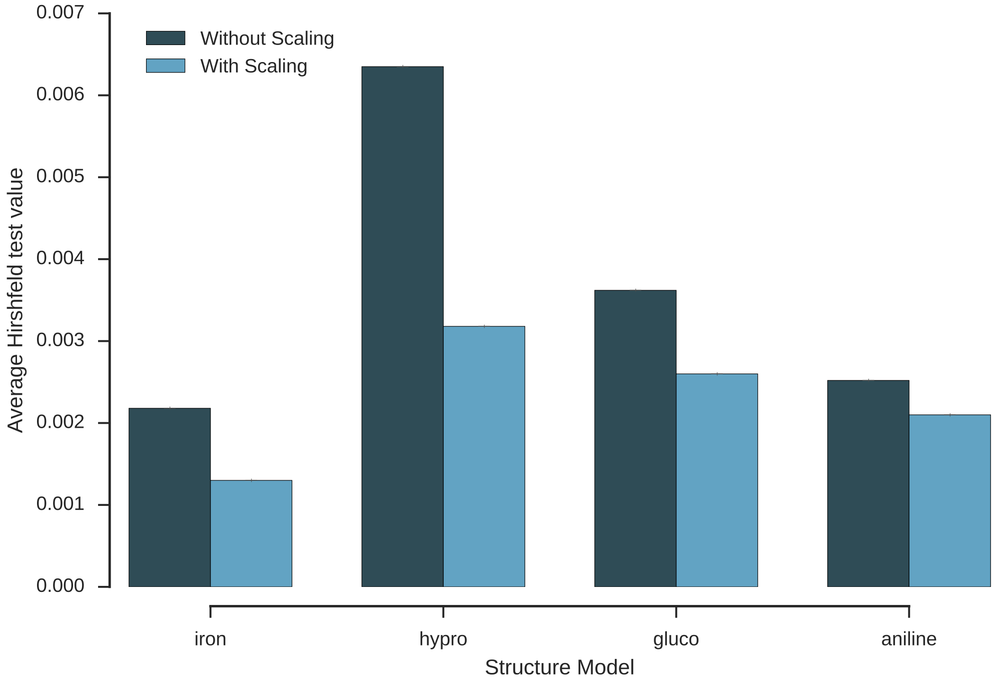


Figure 6.8.: Improvement of the average Hirshfeld test value upon applying the proposed scaling model.

6.2.1. Mass-Adjusted Hirshfeld Test

The minimization of the average Hirshfeld test value is not useful in itself but only if it preserves errors in the model while false positive errors are removed. In order to test that quality of the scaling procedure outlined in section 6.1.2 three atom pairs were analyzed in each structural model. The pairs are the bonds that give rise to the largest Hirshfeld test values before scaling and after scaling. The atom pair corresponding to the largest Hirshfeld test value is also inspected visually: figures 6.9 to 6.13 show the atom pairs for the four test structures. The figures show sections of the molecule. Atoms irrelevant for interpreting the results are omitted for clarity.

Analysis of IRO

The structural model of IRO (Figure 6.9) before scaling indicates that either atom $Fe1$ or atom $H1$ have erroneous ADPs. However, visual inspection shows no indication of an

	Unscaled		Scaled	
Number	Pair	Value	Pair	Value
1	Fe1–H1	0.0086	C5–H5	0.0089
2	Fe1–H2	0.0070	C8–H8	0.0061
3	C31–H31	0.0061	C7–H7	0.0060

Table 6.2.: Hirshfeld test values of the most likely erroneous ADPs before and after scaling for structure model IRO.

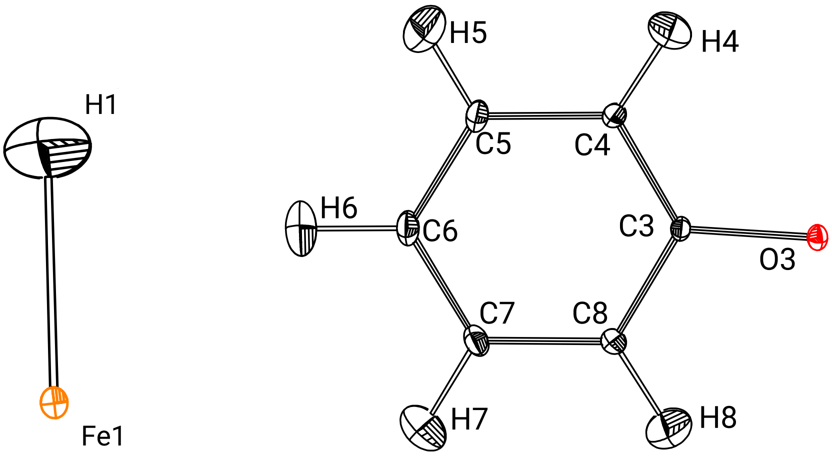


Figure 6.9.: Most likely erroneous ADPs of structure IRO. Left: before scaling. Right: after scaling. Potential errors are discussed in sub-section *Analysis of IRO*.

error. The reason for this false positive is most likely the very large mass difference of the involved nuclei. After scaling, the $Fe1$–$H1$ atom pair is no longer in the list of most likely erroneous ADPs. Instead, three pairs that are part of an aromatic six membered ring give rise to the highest Hirshfeld test values. And indeed, the ADPs of atoms $C5$, $C8$, $C7$ and their corresponding hydrogen atoms seem to be misaligned upon visual inspection. In this case the scaling procedure proves to be a significant improvement to Hirshfeld test.

Analysis of HYP

Before scaling structure HYP (Figure 6.10), the Hirshfeld test indicates an error in the ADPs of either atom $C3$ or $H4$. Visual inspection reveals that the ADPs of both atoms are not perfectly aligned but they do seem to be plausible when compared to the ADPs of neighboring atoms. After the scaling procedure, the most pronounced error is indicated for either atom $C2$ or atom $C3$. Visual inspection reveals that the ADP of atom $C2$

	Unscaled		Scaled	
Number	Pair	Value	Pair	Value
1	C3–H4	0.017	C2–C3	0.010
2	O2–H1	0.016	O5–H12	0.007
3	C5–H7	0.015	O2–H1	0.007

Table 6.3.: Hirshfeld test values of the most likely erroneous ADPs before and after scaling for structure model HYP.

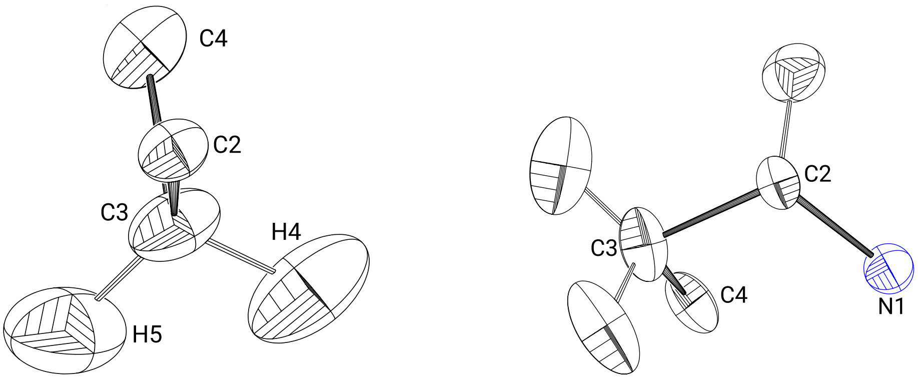


Figure 6.10.: Most likely erroneous ADPs of structure HYP. Left: before scaling. Right: after scaling.

is significantly smaller than expected considering the ADPs of its immediate surrounding. It is unlikely that the ADP of a carbon atom bonded to another carbon atom and a nitrogen atom is smaller that the ADPs of its bonding partners.

The atom pair $O2$–$H1$ is in the list of worst offenders both before and after scaling. Therefore the atom pair was also visually inspected to check if the testing procedure yields reasonable results. Figure 6.11 shows the relevant atoms and shows that the hydrogen atom is not well aligned to the oxygen atom. This indicates that the scaling procedure does not obscure errors in the structure model by absorbing them into its scaling parameters.

Analysis of GLU

The case of structure model GLU (Figure 6.12) shows a potentially misaligned ADP for atom $H6B$ for both the scaled and the unscaled procedure. More significant is however the atom pair yielding the highest Hirshfeld test value for the unscaled structure: atom pair $O3$–$H3$. Visual inspection clearly shows a very large ADP for atom $O3$ that is not justified when compared to the neighboring atom $C3$. This is a potential error that was

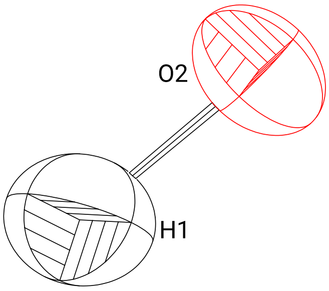


Figure 6.11.: ADPs of atom $O2$ and atom $H1$ of structure HYP.

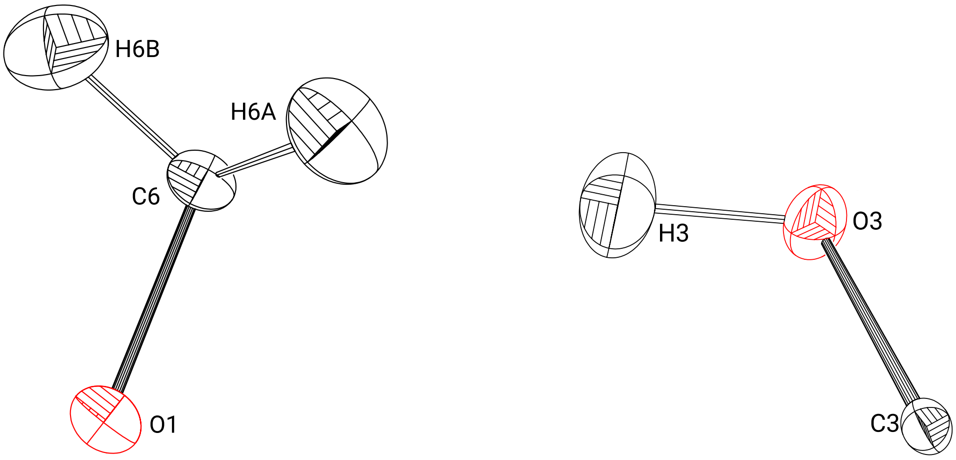


Figure 6.12.: Most likely erroneous ADPs of structure GLU. Left: before scaling. Right: after scaling.

	Unscaled		Scaled	
Number	Pair	Value	Pair	Value
1	C6–H6B	0.0107	O3–H3	0.0113
2	C1–H1	0.0087	O4–H4	0.0065
3	C5–H5	0.0086	C6–H6B	0.0064

Table 6.4.: Hirshfeld test values of the most likely erroneous ADPs before and after scaling for structure model GLU.

	Unscaled		Scaled	
Number	Pair	Value	Pair	Value
1	C14–H14A	0.0085	C14–H14C	0.0121
2	C9–H9	0.0084	C14–H14B	0.0091
3	C13–H13	0.0073	C7–H7	0.0059

Table 6.5.: Hirshfeld test values of the most likely erroneous ADPs before and after scaling for structure model ANI.

missed by the test in the absence of scaling.

Analysis of ANI

The case of ANI (Figure 6.13) strongly suggests a un-modeled rotational disorder of the $C14$ methyl group. Both the scaled and the unscaled test agree in that regard. The difference is that the unscaled test suggests that the $C14$–$H14A$ bond is problematic and the bonds to the other two hydrogen atoms are fine, while the scaled test suggests the opposite. This can be explained with the shape of the carbon atom's ADP. It is smaller in the direction of the $C14$–$H14A$ bond which leads to a larger Hirshfeld test value for the test against the large $H14A$ ADP. Since the carbon atom ADP is bigger in the direction of the other bonds, the Hirshfeld test computes a smaller difference in ADP size. The scaling procedure enlarges the carbon atom's ADP yielding a different result. Overall, both procedures indicated a case of unmodeled disorder which is the desired result.

Summary

Analyzing the test results as a whole one can conclude that the scaling procedure yields satisfactory results. However, it should be noted that none of the tested structures had any significant model errors. The test shows the most striking discrepancies in the model on a scale relative to the rest of the molecule. While this should not be

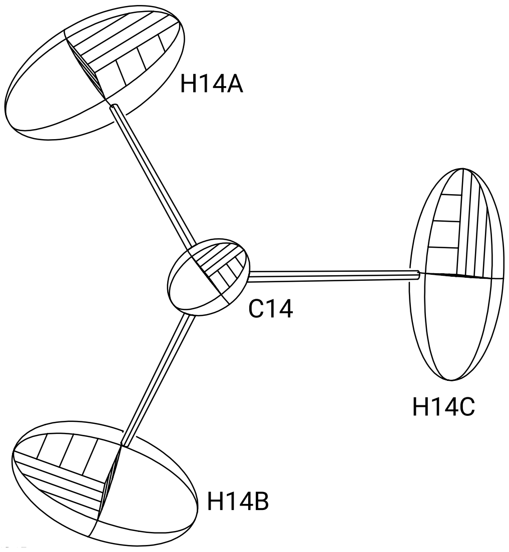


Figure 6.13.: Most likely erroneous ADPs of structure ANI.

Number	Atom Pair	Value
1	C6–C7	0.0036
2	C5–H5	0.0036
3	C7–H7	0.0027

Table 6.6.: Highest H'_{ij} values for structure model IRO.

considered a real *field test*, it is reasonable to assume that it is more challenging to find minor discrepancies in an overall good model than to find major discrepancies in a less well modeled structure.

It would be desirable to repeat the testing procedure with less well modelled structures, but structures optimized against Neutron diffraction data are rare and are usually carefully checked for erroneous ADPs prior to publication. Therefore an appropriate *worst-case* data set could not be found.

6.2.2. 3D Hirshfeld Test

The modifications to improve the analysis of ADPs in three dimensions where tested by applying both proposed methods to the structure models *IRO* and *HYP*.

Number	Atom Pair	Value
1	C4–C5	0.0129
2	O5–H12	0.0120
3	C3–C4	0.0078

Table 6.7.: Highest H'_{ij} values for structure model HYP.

Number	Atom	Value
1	H7	0.003039
2	H37	0.002625
3	C7	0.002072

Table 6.8.: Highest BEEF values for structure model IRO.

Distorted Projection

The distored projection test for structure IRO (Table 6.6) indicates similar potential errors in the structural model as the standard Hirshfeld test does. Figure 6.9 shows all atoms involved in the three most likely erroneous bonds. It is difficult to judge wether the displacement ellipsoids are reasonable or not. However, the test values indicated only minor discrepancies between the structural model on the expected values. Therefore these results are not unexpected. The test remains inconclusive in this case which could simply indicate that the structural model is perfectly fine.

The distored projection test for structure HYP (Table 6.7) indicates that either atom C5 or atom C4 are potentially un-physical. Visual inspection reveals that their displacement ellipsoids do not agree well with the rigid body approximation. The primary displacment directions of both atoms are not reasonable. The standard Hirshfeld test on the other hand provides a test value of 0.000005 Å^2, thereby indicating no potential problem at all. This is due to the fact that the expansion of both ellipsoids in bond direction is almost perfectly equal, which can be the case even for un-physical ellipsoids. Here, the distorted projection test revealed a potential error in the model that would remain hidded from the standard Hirshfeld test.

A possible explanation for the unually oriented displacement ellipsoids could be ring puckering motion. However, no signs for such motion could be found when inspecting other atoms in the ring.

Number	Atom	Value
1	H4	0.041312
2	H8	0.010688
3	C4	0.010185

Table 6.9.: Highest BEEF values for structure model HYP.

RIGU-Based Testing

Table 6.8 lists the most likely errors in structure IRO based on the BEEF. In addition to atoms $C7$ and $H7$, which also show some of the highest values during the mass-scaled Hirshfeld test, the test highlights atom $H37$ to be potentially problematic. Figure 6.15 shows the relevant part of the structure model. The ADP of atom $H37$ shows some misalignment. The shortest principal axis is clearly not well aligned with the bond vector. The atoms $C7$ and $H7$ are displayed in figure 6.4 and show similar characteristics. Visual inspection of the rest of the structure indicates that the most significant ADP errors are found by the testing algorithm. The RIGU-based testing procedure produces valuable results in this case.

BEEF values for model *HYP* are listed in table 6.9 and are significantly higher than the values computed for model *IRO*. Figure 6.16 shows the atoms in question. It is clear that the tabulated atoms' ADPs are not aligned very well and that the ADP of the tertiary carbon atom $C4$ seems to be randomly oriented. However, there is no clear indication that the atoms in the list of most disagreeable ADPs are significantly worse than other ADPs in the structural model. This indicates that even though the algorithm is able to detect unusual displacement ellipsoids, it is not yet suitable for evaluating structure models automatically. Presumably, the reason for this is connected to rigid body vibration. Rigid body motion of bonded atom pairs can lead to ADPs that are not aligned with the bond vector but are still perfectly reasonable. To address this issue, the difference ellipsoids ΔU_{ij} could be analyzed instead of analyzing U_i and U_j separately, in a similar manner to how the distorted projection method is implemented. However, this would obscure potential errors when mirror symmetry is present as discussed in section 6.1.3.

Conclusion

Both proposed methods for analyzing ADPs in three spatial dimensions appear to work well for some of the test structures, but fail to provide useful information for others. This is not surprising since both have known deficiencies that can plausibly explain why the

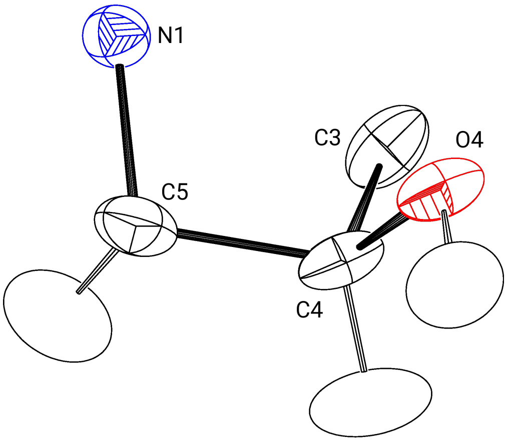


Figure 6.14.: ADPs of structure model HYP that are most likely erroneous based on the distorted projection method.

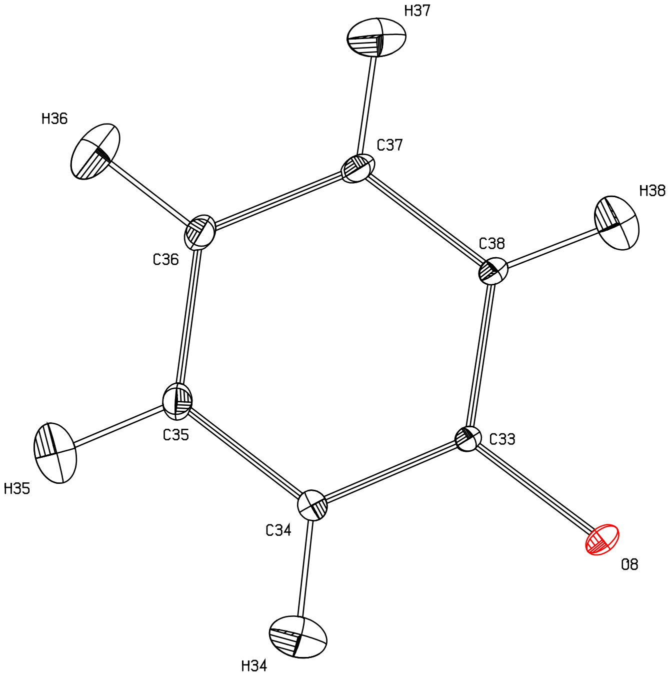


Figure 6.15.: ADPs of structure model IRO that are most likely erroneous based on the *BEEF*.

tests failed to produce valuable results. If a highly detailed analysis of displacement ellipsoids is required, it can be useful to use both tests in conjunction to reveal potential problems in the structure model. However, a fully automated testing procedure can not be provided at this point. In conclusion, a solution combining the advantages of both proposed methods is needed. Appropriate solutions are still being investigated.

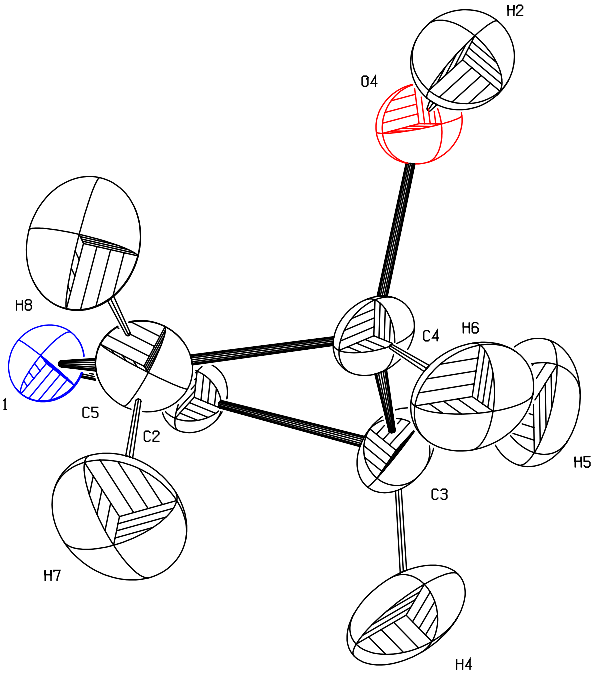


Figure 6.16.: ADPs of structure model HYP that are most likely erroneous based on the *BEEF*.

7. Cross Validation for Small Molecule Structures

The X-ray diffraction experiment does not give the scientist access to the phase information of the diffracted beam. This problem makes the analytical determination of crystal structures impossible. Instead the crystallographer must rely on careful refinement of a structure model against the collected data trying to minimize some form of cost function that evaluates the agreement between model and data. As long as the system is severely overdetermined – meaning several times as many data points were collected than there are parameters to be optimized – this procedure yields reliable results (Kleywegt and Brunger, 1996, Kleywegt and Jones, 1995).

This procedure works reasonably well for small molecules. However, bigger structural models require more parameters to be fitted against the collected data which is usually less precise and less accurate the bigger the studied structure becomes. This leads to the challenge that many structural models are defined by more parameters than there are data points the parameters can be refined against. This results in an under-determined optimization problem. The most common way of working around that challenge is the introduction of side conditions that restrain some parameters to known values basically acting as additional data points. Another necessity is to simplify the structural model, resulting in less parameters to be determined. For example a typical small molecule structure model includes freely refined anisotropic ADPs and atomic positions for most atoms. Structural models of proteins on the other hand use a more simplified rigid-body model for atomic displacement and heavily restrained inter-atomic distances on the amino acid level. Alternatively, single parameters can be removed from the optimization procedure by setting them to fixed values – commonly known as constraints in the field of crystallography. Still, the question remains how many parameters are acceptable and how many side conditions should be introduced to avoid overfitting. The well established solution to this in the world of protein crystallography is the cross validation implementation called R_{free} (Brunger, 1992).

R_{free} works by omitting a randomly selected set of data points from the refinement protocol and validating the refinement results against those omitted reflections. If the model parametrization is chosen appropriately, the agreement between the model and the omitted data should be similar to the agreement between the model and the data it was refined against. If it is not, the data is overfitted and a less flexible parametrization should be chosen. This method works well but has the downside that not all data can be utilized for model optimization, limiting its application to problems where a lot of data points are available and the overall model quality is not significantly affected by the omission of data points. Unfortunately, this usually means that the application of R_{free} is limited to protein X-ray crystallography.

Routine small molecule crystallography does not deal with this problem. A data to parameter ratio for a typical small molecule data set is usually greater than ten. This reduces the risk of overfitting drastically to a point where it can be ignored for most applications. However, not all data points are equivalent. The low scattering amplitude contribution of certain features in the electron density of a molecule can make the overall data to parameter ratio basically meaningless. For example the deformation density – the difference between the IAM density and the measured density – is relatively flat when compared to the density modeled by the IAM. This means that above a certain scattering angle there is basically no contribution of those features to the diffracted intensity meaning that reflections of higher resolution do not contain information about those features.[1] This also implies that even if the overall data parameter to ratio might be well above 10 (indicating a well overdetermined optimization problem) certain parameters might still overfit the data because only a fraction of the data points actually contains information relevant to these parameters. Another case – and the reason why this section is included in this thesis – is the parametrization of hydrogen atoms. Similar to bonding density, hydrogen atoms contribute very little density to the overall charge density distribution, limiting its scattering contribution to reflections of a resolution below 0.8 Å. This means that even high resolution data including reflections with a resolution better than 0.5 Å does include the same information about the hydrogen atoms as the same data set limited to a resolution of 0.8 Å. In conclusion, even if a scattering model appropriate for modeling and refining hydrogen atom parameters is chosen, and very high resolution data is available, the crystallographer needs to be very careful when refining those parameters and should regularly check if some parameters are fitted to errors in the data instead of actual features.

This section discusses an alternative implementation of cross-validation that is not

[1] High resolution reflections are still necessary to appropriately model these density features. However, these reflections don't provide information about the features directly.

limited by the amount of experimental data available and works equally well for all crystallographic structure optimization problems. The discussed protocol was developed in collaboration with Tim Grüne (Lübben and Gruene, 2015) and is an implementation of a protocol that was suggested by Brunger (1992).

7.1. Methods

The goal of $R_{complete}$ is to provide a structure quality indicator for crystal structure models that is free of overfitting and affected by model bias as little as possible. $R_{complete}$ is closely related to *k-fold cross validation* (Efron and Gong, 1983, Kohavi *et al.*, 1995).

7.1.1. Cross-Validation

The challenge of estimating the validity of a mathematical model is not specific to crystallography. Every time a model is optimized against a set of non-ideal data the model is affected by the errors in the data. Cross-validation is used to check how trustworthy the optimized model is. Several flavors of cross-validation exist. The basic concept in most of the available techniques is similar:

1. Split data D into training set T and validation set V.
2. Optimize model M against training set.
3. Compute statistics by comparing M to V.[2]

The critical decision to be made is: how to split D into T and V. The most commonly used implementation of cross-validation in crystallography – R_{free} – randomly selects n reflections to be omitted from model building. If p reflections were collected, the model is optimized against $p - n$ reflections and statistics are computed against n reflections. This means that T and V are defined at the beginning of model building and are not changed throughout the whole procedure. This provides the critical advantage of having a validation set V that is not affected by bias and only needing to optimize the model against one training set T – the latter being particular important because computational power is limited and more exhaustive validation techniques were not feasible 20 years ago. The main disadvantage of this method is that the computed statistics' stability

[2] Some validation techniques e.g. Jackknife compute the statistics against T by averaging values obtained from different ways to split D into T and V. However, the techniques used in crystallography share the characteristic that quality indicators are computed against V.

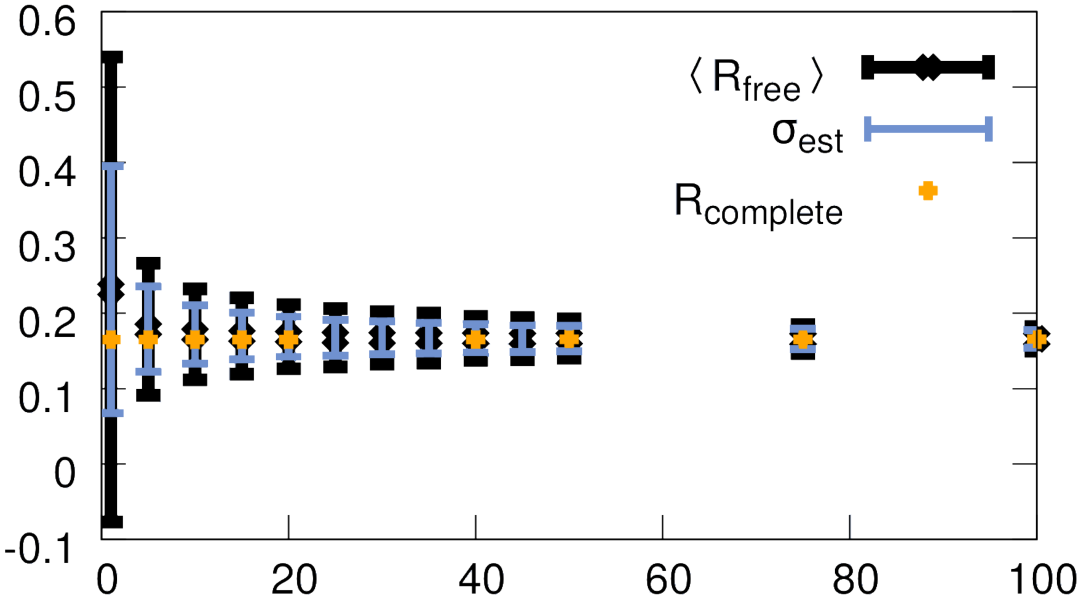


Figure 7.1.: Dependence of $R_{complete}$ on the validation set size n. $R_{complete}$ was computed for all possible sets V_i with $0 < i < k = p/n$. Data set Hormaomycin was used for this study. Figure from Lübben and Gruene (2015).

depends on the size of V. If V is small, the variance of statistics computed against V is large. The standard deviation of R_{free} is estimated (Tickle *et al.*, 1998) to be

$$\sigma_{R_{free}} = \frac{R_1}{\sqrt{n}}. \tag{7.1}$$

However, if V is large, the model optimization procedure itself might become unstable because the data to parameter ratio is negatively affected by a large set V. In practice, $n \geq 500$ is strongly recommended for the R_{free} technique. In cases where this is not possible due to model instability, R_{free} becomes unreliable. Figure 7.1 shows the dependence of the variance on the set size n. The figure also shows that $\sigma_{R_{free}}$ is underestimated for small set sizes.

$R_{complete}$, originally proposed by Brunger (1992), implements a k-fold cross validation-like technique for crystal structure model validation. Instead of splitting D into two static sets T and V, D is split into k set pairs V_i and T_i. V_i contains n data points with $n = p/k$ and T_i consists of all remaining data. The sets are selected in a way that each data point is part of exactly one V_i set and part of $(k-1)$ T_i sets. The model validation procedure now requires k model optimizations, each against one set of T_i. Statistics are then computed against the sum of all V_i. The statistic relevant in this context $R_{complete}$ can than be computed as

$$R_{complete} = \frac{\sum\limits_{i}\sum\limits_{h \in V_i} ||F_{obs}(h)| - |F_{calc}(h)||}{\sum\limits_{i}\sum\limits_{h \in V_i} |F_{obs}(h)|} \tag{7.2}$$

$$R_1 = \frac{\sum\limits_{h} ||F_{obs}(h)| - |F_{calc}(h)||}{\sum\limits_{h} |F_{obs}(h)|} \tag{7.3}$$

$$R_{free} = \frac{\sum\limits_{h \in T} ||F_{obs}(h)| - |F_{calc}(h)||}{\sum\limits_{h \in T} |F_{obs}(h)|} \tag{7.4}$$

Equations 7.3 and 7.4 show the definition of $R1$ and R_{free} respectively for reference. $R_{complete}$ is computed against all data and was shown to be independent of the set size n/k (Lübben and Gruene, 2015) (see figure 7.1). This implies that, short of numerical inaccuracies and potential instability of the model optimization, all possible values of k yield the same value of $R_{complete}$. This also implies that structure optimization problems that require as much data as possible included in the optimization procedure can use as many as $p-1$ reflections, thereby loosing virtually no stability compared to omitting no data at all. The downside is significantly increased computational cost compared to R_{free}. This can be mitigated to some degree by choosing the set size to be as large as possible without affecting the optimization stability negatively. However, even the largest reasonable set size of $n/k = n/(0.5n)$ requires two complete optimization steps while R_{free} will only require one step.

7.1.2. Bias

The bias of a quality indicator like R_1 is the difference between R_1 and the value R_1 would have if the data was free of errors. R_{free} provides a way to quantify bias by omitting the validation set V from the optimization protocol. Since the model is never actually optimized to reproduce V it is not affected by bias and the difference between R_1 and R_{free} is a measure for bias. $R_{complete}$ does not have that advantage. $R_{complete}$ is designed to potentially utilize all data for the optimization procedure which implies that the model was refined against V and bias must be reduced actively during validation.

Two possibilities to achieve this were investigated:

1. Model relaxation by random perturbation of parameters before each of the k refinement steps. The random displacement removes the potential impact the omitted reflections had on the model in previous refinement steps (Joosten *et al.*, 2014, Mihelic *et al.*, 2011, Pražnikar and Turk, 2014).
2. Model relaxation by enforcing a large number of refinement cycles during each step. If model bias is affecting the refinement, refinement until quasi convergence for each of the k refinement steps ensures that the bias is removed prior to quality indicator computation.

7.1.3. Implementation Details

Several ways to implement $R_{complete}$ are possible. The study discussed here is based on an implementation utilizing the least-squares refinement program SHELXL without modification. However, the successful pilot study using that implementation led to the implementation of the $R_{complete}$ computation protocol into SHELXL, thereby reducing the application complexity significantly. The results presented in this chapter are obtained via the prototype implementation which will be discussed here in detail. An outline of the SHELXL implementation will be given to demonstrate how equivalent results can be obtained in a simplified and streamlined manner.

Computing $R_{complete}$ requires the following prerequisites:

- A structural model to be validated in SHELXL's *.res* format.
- A fully[3] merged reflection data file in *HKL* format.
- The number of data points omitted in each cross-validation step n/k.

Using n/k and the reflection data file as input, the utility program *crossflaghkl*, authored by Tim Grüne, will randomly generate k reflection files. Each file contains all data with a procedurally generated set of n/k reflections being flagged as *free* reflections by setting the appropriate flag. Each data point will be flagged as *free* exactly once across all generated files.

Next, the structural model has to be refined against each of the data files. Both the full-matrix least-squares and the conjugate gradient least-squares algorithms are

[3] *fully merged* implies that symmetry equivalents and Friedel pairs are merged. The simplified implementation works without this prerequisite because the merging is performed within SHELXL.

appropriate for this purpose as long as the *nrf* flag is set to -1 in both cases. It is recommended to use the conjugate gradient least-squares algorithm to reduce the computation time. The main disadvantage of the algorithm – the fact that the algorithm does not yield error estimates for optimized parameters – is not relevant in this context. Since the parameters are not refined against all data, the parameters obtained via full-matrix least-squares are not meaningful either. Instead, other means of obtaining error estimations are discussed later in this chapter.

After all refinement processes are finished, the relevant data can be extracted from the output files. The relevant information is the sum of the differences between the observed data points that were flagged as *free* and the corresponding calculated intensities. This represents the inner sum of the numerator in equation 7.2. The second relevant information is the sum of the flagged, observed data points which corresponds to the inner sum of the denominator in the same equation. When this is done for all output files, corresponding information from all files can be added – representing the outer sums in equation 7.2 – and divided by each other to yield $R_{complete}$.

SHELXL Implementation of $R_{complete}$

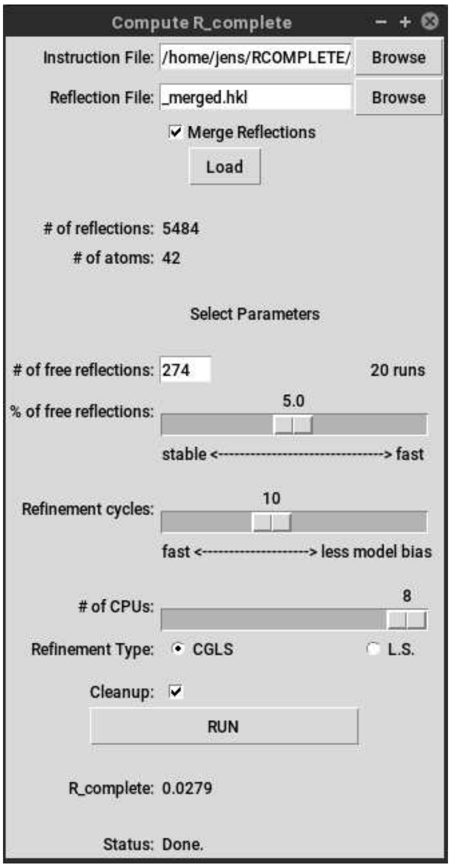
The SHELXL implementation now available simplifies the process significantly by incorporating the generation of the required reflection data files into the refinement program. This implies that no third party programs are needed and no (mostly redundant) data files need to be created. Instead only two numerical parameters are needed in addition to the standard SHELXL input:

k the number of discrete refinement runs. k is put into SHELXL via the command line flag $-g[k]$.

m indication which of the k refinement runs should be executed. m is put into via the command line flag $-m[m]$.

In conclusion, for a k-fold cross-validation of a structure model SHELXL can be started k times, each time with the parameter m incremented by 1 starting at $m = 1$. The procedure to harvest the relevant data is equivalent to the prototype implementation discussed before.

For convenience, a small utility program was coded providing quick access to this functionality via a simple graphical user interface and a command line wrapper that reduces the user input to one single command. Figure 7.2 shows an image of the



Compute R_complete
Browse
Browse
Load
Select Parameters
stable <------------------------------> fast
fast <-------------------> less model bias
RUN

Instruction File: /home/jens/RCOMPLETE/

Reflection File: _merged.hkl

☑ Merge Reflections

# of reflections: 5484

# of atoms: 42

# of free reflections: 274 20 runs

% of free reflections: 5.0

Refinement cycles: 10

# of CPUs: 8

Refinement Type: ● CGLS ○ L.S.

Cleanup: ☑

R_complete: 0.0279

Status: Done.

Figure 7.2.: Depiction of the $R_{complete}$ graphical user interface developed to streamline the computation of $R_{complete}$ with SHELXL

graphical user interface. The application is freely available.[4]

7.1.4. Parameter Error Estimation

In addition to the main application of $R_{complete}$ as a tool to detect overfitting, the $R_{complete}$ procedure provides a way to estimate uncertainties for all optimized parameters. Estimating the uncertainties of optimized parameters is essential to properly analyze the obtained model. Unfortunately, not all optimization techniques give access to estimated uncertainties. The large number of optimized parameters often makes it impossible to apply a full-matrix least-squares algorithm, which provides means to compute error estimates. The required computational resources of said algorithm scale quadratically with the system size which is not feasible for larger structures. The conjugate gradient

[4] https://github.com/JLuebben/R_complete

method makes more efficient use of computer resources, facilitating the optimization of much larger systems, but does not give access to estimated uncertainties. Maximum likelihood based algorithms also do not permit error estimation.

$R_{complete}$ provides a solution for that problem. Every time $R_{complete}$ is computed each of the k refinement steps provides a parameter file listing all model parameters optimized against a subset of the data. After all refinement steps are carried out, the parameter listing files can then be analyzed to find the variance of optimized parameters amongst all files, thereby providing an uncertainty estimate for every parameter. This method of error estimation provides the additional benefit of being virtually independent of the optimization algorithm applied and does therefore allow to directly compare error estimates obtained via different optimization techniques.

7.1.5. Free Density Maps

Similar to how $R_{complete}$ facilitates the estimation of parameter uncertainties, the method can be used to generate density maps that are less affected by overfitting. It is necessary to use the phases computed from the structure model to generate density maps. This implies that errors in the model affect the density map that is subsequently used in additional model building steps. This results in density maps that reproduce the model used to generate them even if the crystal does not contain electron density at the part in question. This is particularly problematic if combined with *human bias*, i.e. bias introduced by the researcher by trying to find certain features in the density. It is therefore desirable to reduce the error in those density maps as much as possible in order to build the least biased model possible.

The protocol to generate these maps is similar to the parameter error estimation protocol: every refinement step produces a *FCF* file listing the observed intensity, its error, the calculated intensity and the phases calculated for the reflection for every reflection flagged as *free* during a given refinement step. Since the phases are computed from a model that was not actually optimized against the reflections listed in the file, they are not subject to overfitting and have reduced bias compared to the default F_{obs} map. The way $R_{complete}$ is computed ensures that each reflection is flagged *free* for exactly one refinement step. This means that the concatenation of all *FCF* files yields a file containing all reflections with reduced bias. That file can then be used to generate model density maps or difference density maps via Fourier synthesis.

7.1.6. Application in Small Molecule Crystallography

As discussed earlier, the main advantage of $R_{complete}$ compared to R_{free} is the possibility to apply the protocol to systems were comparably few data points are available or virtually no data can be excluded from the optimization procedure. This facilitates the application of $R_{complete}$ in small molecule crystallography were smaller unit cells and more flexible model parameterization require to use as much data as possible in the optimization. The applicability in small molecule crystallography was investigated by analyzing the effect of freely refined hydrogen atom ADPs on $R_{complete}$.

Supramolecular Structures

The treatment of hydrogen atoms in routine crystal structure determination is well standardized and common validation practices e.g. CHECKCIF are efficient tools for detecting potential errors in the structural model. The concept of cross-validation is still useful for small molecule XRD studies that use non-standard models or refine against less complete data. This is often the case when analyzing supramolecular structures.

The studied supramolecular structures share some characteristics that are typical for this type of structure. They have a comparably high solvent content that is usually disordered making modeling of parts of the structure challenging and in some cases even impossible. The solvent content, together with flexible parts of the organic ligands result in badly crystallizing compounds. The small size of the crystals often requires the data to be collected at synchrotrons where the beam intensity is high enough to get measurable diffracted intensities. Limited goniometer flexibility at synchrotron beam lines leads to less complete data and the fragility of the sample crystals often results in low redundancy as well. Moreover, avoiding radiation damage and reaching the best possible resolution need to be taken into account in the measurement strategy simultaneously.

The overall poor data quality and deficiencies in the applied structural model to describe disordered or almost flat solvent regions requires a different method for analyzing the influence of different models on the overfitting of a structural model. The previously studied small molecule structure models manifest overfitting in subtle ways, making it necessary to analyze the change of $R_{complete}$ relative to R_1. The supramolecular structures investigated here make it possible to look at the changes of $R_{complete}$ directly. Overfitting can manifest so drastically for these structures that additional parameters result in larger $R_{complete}$ values, which is a much clearer indicator for overfitting than the changes of $R_{complete} - R_1$ that were studied before. However, overfitting can occur even if $R_{complete}$ is dropping, which made the analysis of relative $R_{complete}$ changes necessary in the previously discussed study.

7.2. Results

Two published XRD data sets were used to test the methods described in this chapter. Table 7.1 lists an overview of the data sets used.

7.2.1. Removal of Bias

It was investigated whether it is necessary to randomly perturb model parameters prior to a validation step in order to reduce bias introduced by optimizing the model against the validation set V. This was done by creating multiple sets of perturbed models with different perturbation amplitudes and comparing the convergence behavior to a non-perturbed model. The procedure was tested with data set Insulin.

The model perturbation was implemented via the SHELXL *WIGL* command changing positional and vibrational parameters by a random amount within a defined range representing the perturbation amplitude.

The convergence behavior was monitored by computing $R_{complete}$ for the reference model and each of the perturbation levels after each refinement cycle. The attributes of interest are the value of $R_{complete}$ when quasi convergence is reached, and the number of refinement cycles required to reach quasi convergence.

Figure 7.3 indicates that the final value of $R_{complete}$ is not affected by the perturbation of parameters. All lines converge to the same value within the accuracy of the method. It is therefore reasonable to conclude that it is not necessary to perturb parameters prior to the validation procedure to remove bias from the optimized model. A number of refinement cycles between 20 and 50 should be more than sufficient to obtain a practically unbiased quality indicator. However, it is worth considering to apply a small perturbation nevertheless to increase the convergence rate or have a starting $R_{complete}$ value higher than the bias free value. The figure suggests that a small perturbation

Name	Space Group	Resolution	No. of Atoms	No. of Data
Insulin	$I2_13$	1.1 Å	436	$32,598$
Hormaomycin	$P2_1$	1.02 Å	215	$7,800$

Table 7.1.: Overview of the data used for testing $R_{complete}$

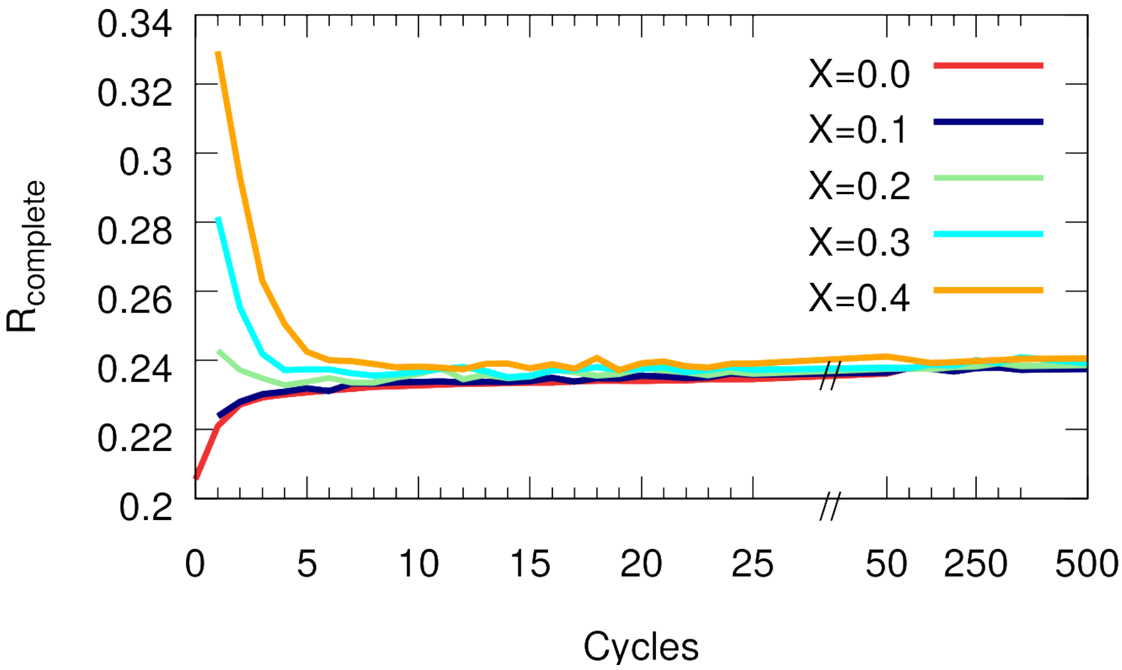


Figure 7.3.: Plot of the $R_{complete}$ value against the number of refinement cycles. Each line corresponds to a different random perturbation amplitude in Å. Figure from Lübben and Gruene (2015).

Type	ID	Resi	#	ESD	ESD: X Y Z	Average Position
O	847	HOH	23	0.6	0.5 0.2 0.5	28.1 50.4 28.4
O	850	HOH	26	0.3	0.4 0.2 0.2	32.7 36.2 31.6
O	852	HOH	28	0.2	0.4 0.1 0.2	31.3 36.9 29.4
O	845	HOH	21	0.2	0.2 0.2 0.2	25.3 51.8 33.1
HZ3	816	LYS	2029	0.2	0.2 0.2 0.2	17.1 54.3 24.3
HZ2	815	LYS	2029	0.2	0.1 0.2 0.2	18.0 53.2 24.1
HZ1	814	LYS	2029	0.2	0.1 0.2 0.1	18.0 53.9 25.3
HE1	811	LYS	2029	0.2	0.1 0.2 0.1	15.9 53.1 25.8
HE2	812	LYS	2029	0.2	0.1 0.2 0.2	16.9 52.0 25.6

Table 7.2.: Example ESD analysis of Insulin test structure.

amplitude of 0.2 Å can be beneficial to increase the convergence rate although there is no reason to assume that the optimal amplitude will be independent of the system that is investigated. It is reasonable to assume that this optimal amplitude depends on the resolution and accuracy of the data the model is optimized against, although no experiments to verify that were carried out.

7.2.2. Parameter Error Estimation

Data set Insulin was analyzed to estimate uncertainties for optimized positional parameters. Table 7.2 shows the result of that study listing the 9 atoms with the highest variance in their positional parameters. The table shows that the atoms with the highest variance in their positional parameters correspond to a number of solvent molecules. Solvent molecule positions are often less well determined than the positions of the protein backbone which is consistent with the obtained results. The other five atoms listed in the table are part of a lysine residue that is part of a flexible part of the protein. The estimated high uncertainty is plausible in this case too.

This study demonstrates that the procedure can be employed to quickly identify less well defined parts of the structural model. This can be useful to determine whether solvent atoms should be removed at a certain position or if parts of the structure are disordered in a way that requires modeling of multiple conformations. This method also provides a way to compare the quality of similar structural models optimized against different data sets or to compare similar but not identical models. The R_1 value is not useful in those cases because it only indicates the agreement between model and data and does not allow judgment on how accurate the model actually is.

It should be noted that the uncertainties estimated by this method do not directly

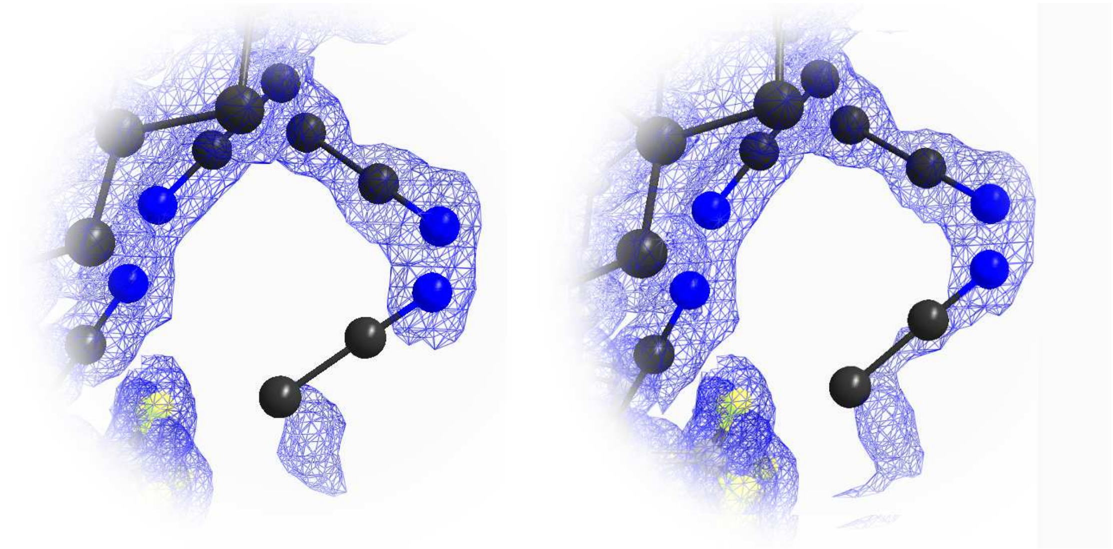


Figure 7.4.: Comparison of a free F_{obs} map (left) and a standard F_{obs} map (right). Both maps are rendered with an iso level of 0.34 e/Å^3.

correspond to uncertainties obtained by full-matrix least-squares optimization against the whole data set.

7.2.3. Free Density Maps

A density map with reduced bias was computed for the structure *MX01*. The structure contains an unknown amount of C_2NH_3 (acetonitrile) molecules that are highly disordered. Commonly applied model building techniques do not provide any measure to estimate whether a certain conformation is modeled appropriately. A free density map is a promising tool to help in that regard. Figure 7.4 shows a particular strongly disordered part of the solvent region. The model density F_{obs} and the difference density $F_{calc} - F_{obs}$ do not provide enough information to make an educated decision where to place solvent molecules. Instead a free F_{obs} map was computed and is displayed at a cut-off level of 0.34 e/Å^3. The figure indicates that one of the solvent molecules might have been placed wrongly.

It should be noted that the differences between the standard F_{obs} map and the free F_{obs} map are very subtle. At different map iso levels the difference between them is hardly visible. While free maps can provide a useful tool for modeling flat density regions, the influence of overfitting bias on density maps is very small in the cases studied.

Name	Resolution	Data/Parameter Ratio	Atoms in Asymmetric Unit	Reference
MBADNP	0.55 Å	41.2	33	(Cole *et al.*, 2002)
Xylitol	0.41 Å	109.3	17	(Madsen *et al.*, 2003)
Maleate	0.45 Å	43.2	35	(Grabowsky *et al.*, 2014)
Squarate	0.45 Å	69.3	40	(Şerb *et al.*, 2014)

Table 7.3.: Overview of small molecule structure investigated.

7.2.4. Application in Small Molecule Crystallography

The applicability of $R_{complete}$ in small molecule crystallography was demonstrated by investigating the commonly used model quality indicator R_1 and the corresponding $R_{complete}$[5] for differently parametrized structural models. 4 differently parametrized models were analyzed.

Isotropic Heavy Atoms All atoms refined with isotropic displacement parameters. A riding atom model is used for hydrogen atoms.

Riding Atom Model Hydrogen atoms are modeled with the riding atom model. All other atoms are modeled with anisotropic ADPs.

Isotropic H-Atoms Hydrogen atom positions are refined freely. Hydrogen atom ADPs are refined isotropically. All other atoms are modeled with anisotropic ADPs.

Anisotropic H-Atoms All atoms are refined with anisotropic ADPs. All atomic positions are refined freely.

The goal of this study was to find the model for hydrogen atoms that yields the least biased model. Bias was quantified here by computing

$$b_{rel} = b - b_0 \tag{7.5}$$

with

$$b = R_{complete} - R_1. \tag{7.6}$$

b_0 is the value of b for the model *Isotropic Heavy Atoms*. The normalization (Equation 7.5) was performed to bring all data sets onto the same scale thus making the plots easier to read.

[5] The size of the validation set V is 10 for all structures investigated resulting in 800–1800 refinement steps for each structure depending on the number of available reflections. The validation was executed on an Intel Xeon X5570 CPU (8 Cores @ 2.93 GHz) and took less than one minute for each structure.

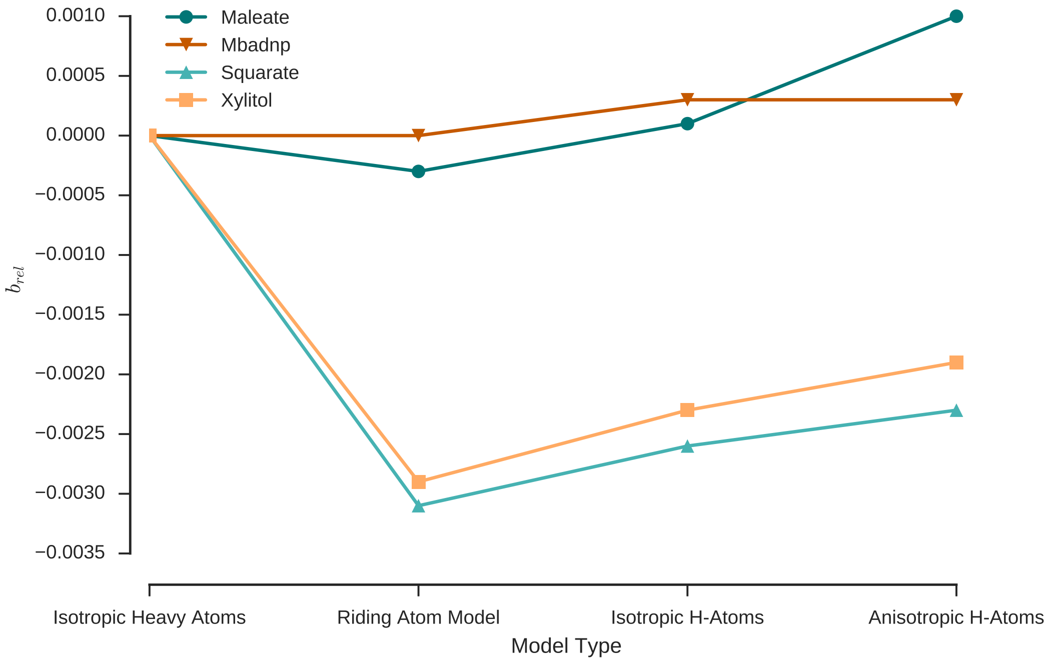


Figure 7.5.: Bias of differently parametrized hydrogen atom models. The least biased model corresponds to the minimum in the plot which is the *Riding Atom Model* in all four cases.

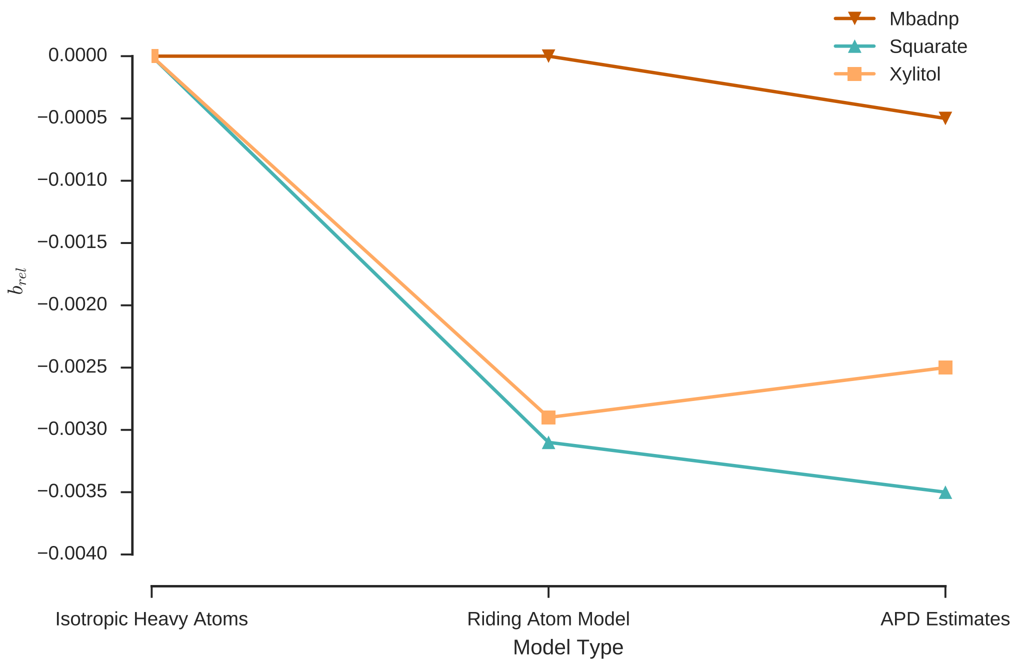


Figure 7.6.: Bias of differently parametrized hydrogen atom models. In contrast to figure 7.5 the ADPs of hydrogen atoms are estimated with the method described in section 4. This modeling technique does not introduce additional parameters to the structural model and is therefore less likely to introduce overfitting.

Figure 7.5 shows that the common practice of modeling hydrogen atoms with the riding atom model is in general appropriate for XRD diffraction studies. Going from a less flexible structural model to a more flexible one – going from left to right in Figure 7.5 – should lower $R_{complete}$ by the same amount as R_1 is lowered. Otherwise the additional parameters overfit the data significantly. The latter situation is the case for all test structures when going from the *Riding Atom Model* to the *Isotropic H-Atoms*.

If a study requires a more flexible model than the *Riding Atom Model*, $R_{complete}$ can be a useful tool to determine the most detailed structural model with the least amount of overfitting. Generally speaking, the minimum in the plot corresponds to the model that fits the data best without overfitting the data.

Name	Resolution	Data/Parameter Ratio	Atoms in Asymmetric Unit	Reference
SL_ADA	1.1 Å	4.4	103	(Löffler *et al.*, 2016)
SL_ACR	1.3 Å	2.8	345	(Löffler *et al.*, 2015)
SL_123	1.1 Å	2.3	227	(Löffler *et al.*, 2016)
MX01	1.1 Å	2.3	227	(Zhu *et al.*, 2015)

Table 7.4.: Overview of the investigated structures.

Figure 7.6 shows that the structural model can be improved with the method described in section 4.[6] $R_{complete}$ indicates that the structural model of *Xylitol* is not improved significantly by estimating hydrogen atom ADPs. This is most likely due to the fact that half of the hydrogen atoms in the model are involved in hydrogen bonding that is not taken into account in the ADP estimation procedure. Since no additional parameters are introduced by the ADP estimation the observed small increase in the relative drop of $R_{complete}$ is probably not an effect of overfitting. Instead, the variation of $R_{complete}$ could be due to limitations of the accuracy of the applied methods or poor accuracy of the diffraction data.

Supramolecular Structures

A set of three supramolecular structures was selected to investigate the application of $R_{complete}$ to this type of structure. Table 7.4 lists the most relevant characteristics of the collected data sets and their corresponding structural models.

4 differently parametrized models were analyzed.

Isotropic ADPs All atoms were refined with isotropic displacement parameters. A riding atom model is used for hydrogen atoms.

ADP-Restraints(RIGU)I All non-hydrogen were atoms refined with anisotropic displacement parameters. All bonded atom pairs were restraint with RIGU restraints. SIMU restraints were applied when necessary. A riding atom model is used for hydrogen atoms.

ADP-Restraints All non-hydrogen were atoms refined with anisotropic displacement parameters. All bonded atom pairs were restraint with DELU restraints. SIMU

[6] The structural model *Maleate* was excluded from this study because the model contains a disordered hydrogen atom that cannot be modeled consistently across all structural models.

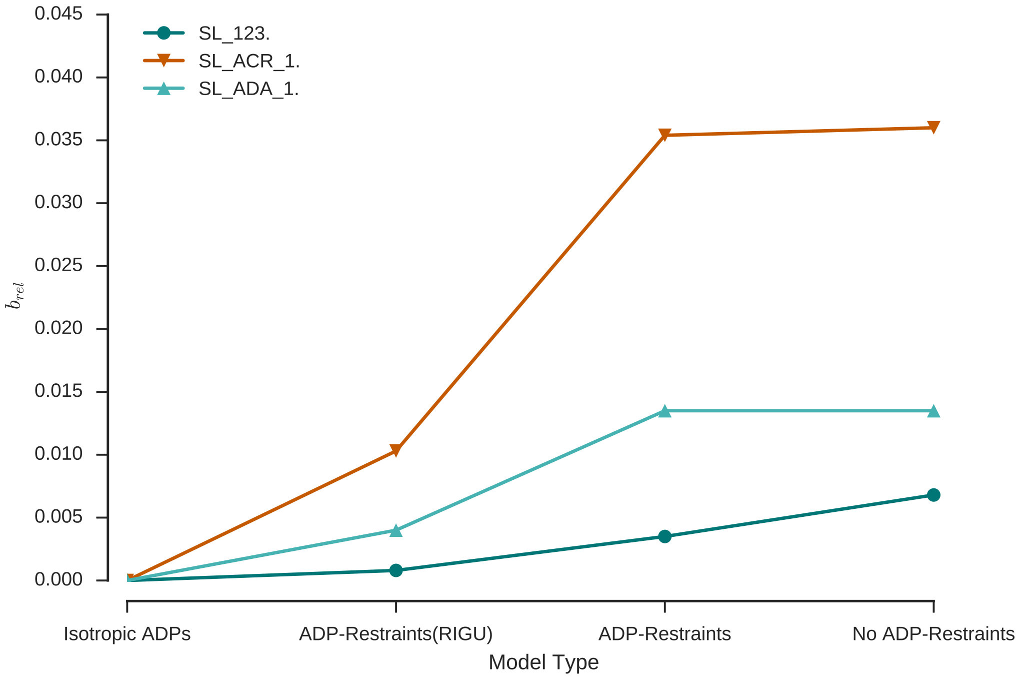


Figure 7.7.: Bias of differently parametrized hydrogen atom models. The least biased model corresponds to the minimum in the plot. In contrast to the plots shown previously in this chapter, plotting b_{rel} does not provide valuable information for these structures.

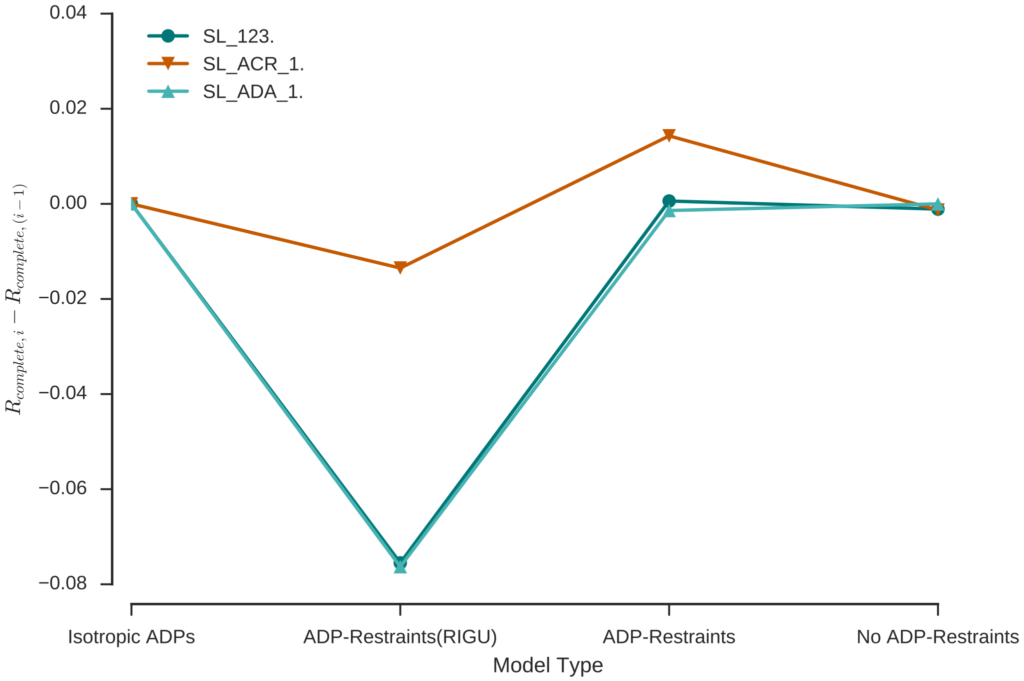


Figure 7.8.: Change of $R_{complete}$ relative to the previous model. A negative value indicates that $R_{complete}$ dropped compared to the previous model. Positive values indicate that $R_{complete}$ increased compared to the previous model which is a clear sign of overfitting.

restraints were applied when necessary. A riding atom model is used for hydrogen atoms.

No ADP-Restraints Al non-hydrogenl atoms were refined with anisotropic ADPs. A riding atom model is used for hydrogen atoms.

Figure 7.8 shows that most appropriate model for the studied structures uses anisotropic ADPs for non-hydrogen atoms and restraints them with appropriate ADP restraints including the advanced rigid bond restraint *RIGU* available in the SHELXL software. Dropping the *RIGU* restraint – even with other rigid bond restraints (DELU) still in place – introduces overfitting to the structural model. Data obtained from particularly bad diffracting crystals e.g. the compound SL_ACR_1, might even require to drop the anisotropic parametrization of atomic displacement entirely in favor of an isotropic description.

This clearly indicates that the validation criteria commonly used for small molecule XRD studies are not appropriate for this kind of structure. It is therefore recommended to apply cross-validation to find the ideal parametrization model for each compound studied.

Part IV.

Conclusion & Outlook

Conclusion & Outlook

The aim of this thesis was to improve common crystal structure modeling techniques and to provide tools to assert an optimized model's validity. This was achieved by combining experimental results with those from theoretical computations and by employing statistical tools for validation purposes.

The first major focus of this thesis – the improved modeling of hydrogen atoms in XRD studies – proved useful to increase model accuracy without introducing additional model parameters to the refinement procedure. Consequently, the application of hydrogen ADP estimation is a valid and recommended procedure independent of the available data accuracy. It was demonstrated that the most commonly used model for hydrogen atoms – the riding atom model – yields inaccurate results at temperatures below 100 Å and thus affects thousands of strurctures deposited in the CCDC although the errors introduced by the constrained model are small. As most XRD data sets are nowadays measured at 100 Å, it was deemed necessary to provide modeling tools that work around that deficiency. An empirical temperature dependent correction factor was published (Madsen and Hoser, 2015) in response to the original publication (Lübben *et al.*, 2015). Studies of bond length accuracy demonstrated that ADP estimation results in more accurate models. The proposed method of estimation was shown to yield results comparable to established estimation techniques and is able to leverage the flexibility of the invariom database. Possible future developments include extending the automation capabilities of the implementation to other structure types like polymers and molecules on special positions, as well as the addition of an anharmonic displacement description for terminal atoms.

The second focus of this thesis is the validation of structural models optimized against experimental data. Inaccuracies in the Hirshfeld test procedure were addressed. Possible solutions were presented and discussed. The proposed modifications to the testing procedure facilitate validation of ADPs of atoms with significantly different atomic masses. Solutions for validating ADPs in special bonding environments were discussed but no satisfactory solution can be provided yet. The presented scaling procedure requires integration into existing validation protocols and a general solution for treating special bonding situations needs to be found.

Further, it was demonstrated that the accuracy of experimental diffraction data severely limits the flexibility of hydrogen atom models possibly leading to overfitting already with positional and isotropic discplacement parameters refined. Although it was shown in the first part that the riding atom model is not appropriate at low temperatures, limited data accuracy does not justify the free optimization of a more appropriate model in general. Conclusively, estimating and subsequently constraining hydrogen atom parameters is the best available solution to this problem. In several test studies, the introduction of estimated hydrogen ADPs in fact reduced the amount of overfitting in the structural model. This is a strong indication for the validity of the estimation procedure introduced. It was also shown that refinement of hydrogen atom parameters is not justified even against very high resolution data.

The presented structure validation technique – $R_{complete}$ – also proved useful in the context of choosing the best parametrization model. This is advantageous in cases were established techniques like *CHECKCIF* are not conclusive due to low data accuracy. This is often a challenge when analyzing XRD data of supra-molecular structures where crystal sensitivity and a poorly crystallizing compounds limit the accuracy of XRD data. It was demonstrated that it can be advisable to use less flexible thermal displacement models for these compounds even for atoms like carbon. $R_{complete}$ can be a valuable tool to select an appropriate model. In addition to its validation capabilities, $R_{complete}$ facilitates estimation of parameter errors and allows density maps with reduced bias to be generated.

Bibliography

Bader, R. F. W. (1990). *Atoms in Molecules: A Quantum Theory*. No. 22 in The International Series of Monographs on Chemistry. Oxford: Clarendon Press, 1st ed.

Becke, A. D. (1988). *Physical Review A*, **38**(6), 3098–3100.

Besl, P. J. and McKay, N. D., (1992). Method for registration of 3-d shapes.

Betteridge, P. W., Carruthers, J. R., Cooper, R. I., Prout, K. and Watkin, D. J. (2003). *J. Appl. Cryst.* **36**, 1487.

Blessing, R. H. (1995). *Acta Cryst. B*, **51**, 816–823.

Blom, R. and Haaland, A. (1985). *J. Mol. Struct.* **128**, 21–27.

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

Brunger, A. T. (1992). **355**, 472–475.

Bürgi, H. B. and Capelli, S. C. (2000). *Acta Cryst. A*, **56**, 403–412.

Burnett, M. N. and Johnson, C. K. (1996). *ORTEP-III, Oak Ridge Thermal Ellipsoid Plot Program for Crystal Structure Illustrations*. Tech. rep. Oak Ridge National Laboratory Report ORNL-6895, Oak Ridge, Tennessee.

Busing, W. R. and Levy, H. A. (1964). *Acta Cryst.* **17**, 142–146.

Campbell, J. W. (1995). *J. Appl. Cryst.* **28**, 228–236.

Capelli, S. C., Bürgi, H. B., Dittrich, B., Grabowsky, S. and Jayatilaka, D. (2014). *IUCrJ.* **1**, 361–379.

Clark, R. C. and Reid, J. S. (1995). *Acta Cryst. A*, **51**, 887–897.

Cole, J. M., Goeta, A. E., Howard, J. A. K. and McIntyre, G. J. (2002). *Acta Cryst. B*, **58**, 690–700.

Cole, J. M., Howard, J. A. K. and McIntyre, G. J. (2001). *Acta Crystallographica Section B*, **57**(3), 410–414.

Cruickshank, D. W. J. (1956*a*). *Acta Cryst.* **9**, 747–753.

Cruickshank, D. W. J. (1956*b*). *Acta Cryst.* **9**, 757.

Dittrich, B. (2009). BAERLAUCH*, A program to prepare input files for QM/MM and other cluster calculations*. Tech. rep. University of Göttingen, Göttingen.

Dittrich, B., Hübschle, C. B., Pröpper, K., Dietrich, F., Stolper, T. and Holstein, J. J. (2013). *Acta Cryst. B*, **69**, 91–104.

Dittrich, B., Sze, E., Holstein, J. J., Hübschle, C. B. and Jayatilaka, D. (2012). *Acta Cryst. A*, **68**, 435–442.

Dittrich, B., Warren, J. and McKinnon, J. J. (2008). *Acta Cryst. B*, **64**, 750–759.

Dominiak, P. M., Espinosa, E. and Ángyán, J. G. (2012). *Intermolecular Interaction Energies from Experimental Charge Density Studies*, pp. 387–433. Dordrecht: Springer Netherlands.

Dunitz, J. D. (1979). *X-ray Analysis and the Structure of Organic Molecules*. London: Cornell University Press, 1st ed.

Dunitz, J. D. and White, D. N. J. (1973). *Acta Cryst. A*, **29**, 93–94.

Efron, B. and Gong, G. (1983). *The American Statistician*, **37**(1), 36–48.

Fischer, R. X. and Tillmanns, E. (1988). *Acta Cryst. C*, **44**, 775–776.

Flaig, R., Koritsanszky, T., Dittrich, B., Wagner, A. and Luger, P. (2002). *J. Am. Chem. Soc.* **124**, 3407–3417.

Flaig, R., Koritsánszky, T., Janczak, J., Krane, H.-G., Morgenroth, W. and Luger, P. (1999). *Angew. Chem. Int. Ed.* **38**(10), 1397–1400.

Frey, M. N., Lehmann, M. S., Koetzle, T. F. and Hamilton, W. C. (1973). *Acta Cryst. B*, **29**, 876–884.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Keith, T., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Adamo, V. B. C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J. and Fox, D. J. (2013). *Gaussian 09, Revision D.01*. Tech. rep. Gaussian, Inc., Pittsburgh PA.

Giacovazzo, C., Monaco, H. L., Viterbo, D., Scordari, F., Gilli, G., Zanotti, G. and Catti, M. (1992). *Fundamentals of Crystallography*. No. 2 in IUCr Texts on Crystallography. Oxford: Oxford University Press, 1st ed.

Grabowsky, S., Woinska, M., Jayatilaka, D., Spackman, M. A., Edwards, A. J., Dominiak, P. M., Wozniak, K., Nishibori, E. and Sugimoto, K. (2014). *Acta Cryst. A*, **70**, 483–498.

Grosse-Kunstleve, R. W. and Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 477–480.

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

Hirshfeld, F. L. (1976). *Acta Cryst. A*, **32**, 239–244.

Hirshfeld, F. L. (1977). *Isr. J. Chem.* **16**, 198–201.

Ho, N. N., Bau, R. and Mason, S. A. (2003). *Journal of Organometallic Chemistry*, **676**(12), 85 – 88.

Hospital, M., Courseille, C. and Leroy, F. (1979). *Biopolymers*, **18**, 1141–1148.

Hummel, W., Hauser, J. and Bürgi, H.-B. (1990). *J. Mol. Graphics*, **8**, 214–218.

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

Jensen, F. (1994). *Introduction to Computational Chemistry*. Hoboken, NJ: Wiley, 2nd ed.

Johnas, S. K. J., Morgenroth, W. and Weckert, E. (2006). *Hasylab Jahresbericht*.

Johnson, C. K. (1969). *Acta Cryst. A*, **25**, 187–194.

Joosten, R. P., Long, F., Murshudov, G. N. and Perrakis, A. (2014). *IUCrJ*, **1**(4), 213–220.

Kabsch, W. (2010). *Acta Cryst. D*, **66**, 125–132.

Kleywegt, G. J. and Brunger, A. T. (1996). *Structure*, **4**(8), 897 – 904.

Kleywegt, G. J. and Jones, T. (1995). *Structure*, **3**(6), 535 – 540.

Koetzle, T. F., Frey, M. N., Lehmann, M. S. and Hamilton, W. C. (1973). *Acta Cryst. B*, **29**, 2571–2575.

Kohavi, R. *et al.* (1995). In *Ijcai*, vol. 14, pp. 1137–1145.

Koritsánszky, T., Flaig, R., Zobel, D., Krane, H.-G., Morgenroth, W. and Luger, P. (1998). *Science*, **279**, 356–358.

Kratzert, D., Leusser, D., Holstein, J. J., Dittrich, B., Abersfelder, K., Scheschkewitz, D. and Stalke, D. (2013). *Angew. Chem.* **125**, 4574–4578.

Löffler, S., Lübben, J., Krause, L., Stalke, D., Dittrich, B. and Clever, G. H. (2015). *Journal of the American Chemical Society*, **137**(3), 1060–1063.

Löffler, S., Lübben, J., Wuttke, A., Mata, R. A., John, M., Dittrich, B. and Clever, G. H. (2016). *Chem. Sci.* **7**, 4676–4684.

Lübben, J., Bourhis, L. J. and Dittrich, B. (2015). *Journal of Applied Crystallography*, **48**(6), 1785–1793.

Lübben, J. and Gruene, T. (2015). *Proceedings of the National Academy of Sciences*, **112**, 29.

Lübben, J., Volkmann, C., Grabowsky, S., Edwards, A., Morgenroth, W., Fabbiani, F. P. A., Sheldrick, G. M. and Dittrich, B. (2014). *Acta Cryst. A*, **70**, 309–316.

Luger, P. (1980). *Modern X-Ray Analysis on Single Crystals*. Heidelberg and Berlin: W. de Gruyter.

Madsen, A. Ø. (2006). *J. Appl. Cryst.* **39**, 757–758.

Madsen, A. Ø., Civalleri, B., Ferrabone, M., Pascale, F. and Erba, A. (2013). *Acta Cryst. A*, **69**, 309–321.

Madsen, A. Ø. and Hoser, A. A. (2014). *Journal of Applied Crystallography*, **47**(6), 2100–2104.

Madsen, A. Ø. and Hoser, A. A. (2015). *Acta Crystallographica Section A*, **71**(2), 169–174.

Madsen, A. Ø. and Larsen, S. (2007). *Angew. Chem. Int. Ed.* **46**, 8609–8613.

Madsen, A. Ø., Mason, S. and Larsen, S. (2003). *Acta Cryst. B*, **59**, 653–663.

Massa, W. (1996). *Kristallstrukturbestimmung*. Teubner Studienbücher. Stuttgart: B. G. Teubner, 2nd ed.

Mebs, S., Messerschmidt, M. and Luger, P. (2006). *Z. Kristallogr.* **221**, 656–664.

Merritt, E. A. (1999). *Acta Cryst. D*, **55**, 1109–1117.

Mihelic, M., Bedrac, L., Renko, M., Besenicar, M. and Turk, D. (2011). *Acta Cryst*, **67**, C480–C481.

Müller, P., Herbst-Irmer, R., Spek, A., Schneider, T. and Sawaya, M. (2006). *Crystal Structure Refinement: A Crystallographer's Guide to SHELXL*. New York: Oxford University Press, 1st ed.

Oxford-Diffraction-Ltd. (2006). *CrysAlis CCD and RED, Version 1.171.31.5*. Tech. rep. Oxford Diffraction, Ltd., Oxford.

Pražnikar, J. and Turk, D. (2014). *Acta Crystallographica Section D: Biological Crystallography*, **70**(12), 3124–3134.

Ramanadham, M., Sikka, S. K. and Chidambaram, R. (1973). *Pranama*, **1**(6), 247–259.

Rez, D., Rez, P. and Grant, I. (1994). *Acta Cryst. A*, **50**, 481–497.

Rosenfield, R. E., Trueblood, K. and Dunitz, J. D. (1978). *Acta Cryst. A*, **34**, 828–829.

Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice and Application to Structural Biology*. New York: Garland Science.

Schomaker, V. and Trueblood, K. N. (1968). *Acta Cryst. B*, **24**, 63–76.

Schomaker, V. and Trueblood, K. N. (1998). *Acta Cryst. B*, **54**, 507–514.

Şerb, M.-D., Kalf, I. and Englert, U. (2014). *CrystEngComm*, **16**(46), 10631–10639.

Sheldrick, G. M. (2008). *Acta Cryst. A*, **64**, 112–122.

Sheldrick, G. M. (2015*a*). *Acta Crystallographica Section A*, **71**(1), 3–8.

Sheldrick, G. M. (2015*b*). *Acta Crystallographica Section C*, **71**(1), 3–8.

Smrčok, Ľ., Sládkovičová, M., Langer, V., Wilson, C. C. and Koóš, M. (2006). *Acta Crystallographica Section B*, **62**(5), 912–918.

Sørensen, H. O., Stewart, R. F., McIntyre, G. J. and Larsen, S. (2003). *Acta Cryst. A*, **59**, 540–550.

Spek, A. L. (2009). *Acta Cryst. D*, **65**, 148–155.

Steward, R. F. (1972). *J. Phys. Chem.* **57**, 1664–1668.

Stewart, R. F. (1969). *J. Chem. Phys.* **51**(10), 4569–4577.

Stewart, R. F. (1976). *Acta Cryst. A*, **32**, 565–574.

Stewart, R. F., Davidson, E. R. and Simpson, W. T. (1965). *J. Chem. Phys.* **42**(9), 3175–3187.

Svensson, M., Humbel, S., Froese, R. D. J., Matsubara, T., Sieber, S. and Morokuma, K. (1996). *J. Phys. Chem.* **100**, 19357–19363.

Tickle, I. J., Laskowski, R. A. and Moss, D. S. (1998). *Acta Crystallographica Section D*, **54**(4), 547–557.

Verbist, J. J., Lehmann, M. S., Koetzle, T. F. and Hamilton, W. C. (1972). *Acta Cryst. B*, **28**, 3006–3013.

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

Wagner, A. and Luger, P. (2001). *J. Mol. Struct.* **595**, 39–46.

Whitten, A. E. and Spackman, M. A. (2006). *Acta Cryst. B*, **62**, 875–888.

Wilkinson, C. and Lehmann, M. S. (1991). *Nuclear Instruments and Methods in Physics Research*, **A310**, 411–415.

Woińska, M., Grabowsky, S., Dominiak, P. M., Woźniak, K. and Jayatilaka, D. (2016). *Science Advances*, **2**(5), e1600192.

Wu, G., Rodrigues, B. L. and Coppens, P. (2002). *J. Appl. Cryst.* **35**, 356–359.

Zhu, R., Lübben, J., Dittrich, B. and Clever, G. H. (2015). *Angewandte Chemie International Edition*, **54**(9), 2796–2800.

Zhurov, V. V., Zhurova, E. A., Stash, A. I. and Pinkerton, A. A. (2011). *Acta Cryst. A*, **67**, 160–173.

Appendices

A. Laue-Script

Almost all data processing performed for obtaining the results presented in this thesis was done with *Laue-Script* - a crystallographic data processing library developed for this purpose.[1] *Laue-Script* facilitates super fast prototyping of crystallographic applications with *Python* and provides automation of many necessary prerequisites like coordinate system transformations, application of symmetry operations, file input and output, user interface generation and much more.

Because the library was specifically developed to perform the studies presented in this thesis and therefore is a significant part of the work performed, a short overview of the library's capabilities is given.

A.1. Library Architecture

Laue-Script consists of several components that can be used separately to access certain functionalities or in conjunction.

Plug-in Manager The plug-in manager is the heart of *Laue-Script*. It is the first object to be created when a *Laue-Script* based application is started and implements the user interface and the plug-in interface that enables cross plug-in communication.

IO-Interface Crystallographic software reads and writes differently formatted files to store and exchange crystallographic data. *Laue-Script* aims to support the most relevant data formats and provides an abstract interface to access crystallographic data independent of the format it is stored in. Currently supported formats are: SHELXL-INS/RES, XD-INP/RES, PDB and CIF.

Types Several crystallographic data types are available that conveniently bundle functionality relevant for working with types like atoms and molecules.

[1] Laue-Script is available at https://github.com/JLuebben/Laue-Script

Function Library Many crystallographic data processing algorithms share a common base of operations that are relevant for many different purposes. *Laue-Script* includes a collection of useful algorithms and functions that can be freely combined to generate more powerful functionality.

Plug-in Skeleton *Laue-Script* includes a template file that can be copied to quickly create new plug-ins without the need to write boilerplate code.[2]

Databases *Laue-Script* can access a compressed version of the entire invariom database and provides convenient access to most of its attributes.

The intended way of writing software based on *Laue-Script* is to encapsulate every feature within its own plug-in. Ideally, every plug-in is a combination of library functions and every program is a collection of plug-ins. The interconnection of plug-ins to a working whole is controlled by the user via a simple chain of command-line arguments that determines when and how plug-ins are to be executed. To further bundle plug-ins into easy to use building blocks, a plug-in can be as simple as a few lines of code that execute different plug-ins with predefined parameters.

The intention behind the design of *Laue-Script* is to create reusable software blocks that prevail over multiple generations of PhD students and can be easily reused for different projects. Encapsulation of features within plug-ins ensures that legacy code can be maintained easily and the common plug-in interface acts as a safeguard: even if a plug-in can no longer be maintained because the author is no longer available for bug fixing or further development, the feature can be re-implemented in a new plug-in that will seamlessly inter-operate with the rest of the code base as long as it uses the same command-line parameters. Strictly enforcing separation of features via the plug-in interface ensures that no unforeseen side effects occur whenever a legacy plug-in is dropped in favor of a new implementation.

Another advantage of the plug-in architecture is that development of new data processing protocols does not require advanced programming skills. If the available plug-ins are designed with flexibility in mind, they can be quickly recombined to serve completely different purposes than they were originally designed for. An example for this are the plug-ins *micro*, *T*, *S*, *W* and *compare*. Depending on the order of execution they can be used to compare invariom database based ADPs to those refined against

[2] A plug-requires certain variables and functions to be defined in order to work properly. Since the functions signatures and variables names are constant for each plug-in, this code is can simply be duplicated to avoid rewriting of identical code.

Neutron diffraction data, compare ONIOM based ADPs to XRD data, write them to files to refine model against the data of another and much more.

In addition to larger software packages like *APD-Toolkit*, *Laue-Script* can also be useful for super rapid prototyping. Since the IO-interlace provides very convenient access to common crystallographic data files (one single line needs to be written to read data from a file with arbitrary format), one can immediately start testing algorithms without the need to deal with boilerplate code like reading and writing of data files, transforming coordinate systems, rotating atoms etc.. The function library is a helpful tool for designing data processing algorithms. While not having been optimized for speed, most functions are helpful in many contexts.

The function library includes algorithms for 3D shape registration, 3D coordinate transformations, topological analysis of graphs, filtering algorithms, global optimization algorithms and similarity search algorithms. In general, algorithms provide interfaces to deal with raw data like arrays and scalar values as well as abstract interfaces to operate directly on crystallographic data types like atoms and molecules.

B. APD-Toolkit

The *APD-Toolkit* is the first application developed on the basis of *Laue-Script* and implements the methods presented in sections chapters 3 and 4. To facilitate the reproduction of all results presented in this thesis, a short overview on the features and usage of the application is provided in this appendix section.

The main concept behind *APD-Toolkit* is flexibility implemented with *Laue-Script*'s plug-in interface. Most features of *APD-Toolkit* are encapsulated within their own plug-in that can seamlessly interface with other plug-ins to construct a tool chain highly specific to a given application case. This makes it easy to adapt a feature developed with a specific application in mind to completely different scenarios that require a similar *step* in their data processing chain.

B.1. Program Structure

The main purpose of *APD-Toolkit* is the generation of ADPs via different methods and to facilitate the transfer of atomic parameters to equivalent chemical environments based on the invariom model. ADPs can be generated from the output of quantum-chemical frequency computations. Both gas phase computations and crystal lattice approximation in the form of ONIOM computations are supported. The standard procedure for computing ADPs based on gas phase computations is discussed first.

B.1.1. Frequency Information from Gas Phase Computations

Gas phase frequency information is part of the invariom database which is a collection of molecules with quantum-chemically optimized geometry data and corresponding frequency information. The database contains thousands of molecules and is several hundred gigabytes big. To keep frequent I/O operations to a minimum, *APD-Toolkit* re-compiles the database into an intermediate database file that caches all relevant information and discards everything not relevant for further processing. This is achieved via Python's native object serialization protocol *pickle*. The resulting file is about $100\ \mathrm{Mb}$ big and is effectively a collection of the invariom database's frequency information.

Based on this intermediate file (database.pkl), *APD-Toolkit* is able to compute ADP representations of atoms' vibrational behavior at a given temperature. Many diffraction experiments are carried out at similar temperatures. Therefore the ADP representations are cached for a given temperature for future application.[1] Whenever *APD-Toolkit* is used again for a structure measured at the same temperature, cached ADPs are used. In addition to ADP information, the cache files include information about the orientation of an atom relative to its immediate chemical environment. This information is crucial for transferring the anisotropic vibrational information to different models.

The cached ADP data encodes the vibrational behavior of atoms in the gas phase. To yield reasonable results, the crystal lattice effects on atomic vibrations must be approximated during the transfer process. This is achieved via an TLS+ARG fit – a method to approximate lattice vibrations and low frequency framework vibrations.[2] Before performing the fit, the cached ADPs are transferred to their corresponding equivalent atoms in the structural model studied (See section 4.1.2 for details). It is assumed that the external vibrations are the difference between the measured[3] ADPs and the ADPs computed from theory (internal ADPs) plus potential errors absorbed by the ADP parameters during refinement. To enforce a physically reasonable model for external vibrations, internal ADPs are subtracted from the measured ADPs and the TLS+ARG fit is performed against the difference ADPs.

At this point *ADP-Toolkit* has stored internal and external ADPs data int their respective coordinate system of each atom in the studied structure. ADP data of atoms with ADPs that cannot be determined via refinement against experiment data (e.g. hydrogen atoms) is extrapolated based on their positional data and the TLS+ARG parametrization.

Estimated ADPs can now be written to a file for all atoms by summing internal and external ADPs.

B.1.2. Frequency Information from ONIOM Computations

ONIOM computations approximate crystal packing effects by placing the asymmetric unit in one or more shells of symmetry equivalent units. This provides additional insight

[1] Computing ADP representations for the whole invariom database based on the database intermediate representations takes about one minute on a modern 8 core CPU.

[2] Both contributions together are denoted *external vibrations* in this context.

[3] ADPs optimized against experimental data are considered *measured* here.

into the vibrational behavior of atoms since inter-molecular forces can be taken into account. On the other hand, results obtained via this method are highly specific to one crystal structure because the crystal packing is unique to that structure. This implies that the procedure described in the previous section must be altered to process ONIOM data.

In the previously described procedure ADP representations are computed from frequency data output and then cached for future application. Caching of intermediate results is not useful here due to the data being specific to exactly one structure.[4] Therefore, a local database file is generated that contains all relevant information for estimating ADPs. The omission of caching is not problematic here because this *micro database* contains the information on one molecule instead of thousands. Re-computation of all ADP representations takes a fraction of a second on a modern computer.[5]

The transfer protocol to match equivalent atoms in both the output of the quantum-chemical computation and the studied structural model is altered to provide the most detailed information possible. The transfer protocol for gas phase computations is based on invariom partitioning, meaning that the chemical environment of an atom is specified in a specific way and all atoms with identical environmental descriptions are deemed equal. For example a hydrogen atom of a phenyl group will be deemed equal to all hydrogen atoms of all phenyl groups. This makes sense for the gas phase transfer protocol but not for the ONIOM protocol, where one phenyl hydrogen atom might be involved in hydrogen bonding to a neighboring asymmetric unit and another one is not. Simply making all phenyl hydrogen atoms equal implies a loss of information. To circumvent this, a geometry matching algorithm is applied instead of invariom based equivalence determination. The algorithm is an implementation of the iterative closest point algorithm published by Besl and McKay (1992) and finds the best superposition of the structure motive in the studied structure and the atomic coordinates of the ONIOM output.

From this point on, the procedure is equivalent to the previously described protocol.

B.1.3. Further Analysis

It can be desirable to perform further analysis with the estimated ADPs. *APD-Toolkit* provides several tools to facilitate that.

[4] Even the same crystal measured at a different temperature can require re-computation due to changes in atomic coordinates.

[5] At this point it is possible to filter the frequency data to approximate certain vibrational features. The nature of the filtering will not be discussed here and is subject to further research.

- ADPs can be scaled to reference data sets via an appropriate least-squares procedure.

- ADPs can be compared to each other to yield scalar similarity indicators.

- ADP data can be exported to visual inspection software like *PEANUT* (Hummel *et al.*, 1990).

- All data is exposed to the plug-in system of *Laue-Script* and can be accessed globally. This implies that custom data analysis routines can be implemented quickly.

B.2. Reproducing Results in this Thesis

The results presented in this thesis require substantial programming effort to reproduce. Therefore *APD-Toolkit* is made freely available to aid in reproducing them.[6] Also, a short description is given explaining the required steps to estimate hydrogen ADPs with *APD-Toolkit*:

- If the user does not specify an input file containing a crystallographic structural model, the program will search for appropriate data files in the current working directory. If multiple files are found, the most recently written one will be used. To override this behavior, the option *load* can be used e.g. *apdtoolkit load <fileName>*.

- This will trigger the program to load the structural model defined in the data file. Subsequently the program will perform invariom partitioning (generating invariom names for all atoms) and then transfer internal ADPs from the invariom database to the loaded model. If the appropriate cache files are missing, a new cache file will be computed based on the temperature specified in the data file.[7]

- At this point the program will exit. In order to perform useful operations with *APD-Toolkit* the user must tell the program what to do via plug-in call commands. Each plug-in has an associated key that is used to trigger its execution. This is done by passing the key prefixed with a '-' character as an command line argument to

[6] https://github.com/JLuebben/APD-Toolkit

[7] Not all supported data formats include the diffraction temperature. In these cases a default of $100\ \mathrm{K}$ is assumed. This value can be overridden with the option *temp* e.g. *apdtoolkit load <filename> temp 50*.

the program. For example the TLS-analysis program has the key 'T' and can be called by typing *apdtoolkit load <filename> -T*.[8]

- It is most likely desirable to store the results of any computations the program performed on the hard drive. This is done with the Writer plug-in. The Writer plug-in creates a copy of the input data file with modified parameters depending on the operations performed before calling the Writer. To combine the internal ADPs from the invariom database with TLS estimations from the TLS-analysis plug-in the following commands can be used: *apdtoolkit load <filename> -T -W*.[9]

- For most default applications the command *apdtoolkit load <filename> -A -W* will be appropriate. The key 'A' triggers the Autosegment plug-in which performs an automated rigid body segmentation and then triggers an appropriately configured TLS-analysis.

B.3. Plug-in Documentation

This section lists a selection plug-ins available for *APD-Toolkit*. A short description for each plug-in is provided. For more detailed documentation the corresponding plug-in files should be consulted. The key triggering the plug-in execution is provided for each plug-in after its name separated by a '–' character. Some plug-ins take references to *ADP-Keys* as arguments. *ADP-Keys* are names given to certain representations of ADPs (fractional space, Cartesian space) and/or parts of an ADP (internal, external). Table B.1 lists all available *ADP-Keys*. Similar to *ADP-Keys* some plug-ins require references to a specific data set. The main data set (specified via *apdtoolkit load <FileName>*) is always stored with the key *exp*. Additional data sets can be stored at any time and referenced by their given name. For example a model compound's geometry is always loaded together with the invariom taken from that compound. In that case the model compound data set is stored with the compound's name as its key.

Some options require multiple arguments. If that is the case each argument is separated by a ':' character.

[8] Plug-in keys can be multiple characters long. The most commonly used plug-ins were given single character keys to reduce the amount of typing required to execute the program.

[9] The default behavior of the Writer plug-in assumes that this is what the user wants to do. The behavior of plug-ins can be fine-tuned with dedicated commands documented for each plug-in.

Key	Description
cart_int	Internal ADP in Cartesian space
frac_int	Internal ADP in fractional space
cart_ext	External ADP in Cartesian space
frac_ext	External ADP in fractional space
cart_sum	Sum of internal and external ADP in Cartesian space
frac_sum	Sum of internal and external ADP in fractional space
cart_meas	ADP read from a data file in Cartesian space
frac_meas	ADP read from a data file in fractional space

Table B.1.: Description of the most common *ADP-Keys*.

Autosegment – A Plug-in for automatically segmenting a molecule (or multiple molecules) into an ARG model. Subsequently, the TLS plug-in is called to perform an appropriately configured TLS+ARG-Analysis. The analysis will correct for correlation between internal and external anisotropic proton displacement (APD)s and will perform a single fit for each molecule in the asymmetric unit.

Compare – compare Plug-in for comparing ADPs of two similar structural models of the same compound. A scalar comparison value (see section 4.1.3 for details) is computed for each pair of equivalent atoms in both models. Equivalent atoms must be named equally in both models.
Options:

load <**filename**> name of the data file specifying the second structural model.

use <**ADP-Key1**>**:**<**ADP-Key1**> ADPs with *ADP-Key1* from the main data file are compared to ADPs with *ADP-Key2* from the data file specified via the *load* option.

CrossCheck – C Plug-in for estimating parameter standard deviations based on $R_{complete}$ computations.

path <**somePath**> Directory the program is looking for SHELXL output files.

mask <**partialFileName**> Files that do not start with <partialFileName> will be omitted.

list <**number**> The output provides a list of atoms starting with the atom with the largest variance in positional parameters. <number> specifies the number of atoms that are listed.

gt <**number**> The output list can be truncated to list only atoms with a variance greater than <number>.

residue <**name**> Limit the output to atoms belonging to residues with the specified <name>.

type <**element**> Limit the output to atoms of the specified <element>.

sigma <**number**> Writes a PDB file listing all atoms with a variance greater than <number>.

cutoff <**number**> Writes a PDB file listing all atoms with a variance smaller than <number>.

Descent – descent Writes a file listing all atoms and their corresponding invariom names and model compounds.

Expander – expand Expands the asymmetric unit to fill a whole unit cell.

GetHDist* – gethdist Generates a database file listing all X–H distances in the invariom database.

Hirshfeld – H Computes Hirshfeld test values for all (bonded) atom pairs.

use <**ADP-Key**> Key specifying which ADP representation should be used for the test.

full Triggers computation of test values for all atom pairs. Otherwise only bonded atom pairs are evaluated.

InvCif – cif Plug-in for preparing XD generated CIF file for publication. The plug-in includes features dedicated to invariom refinement. It reads a series of CIF files, joins them and edits them in a way suitable for publication.

load <**FileName**> Name of main CIF file. Defaults to newest CIF file in working directory.

write <**FileName**> File name of the program output.

include <**Path1**>**:**<**Path2**>**:...** A colon separated list of directories that are scanned for additional CIF files.

exclude <**FileName1**>**:**<**FileName2**>**:...** A colon separated list of CIF files that are excluded.

size <**a**>**:**<**b**>**:**<**c**> Crystal dimensions.

authors <name1>:<name2>:... Names of authors. Authors can be added to a database file.

temp <number> Diffraction temperature.

omit <CifKey1>:<CifKey2>:... List of CIF items that are omitted from the final file.

sadabs <Path> Path to a sadabs output file (*.abs) that may contain information about the performed absorption correction.

p4p <Path> Path to a P4P that may contain detailed cell information.

hkl <Path> Path to the xd.hkl file. The file will be embedded into the final CIF file to archive the diffraction data together with the model parameters. Defaults to <./xd.hkl>

nohkl Triggers the omission of the xd.hkl file from the output file.

shelx Switches off features that are not required for processing CIF files written by SHELXL.

nodetails Triggers the omission of the xd.res parameter file from the output file.

Leek – leek Plug-in for estimating anisotropic rigid body vibrations from ONIOM point mass computations.

data <Data-Key> The ADPs of the atoms of data set <Data-Key> will be overridden with estimated ADPs.

Micro – micro Replaces the default interface to the invariom database with an interface suitable for processing ONIOM data.

generate Triggers the database base generation mode. This mode is used to generate a micro database file based on an GAUSSIAN (Frisch *et al.*, 2013) output file.

load <FileName> If in database generation mode, *load* specifies a GAUSSIAN output file that contains the required frequency data. Otherwise it specifies the crystallographic data file containing the structural model the database information is applied to.

cluster <number> Number of molecules in the ONIOM cluster. This number is only used if the algorithm determining the cluster size automatically fails.

match <Key> Legal keys: *geom*, *trust*, *inv*: *geom* sets ADP transfer mode to iterative closest point algorithm. *trust* assumes that the ordering of atoms in the database and the data file are equal. *inv* applies the invariom transfer scheme that is usually not suitable for this application.

Peanut – peanut Plug-in for generating input files for the program PEANUT. The plug-in requires a second structural model to which the main model is compared to.

load <**FileName**> Name of a crystallographic data file.

use <**ADP-Key**> Key specifying which ADP data from the main model is compared to the newly loaded model.

PQR – pqr Plug-in for writing a PQR formatted file.

PsiPole – Psi Prototype implementation of the BODD model.

RealResp – realresp Plugin for estimating RESP charges based on the invariom database.

Resp – resp Plugin for estimating RESP charges based on the invariom database.

Restrain – restrain Plug-in for generating geometry restraints from the invariom database.

write <**FileName**> Name of the output file name containing a listing of SHELXL style restraints.

Scale – S Plug-in for scaling the ADPs of the main data set to the ADPs of a reference data set.

load <**FileName**> Name of the data file containing the reference data set.

use <**ADP-Key1**>:<**ADP-Key2**> The ADPs stored as ADP-Key1 is scaled to the reference data set and than saved with the key ADP-Key2.

THMAReader – thma Plug-in for reading THMA output files and storing the ADP information in the main data set.

load <**FileName**> Name of the THMA output file.

TLS – T2 Plug-in for performing TLS+ARG-Analysis.

molecule <**ID**> Integer specifying for which molecule in the asymmetric unit the analysis should be performed.

data <**Data-Key**> Key specifying which data set should be used for the analysis.

correlate By default correlation between internal and external vibrations is corrected by subtracting internal ADPs from the optimized ADPs before performing the analysis. This behavior can be switched off with this trigger.

Write – W Plug-in for writing crystallographic data files.

write <**FileName**> Base of the output file name. The file suffix will be added automatically based on the format of the input file.

use <**ADP-Key**> Key specifying which ADP data should be written to the output file.

data <**Data-Key**> Key specifying which data set should be written to a file.

Acknowledgment

First I would like to thank my supervisors George M. Sheldrick and Birger Dittrich for their support and the freedom to pursue my own research ideas.

I would like to thank Prof. Dr. Ricardo Mata, Dr. Heidrun Sowa, Prof. Dr. Dietmar Stalke and Prof. Dr. Hartmut Laatsch for being part of my examination commission.

I would like to thank Tim Grüne for many helpful discussions and very enjoyable collaboration as well as my collaboration partners in the Clever group and the Roesky group.

I would like to thank Tim Grüne, Claudia Wandtke and Anna Lübben for proofreading and for providing a lot of helpful feedback that helped finalizing this thesis.

Furthermore I would like to thank everyone else who helped, making the last three years as enjoyable as they were. This includes Claudia Wandtke and Anna Lübben with whom I shared an office, Lennard Krause, Felix Engelhardt and Christian Schürmann from the Stalke group, Sofiane Saouane and Rubén Granero from the Fabbiani group, Massimo Sammito, Claudia Milan, Rafael Borges from the Usón group in Barcelona, Tim Grüne from the PSI in Switzerland, Christian Hübschle from the Smaalen group in Bayreuth, Julian Holstein from the Clever group in Dortmund and all the other fellow researchers I had the pleasure of meeting.

Finally I would like to thank my wife Anna for her continuing support.