Algorithms for Crystal Structure Determination in Macromolecular Crystallography

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
Anna V. Lübben
aus Unna

Göttingen, 2019

Betreuungsausschuss

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

Prof. Dr. Kai Tittmann
Department of Molecular Enzymology, Georg-August-Universität Göttingen

Prof. Dr. Dr. med Robert Steinfeld
Klinik für Neurologie, Universitäts-Kinderspital Zürich

Mitglieder der Prüfungskommission

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

Korreferent: Prof. Dr. Kai Tittmann
Department of Molecular Enzymology, Georg-August-Universität Göttingen

Prof. Dr. Dr. med Robert Steinfeld
Klinik für Neurologie, Universitäts-Kinderspital Zürich

Prof. Dr. Inke Siewert
Institut für Anorganische Chemie, Georg-August-Universität Göttingen

Jun.-Prof. Dr. Nathalie Kunkel
Institut für Anorganische Chemie, Georg-August-Universität Göttingen

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

Tag der mündlichen Prüfung: 21. Juni 2019

“Somerville College May 17th, 1931

My dearest Mummy and Daddy,

(...) A few days ago Dr. Joseph wrote to me to say that he had asked Professor Lowry about the possibility of my doing X-ray work on crystals – and whether it was a good thing. (...) And all that sounded very nice – really excellent just then – since the X-ray work would be useful in absolutely anything I decided to do ever afterwards and yet if I did not do it now – I probably should not have the chance again. But at the moment I’m feeling quite appalled at the prospect. There will be such a fearful lot of work – and mathematics – involved. And I was just beginning to rejoice so much in the idea of a nice quiet organic research that would involve no brain whatsoever. As it is, it will be pure brain work – I’m just shivering in my shoes terribly afraid I really am trying to force too much on one poor little brain that is almost non-existent already. (...)
Of course, if I can really do it it will be rather priceless ...”

– Dorothy Crowfoot Hodgkin (1910 - 1994)

Contents

List of Figures

Acronyms

ABCA1 ATP-binding cassette transporter 1

ADP atomic displacement parameter

ApoA1 apolipoprotein A1

ATP adenosine triphosphate

ATPase adenosine triphosphate synthase

BMA β-D-mannose

CTSD cathepsin D

CerS dihydroceramide synthase

CI-MPR cation-independent mannose-6-phosphate receptor

DNA deoxyribonucleic acid

EEG electroencephalogram

EndoH endo-β-*N*-acetylglycosaminidase H

ER endoplasmic reticulum

ERG electroretinogram

e.s.d. estimated standard deviation

GFP green fluorescent protein

GUI graphical user interface

HMM hidden Markov model

LSD lysosomal storage disease

MAN D-mannose

Man-6-P mannose-6-phosphate

MPR mannose-6-phosphate receptor

MRI magnetic resonance imaging

NAG *N*-acetyl-D-glucosamine

NCL neuronal ceroid lipofuscinosis

PDB protein database

PLTS phospholipid transfer protein

PPT1 palmitoyl-protein thioesterase 1

r.m.s. root mean square

RNA ribonucleic acid

SAD single-wavelength anomalous diffraction

saposin sphingolipid activator protein

SDS-PAGE sodium dodecylsulfate polyacrylamide gel electrophoresis

SeMet selenomethionine

s.u. standard uncertainty

TBM template based modeling

TPP1 tripeptidyl-peptidase 1

XRD X-ray diffraction

1. Theoretical Background and Methods

Theoretical Background and Methods

1.1. X-ray diffraction

The discovery of X-rays and the subsequent detection of diffraction of X-rays by crystals by Max von Laue mark the beginning of crystallography (Eckert, 2012). Using X-ray diffraction (XRD), the first crystal structure was solved by William Henry Bragg and his son William Lawrence Bragg in 1915 (Bragg, 1962). The Bragg equation (Equation 1.1) describes the interference of light waves with a point lattice.

$$d_{\mathsf{min}} = \frac{\lambda}{2\sin\theta} \tag{1.1}$$

Macromolecular crystallography evolved into the primary method for determining macromolecular structures and provides insight to the function of proteins and their complex assemblies through detailed structure models.

Still, the availability of well diffracting crystals remains a prerequisite. X-ray radiation interacts with matter through an oscillating electric field vector which interacts with electrons by polarization. As a result of this interaction the electrons emit electromagnetic waves of the same frequency. Superposition of this electromagnetic waves give rise to interference. This diffraction phenomena results in a diffraction pattern as a sum of all scattering events. One has to keep in mind that electrons are not located at grid points in a lattice but moving around atomic nuclei which are distributed in the crystal in an ordered fashion.

The diffraction pattern is recorded in the form of Bragg reflections hkl with specific intensities. The intensities of all unique reflections are used to reconstruct the electron density of the molecule that gave rise to the diffraction pattern. All scattering contributions of each atom in all unit cells to a reflection are described in the form of a structure factor. The amplitude of the structure factor of a reflection is used to reconstruct the electron density in the crystal by calculating the Fourier summation. The Fourier transformation needs both the amplitudes and the phases of the experiment to reconstruct the electron density. Since only the amplitudes are measured during XRD, the phases must be

recovered – this phenomenon is known as the phase problem.

A comprehensive overview of the fundamentals of crystallography is provided by Giacovazzo *et al.* (2011), Massa (2009), Borchardt-Ott (2012), and Rupp (2011).

1.1.1. Anomalous diffraction

Friedel pairs are reflections related by inversion through the origin (North, 1965). The intensities of a Friedel pair of reflections (hkl and $\bar{h}\bar{k}\bar{l}$) are approximately equivalent and their phase is opposite, known as Friedel's law.

Absorption and subsequent re-emission of X-ray photons coupled with ionisation of the atom is known as anomalous scattering. Anomalous scattering generates anomalous differences in the structure factor amplitudes of symmetry related reflections. Due to the anomalous scattering contribution, Friedel's law breaks down and the reflections are forming Bijvoet pairs with different amplitudes.

The anomalous differences can be utilized for experimental phasing of crystal structures. The absorption of X-ray radiation decreases with increasing wavelength. At a given wavelength the mass absorption coefficient is higher for heavier elements than for light elements. This correlation is used in the structure determination of macromolecular structures (Rossmann, 1961). Heavy elements or metals can be incorporated into the crystal to facilitate anomalous diffraction at shorter wavelength that are experimentally more easy to access (Garman and Murray, 2003). For some biomolecules intrinsic elements can be used, such as phosphorus in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) and sulfur in proteins (North, 1965). Hendrickson and Teeter (1981) solved the structure of Crambin solely from anomalous signal of intrinsic sulfur atoms. Other methods for the incorporation of anomalous scatterers are co-crystallization or soaking of the crystal with suitable compounds (Taylor, 2010).

1.2. Quality indicators in X-ray diffraction[1]

Collecting accurate intensities for all Bragg reflections hkl is a critical objective of any XRD experiment. Various quality indicating parameters have been established over the years to evaluate the XRD experiment.

[1] The definitions in this chapter are based on the collection as presented in the International Tables for Crystallography, Volume F (Arnold *et al.*, 2012).

1.2.1. Quality indicators in general

Both random error and systematic error influence the measured intensities. It is the goal of early data processing to separate the true signal from the noise. Relative random errors can be reduced by increasing exposure time. Systematic errors are approximately proportional to measured intensities.

Systematic errors can occur due to technical or macroscopic aspects of the experiment, e.g. beam instability in flux or direction, due to the shutter or vibration of the crystal in the cryo-stream. Also radiation damage, absorption of the loop or detector issues such as overloads or shadows can add to the systematic error. These are only a few of the possible problems adding to the observed overall error and are only partially under the control of the experimenter.

Random error is the chance that not the most likely number of diffraction events was observed. This is due to the fact that diffraction events follow a Poisson distribution.

One has to distinguish between accuracy and precision when evaluating the errors influencing the measured intensities. Precision quantifies the average deviation of a series of measurements from its mean value. It can be understood as the internal consistency of the signal. Accuracy describes the deviation of the collected values from the true values. Accuracy is generally more important than precision. While low precision yields imprecise results, low accuracy leads to wrong results. The goal of the experimenter is the optimization of both accuracy and precision.

To describe the quality of a data set, numerous quality indicators have been defined over the time and their convenience is often dependent on the intent of their use. Many indicators of data quality are dependent on estimates for the standard uncertainties (ss.u.s) and therefore on the methods for estimating them. All initial estimates of the s.u.s of each intensity observation are generally underestimated by integration programs (Evans and Murshudov, 2013). It should be noted that most quality indicators can only describe the precision of the collected signal and not the accuracy (Diederichs, 2016). Since the accuracy is not available for direct evaluation, the minimization of undetected errors is significant for data collection and processing.

In the following sections the most commonly used quality indicator, and those used in this thesis are discussed.

Determination of the resolution

Early definitions of the resolution of a data set were subjective and dependent on the experimenter, like the nominal resolution d_{min}. Here, the limit is defined as a fraction of

the unique reflections, e.g. 70%, that are above a threshold, for example set at three times their s.u.s. Another definition is the midpoint of the resolution range of the shell at which the mean signal-to-noise ratio falls below 2.

The true resolution d_{true} is given as the minimum distance between two objects in a crystal that permits their images in the resulting electron density map to be resolved. When two equivalent atoms are represented by Gaussians, they can be considered resolved when the electron density value drops to zero at midpoint between them[2]. A consistent use of the resolution of data sets is favorable since it is often used to determine the resolution cutoff, not only during model refinement but integration as well.

In this work the resolution of a data set is defined by quality indicators that are discussed in the following sections. The resolution is used to define the point to which diffraction images are integrated during data processing.

1.2.2. Data quality of unmerged intensities

Crystallographic residual index factors (R-factors) are widely used as a measure of data or model quality (Einspahr and Weiss, 2012).

Merging R factor

The merging R factor R_{merge} describes the spread of the individual intensity measurements I_i of a reflection hkl around the mean intensity $\langle I(hkl)\rangle$ of this reflection.

$$R_{\text{merge}} = \frac{\sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|}{\sum_{hkl}\sum_i I_i(hkl)} \tag{1.2}$$

The sums run over all observed unique reflections (hkl) and over all individual observations i of a given reflection.

R_{merge} is dependent on the multiplicity of a data set (Diederichs and Karplus, 1997, Weiss, 2001, Weiss and Hilgenfeld, 1997). The R_{merge} will be higher with a higher multiplicity of the data set. This conflicts with the statistical expectations, that averaged intensities values should be more precisely determined. Therefore, the merging R factor proved not useful as general data quality indicator for diffraction data (Evans and Murshudov, 2013, Weiss, 2001). Nonetheless, $R_{\text{merge}} \geq 0.6 - 0.8$ has been used to determine the resolution cutoff for a long time (Karplus and Diederichs, 2012).

[2] This is not a reasonable criteria for atoms in molecules as they are found in real crystals.

Redundancy independent merging R factor

The precision of the individual intensity measurement is better described with the redundancy independent merging R factor (R_{rim} or R_{meas}). The R_{rim} is independent from the number of observations of the individual reflection.

$$R_{\text{meas}} = R_{\text{rim}} = \frac{\sum_{hkl}\{\frac{N(hkl)}{N(hkl)-1}\}^{1/2} \cdot \sum_i |I_i(hkl) - \langle I(hkl)\rangle|}{\sum_{hkl}\sum_i I_i(hkl)} \tag{1.3}$$

$\langle I(hkl)\rangle$ describes the mean of the $N(hkl)$ individual measurements $I_i(hkl)$ of the intensity of a reflection hkl. R_{rim} should be substituted for the conventional R_{merge} (Diederichs and Karplus, 1997, Weiss, 2001, Weiss and Hilgenfeld, 1997).

Mean signal-to-noise ratio

As a measure of the overall precision of a data set, the mean signal-to-noise ratio $\langle I/\sigma(I)\rangle$ (mean I/σ(I)) finds broad application. It describes the statistical significance of a measured intensity and for all reflections the averaged intensity as a multiple of the s.u.s. There are two definitions of $\langle I/\sigma(I)\rangle$ that find application in macromolecular crystallography:

$$\begin{aligned} \frac{\langle I(hkl)\rangle}{\sigma[I(hkl)]} &= \frac{\langle I(hkl)\rangle}{[(\frac{1}{N})|I_i(hkl) - \langle I(hkl)\rangle|^2]^{1/2}} \quad \text{and} \\ \frac{\langle I(hkl)\rangle}{\sigma\langle I(hkl)\rangle} &= \frac{\langle I(hkl)\rangle}{[(\frac{1}{N})\sum_i \sigma_i(hkl)^2]^{1/2}}. \end{aligned} \tag{1.4}$$

The first definition describes the ratio of the mean intensity $\langle I(hkl)\rangle$ to the root mean squared (r.m.s.) scatter of the individual reflections about that mean. It does not take into account multiplicity or redundancy of reflections.

The second definitions describes the average of all observations of the reflection (hkl). Here $\sigma_i(hkl)$ describes the experimental s.u. of the individual measurement and $\sigma\langle I(hkl)\rangle$ is the propagation-of-error combination.

The mean signal-to-noise ratio of the outer resolution shell is often used to define the nominal resolution of a data set. Furthermore, this is used as indicator for the highest resolution shell that should be used for refinement (Wang, 2010). It is common practice to define mean I/σ(I) $\sim$ 2 as integration limit during data reduction. The indicator is depending on σ(I), which can be mis-estimated (Evans, 2006, 2011).

Asymptotic signal-to-noise ratio

A visual representation of the data quality is a plot of $I/\sigma(I)_{asymptotic}$ against the resolution. ISa is the highest observed asymptotic signal-to-noise ratio of a data set (Diederichs, 2010). The sigmoidal curve reaches an upper limit that can serve as a guide for the data quality. The limit should be as high as possible and is sensitive to systematic errors – instrument errors manifesting themselves in the data set. The sensitivity for systematic error arises from the approximate independence of random error for very high values (Diederichs, 2016). The measure is nearly independent from counting statistics, provided radiation damage is negligible. Maximizing ISa should be the goal of data processing and is an indicator of a good data set in general. Diederichs (2016) suggested that an ISa ~ 30 is as high as can be achieved from a charge-coupled device (CCD) detector and indicates a good data set.

1.2.3. Precision of averaged reflections

Precision-indicating merging R factor

Especially to describe the precision of the averaged intensity measurements, the R_{pim} was introduced analogous to R_{rim}.

$$R_{\text{pim}} = \frac{\sum_{hkl}\{\frac{1}{N(hkl)-1}\}^{1/2} \cdot \sum_i |I_i(hkl) - \langle I(hkl)\rangle|}{\sum_{hkl}\sum_i I_i(hkl)} \tag{1.5}$$

The R_{pim} accounts for the increase in precision of the intensities when merging more observations. Decreasing with increasing redundancy, the R_{pim} is also a useful statistic for the estimation of data quality for anomalous diffraction data sets (Weiss, 2001).

Correlation coefficient

The Pearson correlation coefficient was adapted as a measure of data quality in the form of $CC_{1/2}$ (Diederichs and Karplus, 2013, Evans and Murshudov, 2013, Karplus and Diederichs, 2012). This precision indicator describes the correlation between two random subsets of the merged intensities (Evans, 2006). $CC_{1/2}$ is independent of estimated standard deviations of intensities and does not suffer an increase through most systematic errors (Diederichs, 2016, Diederichs and Karplus, 2013). It has been suggested to integrate and process raw data up to a $CC_{1/2}$ of 10% and deposit it with the final structure (Diederichs, 2016, Karplus and Diederichs, 2015). Including all data

until a $CC_{1/2} \sim 20\%$ can improve the solution and resulting model quality (Wang, 2010) compared to data truncated at cutoff limit given by R_{merge}.

1.2.4. Data quality of the anomalous signal

The presence and accuracy of anomalous signal is of special interest for anomalous scattering phasing procedures. When averaging multiple data sets from a single crystal or from multiple crystals (MCA), the quality of the merged data set is often the crucial criterion for the success of the structure solution (Akey *et al.*, 2014). In difficult cases only the use of high-multiplicity data from multiple samples was effective for the solution of the anomalous scatterers' substructure (Akey *et al.*, 2014, Liu *et al.*, 2012). It has become common practice to use the peak wavelength alone to collect data for single-wavelength anomalous diffraction (SAD) phasing (Rice *et al.*, 2000). The efficiency of merging with respect to the anomalous signal is therefore often the crucial step for further success (Akey *et al.*, 2014). The quality of the averaged measurements can be evaluated by indicators specific for the anomalous signal.

Anomalous merging R factor

The anomalous R factor R_{anom} is used to describe the anomalous signal strength.

$$R_{\text{anom}} = \frac{\sum_{hkl} |I(hkl) - I(\bar{h}\bar{k}\bar{l})|}{\sum_{hkl} \langle I(hkl) \rangle} \tag{1.6}$$

with $\langle I(hkl) \rangle = \frac{1}{2}[I(hkl) + I(\bar{h}\bar{k}\bar{l})]$. The ratio between R_{pim} and R_{anom} has been suggested as an indicator of the strength of the anomalous signal of a data set (Panjikar and Tucker, 2002). When the ratio of the precision indication R value and the anomalous merging R value ($R_{\text{pim}}/R_{\text{anom}}$) exceeds 1.5, the substructure solution with the anomalous differences becomes achievable (Weiss, 2001).

Anomalous correlation coefficient

With a better understanding of factors influencing the quality of the data, the anomalous correlation coefficient CC_{anom} was introduced (Zwart, 2005).

$$CC_{\text{anom}} = \frac{\sum (x - \langle x \rangle)(y - \langle y \rangle)}{[\sum (x - \langle x \rangle)^2 \sum (y - \langle y \rangle)^2]^{1/2}}, \tag{1.7}$$

where x and y are the anomalous differences $[I(hkl) - I(\bar{h}\bar{k}\bar{l})]$ in the two data sets and $\langle x \rangle$ and $\langle y \rangle$ are their averages. Also called Pearson's CC, CC_{anom} of 30 or higher is regarded as a good indicator for an adequate anomalous signal. A resolution cutoff for anomalous phasing has been suggested at CC_{anom} of 10 (Schneider and Sheldrick, 2002). Choosing an adequate cutoff for structure solution can be critical for its success (Sarma and Karplus, 2006).

1.3. Data collection and processing

1.3.1. Data collection

XRD images collected on hybrid pixel detectors, such as the Dectris Pilatus and Eiger detectors, have different characteristics when compared to CCD detector images (Hülsen *et al.*, 2006). Hybrid pixel detectors have fast readout times and no readout noise. Fine ϕ-slicing was suggested as strategy for these single photon counting detectors (Mueller *et al.*, 2012).

Using this strategy, the accuracy of strong reflections should improve and the mosaicity can be calculated more accurately (Hülsen *et al.*, 2006, Kraft *et al.*, 2009). For data collection with Pilatus detectors, an oscillation range of $\Delta\phi$ of 0.1 – 0.2° per image is commonly used. The Eiger detector installed at the SLS PXI X06SA undulator beamline is the next generation of single photon counting detectors. With an even shorter dead time and faster frame rates, the detector allows data collection with an ultra-fine ϕ-slicing method (Casanas *et al.*, 2016). With this detector, data is commonly collected with an oscillation range of $\Delta\phi$ of 0.04 – 0.1° per image.

1.3.2. Data reduction and scaling

All frames collected at synchrotron undulator beamlines were processed using the XDS software (Kabsch, 2010). XDS control files were provided by the corresponding beamline. These control files contain all parameters needed for the integration and command the steps the program performs.

XDS determines the initial detector background, strong reflections for indexing, unit cell dimensions, crystal orientation, and the active detector area in the first steps. The next step is the integration of all frames and thereby the estimation of all reflection intensities. In a last step, corrections are applied, the reflections are scaled, and statistics are generated. For all steps output files are written that can be reviewed manually or by

using XDSgui[3]. The correction step provides an output file with an overview of the most common data quality indicators.

Optimization of data integration

For optimization, all data are re-processed as suggest by the XDSwiki[4], by Mueller *et al.* (2012) and Diederichs (2016). First, space group and cell parameters are inspected and transferred into the XDS control file. High and low resolution cutoff can be adjusted depending on the statistics and result files provided by XDS. The integration is performed without merging the Bijvoet pairs.

For each new integration cycle some files need to be renamed and parameters transferred to the XDS control file. The output file containing the latest geometry description is recycled with each iteration, so that the newest file is used by XDS. This results in more reliable statistics and a better anomalous signal. Next, the beam divergence and mosaicity are updated in the XDS control file. These values are refined with each run of XDS and can lead to better R-factors, if recycled (Diederichs, 2016).

The integration and correction step in XDS are repeated at least three times. The output file of the correction step is inspected between each cycle. Optimization should improve the overall statistics of ISa, mean I/σ(I), and $CC_{1/2}$. ISa is used as indicator for systematic error and $CC_{1/2}$ is used to evaluate the precision of the merged intensities. The precision of the unmerged intensities is evaluated using R_{rim} and mean I/σ(I).

The error model estimation is validated via χ^2 as a function of resolution and intensity. A value close to one over the resolution range indicates a good fit of the error model.

When an anomalous scatterer is present, the values of 'SigAno' (anomalous signal-to-noise ratio, d"/σ(d")) and 'AnomalCorr' (CC_{anom}) are also inspected.

During the repeated integration cycles a high resolution cutoff is applied. The cutoff is either chosen as the edge of the detector surface or at the resolution where $CC_{1/2}$ falls below 30 % (Karplus and Diederichs, 2012).

For SAD data, or in general when an anomalous signal is present, the integration is repeated with a focus on the absorption correction. It was reported that the anomalous signal is better described when the absorption correction was applied (Akey *et al.*, 2014). This is tested by comparing the resulting statistics of the correction step, once performed with absorption correction and once without.

[3]W. Brehm, K. Diederichs and M. Hoffer, *https://sourceforge.net/projects/xdsgui/*.
[4]University of Konstanz, *https://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Optimisation*.

Scaling

The programs PHENIX_scale_and_merge (Adams *et al.*, 2010), XPREP[5] and XSCALE (Kabsch, 2010) were used to process the datasets. If the aforementioned programs were used for scaling, the scaling option intrinsic to XDS during the correction step was not employed. XSCALE and PHENIX_scale_and_merge both scale and merge the data sets in one step (see Section 1.4). XPREP is a multi purpose program and does not scale the intensities before data averaging. The script XDS2SAD[6] is used to convert the XDS data file into the input format for the program SADABS[7]. SADABS is used to scale the raw data and convert it to the appropriate format for XPREP. Averaging of equivalent reflections is performed by iterative improvement of weights of the intensities. Friedel pairs are treated as non-equivalent reflections when anomalous signal is present.

1.4. Multi crystal averaging

1.4.1. XSCALE

The program XSCALE scales and averages reflections obtained from XDS. When the single scans are associated with a crystal name, the program can perform zero dose extrapolation to correct for radiation damage (Diederichs *et al.*, 2003). This option was used whenever more than one data set was obtained from one crystal. The output file provides a wealth of information to evaluate the quality of the merging. Parameters such as ISa for the whole data set and $CC_{1/2}$, mean I/σ(I), CC_{anom}, R-value or anomalous density by resolution were reviewed. Some options already available in XDS are implemented in XSCALE as well, such as scaling and absorption correction.

1.4.2. PHENIX_scale_and_merge

PHENIX_scale_and_merge uses local scaling and a multi-step merging approach (Akey *et al.*, 2016). During averaging of equivalent reflections from multiple data sets, the Friedel pairs are excluded and their accuracy is optimized in a separate step. The weight of anomalous differences is optimized by comparing the anomalous differences from individual data sets with those of the merged dataset. Optionally, outlier data sets can be excluded. However, this option was not used in this thesis. The data quality of

[5] *G.M. Sheldrick*, Bruker AXS Inc., Madison, Wisconsin, USA, 2003.
[6] *G.M. Sheldrick*, *www.shelx.uni-goettingen.de*.
[7] *G.M. Sheldrick*, Bruker AXS Inc., Madison, Wisconsin, USA.

the merged file is evaluated by the value of CC_{anom} against different resolution cutoffs, reported as $CC_{\text{cumulative}}$. The values for $CC_{\text{cumulative}}$ at a resolution of 6.0 Å are used as an indicator for the data quality in this work.

1.4.3. XPREP

XPREP is a command line program used for analysis and preparation of data which is also capable of merging and scaling. After space group determination, the scaling factors are determined by least-squares optimization of equivalent reflections and outliers in the data set are down-weighted. Data quality can be reviewed by various data statistic tables and plots, including the anomalous signal-to-noise ratio (d"/σ(d")).

1.4.4. SHELXC

SHELXC (Sheldrick, 2010) was designed as a data-preparation program for structure solution coming from an integration or scaling program. The development version of the program capable of averaging data sets assigns low weighting factors to data sets disagreeing with other data sets, yielding the combined data set with the highest internal consistency. Data quality can be reviewed in the log file giving tables over-viewing data quality indicators such as $CC_{1/2}$, CC_{anom} for each data set and mean I/σ(I), CC_{anom}, d"/σ(d"), χ^2, R_{pim}, and R_{anom} by resolution.

1.5. Evaluation of the anomalous signal

It has been common practice to average the reflections of multiple data sets to improve data precision and phasing. In recent years the advantage of using intrinsic anomalous scatterers in macromolecules has been discussed (Rice *et al.*, 2000). Especially the possibility of using sulfur atoms in proteins is an attractive alternative of heavy metal soaking or modification with e.g. selenomethionine (SeMet). To enhance the anomalous signal of intrinsic scatters, data from many crystals can be merged (Liu *et al.*, 2012, 2013). The CFOM (combined figure of merit) value of a successful run of the data set with SHELXD (Schneider and Sheldrick, 2002) and the averaged anomalous density calculated from ANODE (Thorn and Sheldrick, 2011) can be used to evaluate the quality of a merged data set.

SHELXD employs direct methods and integrated Patterson to determine the marker-atom substructure (Schneider and Sheldrick, 2002). Correct solutions are identified

by correlation coefficients which are combined to the CFOM value. SHELXD takes all anomalous scatters present into account for the calculation of CFOM.

ANODE calculates the phases of the marker atom substructure from the native phases obtained from the structure. With these phases a heavy-atom density map is computed and the averaged anomalous density at specific atomic positions is calculated. ANODE requires the structure and files prepared from the reflections, which are produced by SHELXC or XPREP.

2. Poly(rA)

Poly(rA) – Quality indicators for data merging enhancing the anomalous signal

In 1961, A. Rich, D. Davies, F. Crick and J. Watson proposed a parallel double helix ribonucleic acid (RNA) structure for poly(rA) on the basis of diffuse fiber diffraction photographs (Rich *et al.*, 1961). Over 50 years later the suggested structure was confirmed via single crystal X-ray diffraction (XRD) experiments (Safaee *et al.*, 2013) (Figure 2.1). The formation of a right-handed, parallel nucleic acid duplex was shown to be thermodynamically favorable (Pattabiraman, 1986). The proposed model of the parallel double-helix of poly(rA) was expected to be stabilized by N1 protonation.

2.1. Introduction

The XRD structure of the poly(rA) parallel double helix was described by Safaee *et al.* (2013). The double helix in the structure of $(rA)_{11}$ is comprised of ten base pairs and one unpaired nucleotide at each terminal end. The A-A base pairing is characterized by the involvement of Watson-Crick and Hoogsteen faces of adenine (Holbrook and Kim, 2004, Rich *et al.*, 1961). The first X-ray crystal structure with a parallel double helical structure is the cytidylyl-3',5'-adenosine (CpA) proflavine complex (Westhof and Sundaralingam, 1980). In this structure a base pairing pattern mediated by four hydrogen bonds was described. The parallel duplex of $(rA)_{11}$ displays the same A-A base pairing.

Crystal structures of (deoxy)ribonucleic acid oligomers can be solved by standard methods of macromolecular crystallography. Molecular replacement with an plausible search model and the use of heavy atoms introduced into the structure are common techniques. Direct methods can be employed for small oligomers when well diffracting crystals are available. Dauter and Adamiak (2001) reported the successful structure solution of a DNA oligomer by phasing with the anomalous signal of intrinsic phosphorus.

Phosphorus single-wavelength anomalous diffraction (SAD) phasing remains the simpler choice for sample preparation but does not find broad application (Reyes *et al.*, 2009). The need for accurate measurements of the reflections amplitudes to determine

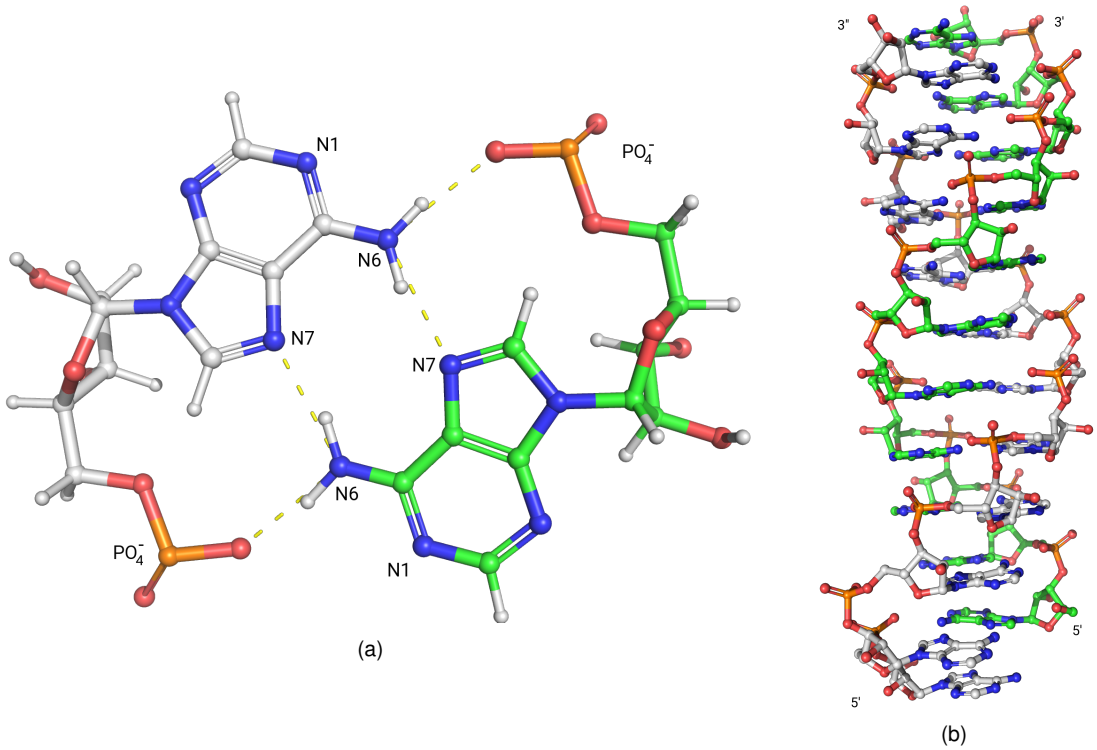


Figure 2.1.: The parallel, right-handed double stranded helix of $r(A)_{11}$. (a) The A-A base pairing driving the formation of the poly(rA) duplex. Four hydrogen bonds are formed including the phosphate groups of the RNA backbone. (b) The $r(A)_{11}$ duplex in ball-and-stick representation. The parallel duplex is comprised of ten base pairs and one nucleotide overhanging at each end.

the anomalous signal is a prerequisite. The average ratio of Bijvoet difference of phosphorus for data collected from DNA crystals is higher as that of sulfur atoms in proteins (Dauter and Adamiak, 2001). Still, data redundancy and adequate data reduction play a vital role in the phasing success of deoxyribonucleic acid (DNA) oligomers via the intrinsic anomalous signal (Dauter and Adamiak, 2001).

2.2. Aim of this work

The availability of highly ordered Poly(rA) crystals evoked the hope for charge density studies on the structure. With this objective, XRD measurements were carried out with the available crystals. In total 37 data sets were collected from eight crystals with a resolution of up to 0.7 Å. Since the analysis of charge density proved infeasible, the focus was shifted to use the data for multi crystal averaging studies.

The collected data are evaluated under different aspects concerning the data quality

and suitable indicators are discussed. XRD data was averaged in a multi-dataset from one crystal (MDS) approach as well as in a multi-crystal approach (MCA). The results of this study aimed at evaluating quality indicators for data averaging.

Furthermore, the program SHELXC by G. M. Sheldrick was extended to include an option allowing the averaging of multiple data sets. The data collected of Poly(rA) provided a suitable test set for the alpha test version of SHELXC with the new feature. The data sets of Poly(rA) are evaluated by different criteria to select which data sets are suitable for multi crystal averaging. Results of these evaluation can provide useful indicators for the implementation of the averaging function in SHELXC.

Additionally, the diffraction data was evaluated for an anomalous signal from the intrinsic phosphorus atoms. Even when the experimental setup was not chosen with a focus on phosphorus-SAD (P-SAD), significant signal was observed. The influence of data processing and averaging on the anomalous signal is evaluated. The systematic averaging of selected data sets to maximize the anomalous signal is discussed to enable P-SAD phasing.

2.3. Materials and methods

2.3.1. Data collection

Crystals of $r(A)_{11}$ forming a parallel double helix were obtained by Jingwei Xie and Nozhat Safaee in the laboratory of Prof. Gehring[1]. The crystals were of large size and diffracted to high resolution – an improvement over the crystals available when the structure was initially solved by Safaee *et al.* (2013).

The crystals were measured at undulator beamlines at DESY (german electron synchrotron, PetraIII P11) and SLS (swiss light source, PXII X10SA, PXI X06SA) at 100 K. Due to one long cell axis of 163 Å and the intent to collect data for charge density studies, a short wavelength was chosen for data collection. The XRD data were collected at the wavelengths 0.6359 Å, 0.6525 Å, 0.796 Å, and 0.7293 Å. The beamline P11 at DESY (PetraIII) allowed measurements at an energy of 19.5 keV, the beamlines X10SA (PXII) and X06SA (PXI) at SLS allowed a high energy limit of 19.0 keV and 17.5 keV, respectively. PetraIII P11 and PXII X10SA were equipped with Dectris Pilatus 6M detectors and PXI X06SA with a Dectris Eiger 16M detector. All beamlines were outfitted with single ϕ rotation axis goniometers.

[1] Department of Biochemistry, McGill University, 3649 promenade Sir-William-Osler, Montreal, Canada.

Table 2.1.: Overview of the data collection from Poly(rA) crystals. The scans are given by number (#) of the data set and sorted by beamline and crystal.

#	crystal	scan	beamline	wavelength [Å]	dd[a] [mm]	rotation [°]	slicing [°]
1	A1	1	P11	0.6525	300	90	0.2
2		2	P11	0.6525	200	180	0.2
3		3	P11	0.6525	180	180	0.2
4		4	P11	0.6525	180	180	0.2
5		5	P11	0.6525	300	180	0.2
6	A3	1	P11	0.6525	300	180	0.2
7		2	P11	0.6525	180	180	0.2
8		3	P11	0.6525	200	180	0.2
9	A4	1	P11	0.6525	155	180	0.2
10		2	P11	0.6525	300	180	0.2
11		3	P11	0.6525	200	180	0.2
12		4	P11	0.6525	230	180	0.2
13		5	P11	0.6525	250	180	0.2
14		6	P11	0.6525	250	180	0.2
15	B3	1	X10SA	0.6359	200	360	0.1
16		2	X10SA	0.6359	400	180	0.1
17		3	X10SA	0.6359	250	180	0.1
18	B4	1	X10SA	0.6359	180	180	0.1
19		2	X10SA	0.6359	400	180	0.1
20		3	X10SA	0.6359	200	180	0.1
21	C3	1	X10SA	0.6358	200	180	0.1
22		2	X10SA	0.6358	400	360	0.1
23		3	X10SA	0.6358	180	180	0.1
24		4	X10SA	0.6358	180	180	0.1
25		5	X10SA	0.6358	200	180	0.1
26		6	X10SA	0.6358	190	180	0.1
27	C2	1	X10SA	0.6358	350	180	0.1
28		2	X10SA	0.6358	200	360	0.1
29		3	X10SA	0.6358	180	180	0.1
30		4	X06SA	0.7293	300	180	0.1
31		5	X06SA	0.7293	140	180	0.1
32		6	X06SA	0.7293	140	180	0.1
33	C1	1	X06SA	0.7293	300	180	0.1
34		2	X06SA	0.7293	135	180	0.1
35		3	X06SA	0.7293	135	180	0.1
36		4	X06SA	0.7293	140	180	0.1
37		5	X06SA	0.7293	135	180	0.1

[a] dd = detector distance.

In total, 37 data sets were measured from eight crystals, an overview of the data collection settings is depicted in Table 2.1. Further data collection information is available in Appendix A.1. Between three and six scans were collected per crystal, depending on the overall size and orientation. The crystals were translated along the goniometer axis for each new scan.

2.3.2. Integration and processing

All frames were integrated using XDS software (Kabsch, 2010). Diffraction images obtained from the PI X06SA beamline with the Dectris Eiger 16M detector were first processed using the script H5ToXds provided by Dectris. The script enables XDS to read the Eiger detector frame format H5. Data integration was performed as described in Section 1.3.

Pilatus detectors are known to suffer from errors in count-rate correction for every strong pixels. This was corrected by flagging all reflections with intensities higher than one tenths of the maximum count rate as overloads. Diffraction images collected on Dectris Pilatus 6M detectors were processed with and without overload correction. The overload correction is evaluated only for images taken with the Pilatus detectors, since the Eiger detector is less effected by this problem.

For optimization, all data were re-processed as suggest by Diederichs (2016) and Mueller *et al.* (2012). The optimization procedure is described in Section 1.3.

2.3.3. Anomalous Signal

Special attention was extended to the anomalous signal of the data sets. At each step during the processing of all data different criteria were evaluated to maximize the anomalous signal.

During data processing, the anomalous signal was optimized to achieve high anomalous correlation (CC_{anom}) to high resolution. Akey *et al.* (2014) reported the positive influence of the strict absorption correction on the anomalous signal strength. Using this correction resulted in a significantly better anomalous signal and allowed structure solution with S-SAD (Akey *et al.*, 2014). Therefore, integration applying the correction was performed and compared to data sets processed without the correction.

2.3.4. Data averaging studies

Measurements of different crystals were selected and merged using various criteria. The goal was to achieve the best possible anomalous signal of the anomalous scatterer phosphorus. The maximization of the phosphorus signal should facilitate substructure solution via P-SAD. The output of the program ANODE was used as primary indicator of the strength of the anomalous signal. It evaluates the averaged anomalous density of a specified element and calculates the anomalous density at the corresponding atomic positions.

The individual ϕ-scans of all crystals were evaluated for their data quality and different quality indicators were selected to rank the data. All data sets were compiled into lists, which where sorted by different criteria such as ISa, mean I/σ(I), resolution, crystal, or beamline. Starting with the data sets ranked best in each category, successively more data was combined into a merged data set. Studies by Terwilliger *et al.* (2016) on merging from the best to the worst data set concluded that adding the worst data degraded the accuracy of the anomalous differences. Here, successively more data sets were combined until the overall quality of the merged data could not be improved upon.

The influence of the program used for the averaging of the reflections was evaluated further. In the interest of comparing the averaging capabilities of the development version SHELXC, the merged files were compared to those of other averaging programs.

2.4. Results and discussion

2.4.1. Evaluation of data quality

Influence of radiation damage

The unit cell parameters were evaluated for changes in their size to exclude radiation damage as source for systematic error. When crystals suffer from radiation damage, an increase in the unit cell dimensions can be noticed (Teng and Moffat, 2000). In Figures 2.2 and A.1 the cell axis for each crystal are plotted against the scan number. The unit cell axis should increase with the scan number if radiation damage is present.

The changes in the unit cell dimension are negligible for all crystals. The crystal C2 was measured at two different beamlines and displays the largest deviation over the range of all measurements. While some measurements show a slight increase in the unit cell parameters, the changes are below 1%. To further evaluate possible radiation

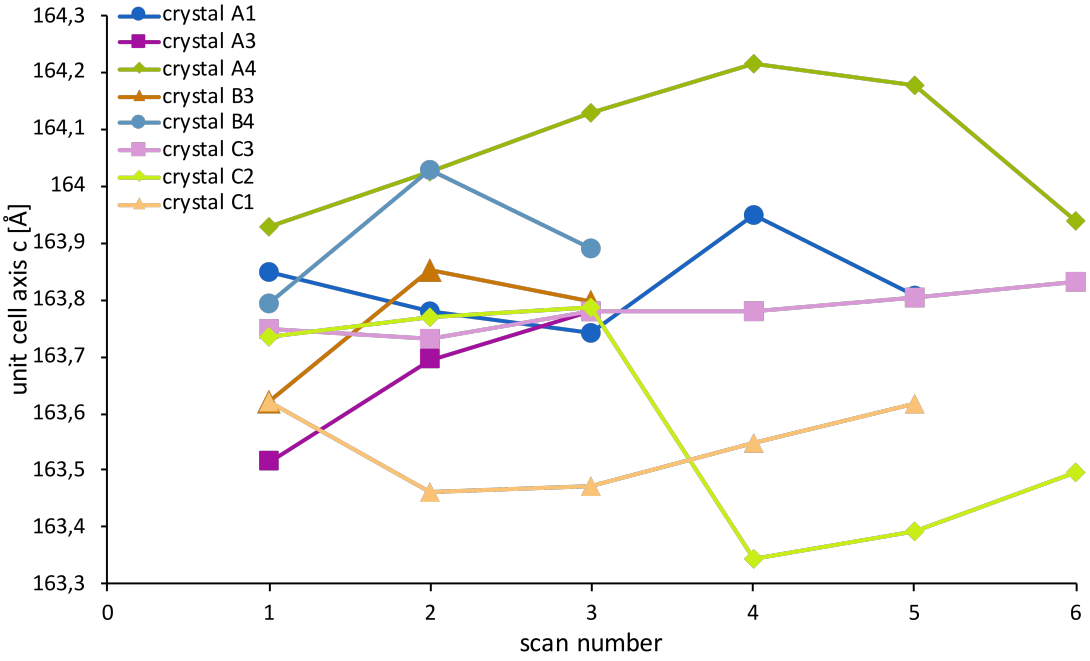


Figure 2.2.: Change of unit cell axis c in consecutive measurements by crystal.

damage, the χ^2 values were be reviewed (data not shown). Overall no radiation damage was found to influence the measurement.

Correction for overload

All measurements were studied for the influence of overloads on data quality. Data sets with and without possible overload pixels factored in were obtained during integration with XDS. The results were compared with a focus on the mean I/σ(I) values calculated by XRPEP, the anomalous signal indicators R_{anom} calculated by PHENIX, and average anomalous density of the phosphorus atoms calculated by ANODE. The results are presented in Figures 2.3, 2.4 and A.2.

For most measurements, the mean I/σ(I) values decrease when the integration is corrected for overload pixels (see Figure 2.3). The decrease in the mean I/σ(I) was the highest in data set 6. Without overload correction the mean I/σ(I) was 9.57 and decreased to 7.30 with the adjustment applied.

In the case of data set 2 the adjustment of the overload rescued the mean I/σ(I) value from 0.98 to 4.65. A similar increase in the mean I/σ(I) can be observed for measurement 8 (from 1.33 to 4.76). An increase can be due to a more accurate description of the

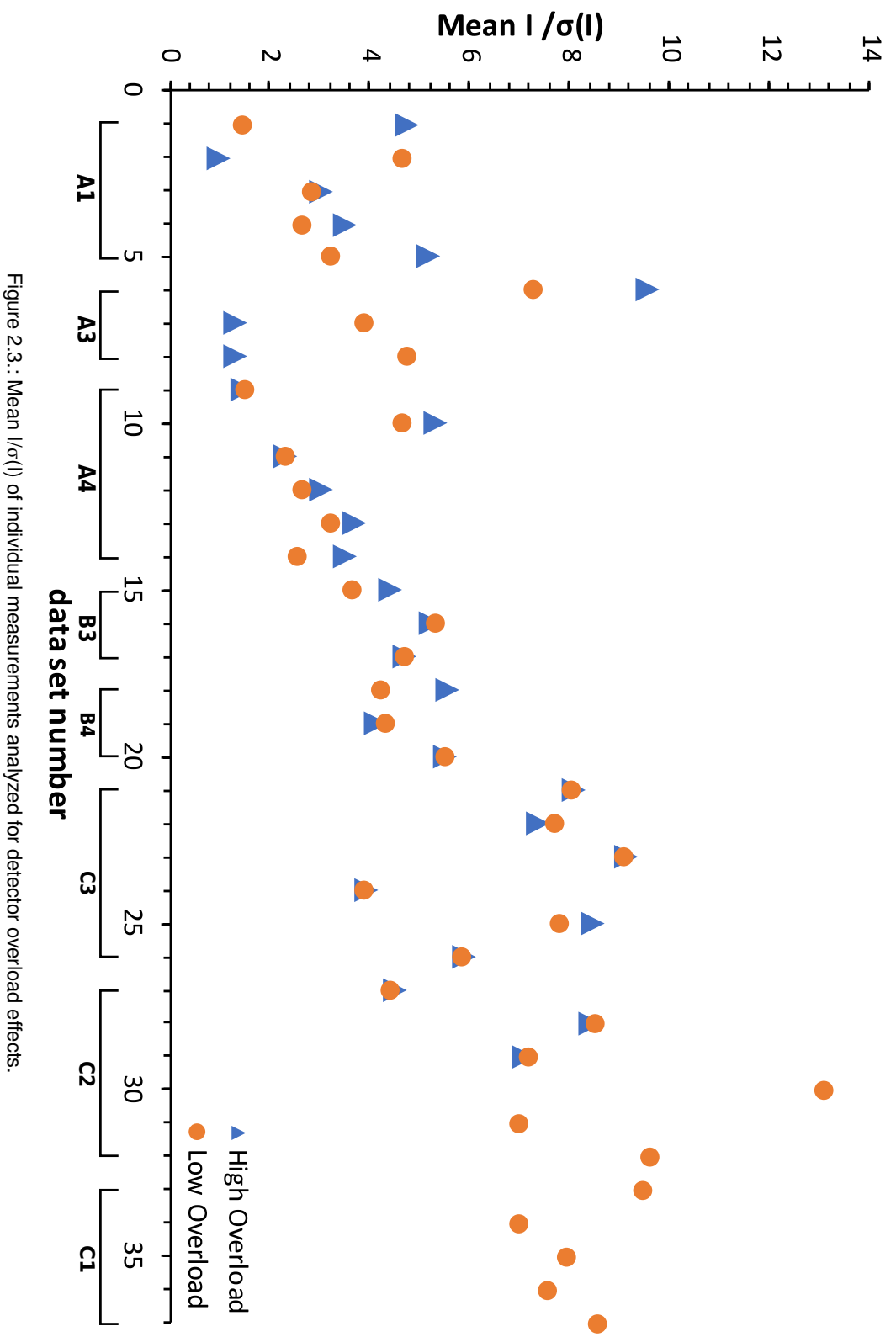


Figure 2.3.: Mean I/σ(I) of individual measurements analyzed for detector overload effects.

reflection profiles or more accurate error estimation.

Overall, the comparison reflects the general expectation that when the overload is considered the mean I/σ(I) can decrease. Only data sets experiencing strong reflections suffering from overload effects or of a similar intensity should be affected by the adjustment. A decrease in the mean I/σ(I) is not necessarily tantamount to a decrease in data quality.

More importantly, the average anomalous density at the position of the phosphorus atoms increased or remained unchanged in almost all data sets after applying the correction (Figure 2.4). For data set 2, 6 and 7 the change was most significant: Without the overload correction no anomalous signal was detectable at all phosphorus positions. When the correction was applied during data reduction, the averaged anomalous signal was not only detectable, but unexpectedly strong for a single measurement. The reverse is true for measurement 15. Here, a very weak averaged anomalous density for phosphorus was reported for the data without overload correction and no signal when the overload correction was applied.

Overall, the overload correction has a positive influence on the anomalous signal strength. For all integrations the overload correction was performed from this point forward.

Correction for absorption

The option to apply an absorption correction during data reduction is available in XDS. Akey *et al.* (2014) reported that the absorption correction influenced the anomalous signal strength. In their study the structure could only be solved via anomalous phasing after the absorption correction had been applied. Here, data sets integrated with and without strict absorption correction are compared.

The influence of the absorption correction was evaluated via the indicators ISa, CC_{anom} and d"/σ(d"). Figure 2.5 displays the CC_{anom} values for each data set at one specific resolution. A full table of the data and figures with the other indicators is available in Section A.1.3. For nearly all data sets the absorption correction influenced the anomalous signal significantly. A value of CC_{anom} of 10 or higher is considered significant (Schneider and Sheldrick, 2002). Ten data sets displayed values below this criteria before the absorption correction and only four afterwards. Overall, the employment of the absorption correction leads to higher values for CC_{anom}.

In conclusion, the absorption correction has a positive influence on the anomalous signal strength. The strict absorption correction was performed during data reduction for

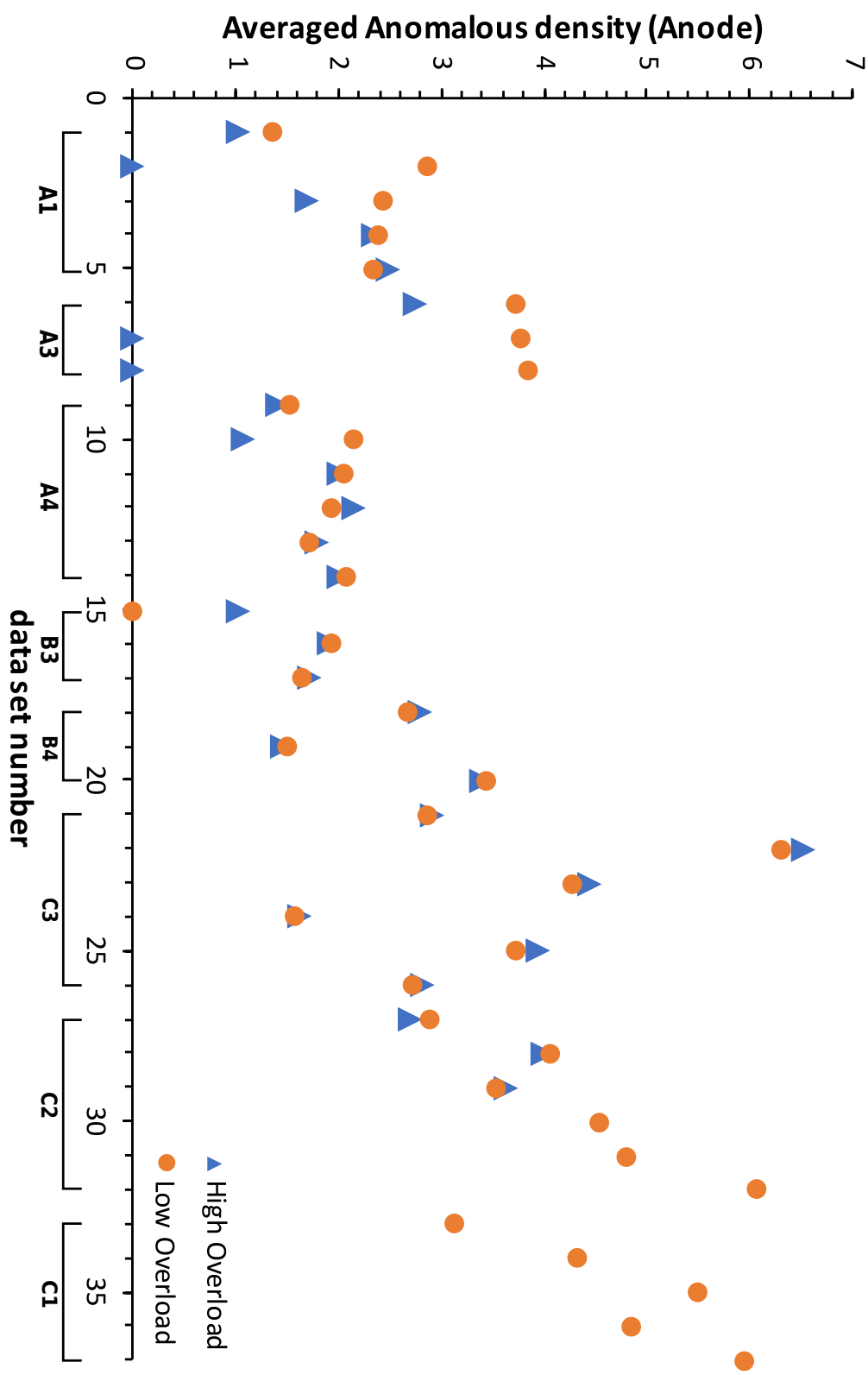


Figure 2.4.: Individual measurements of Poly(rA) were analyzed for detector overload effects on the anomalous signal strength.

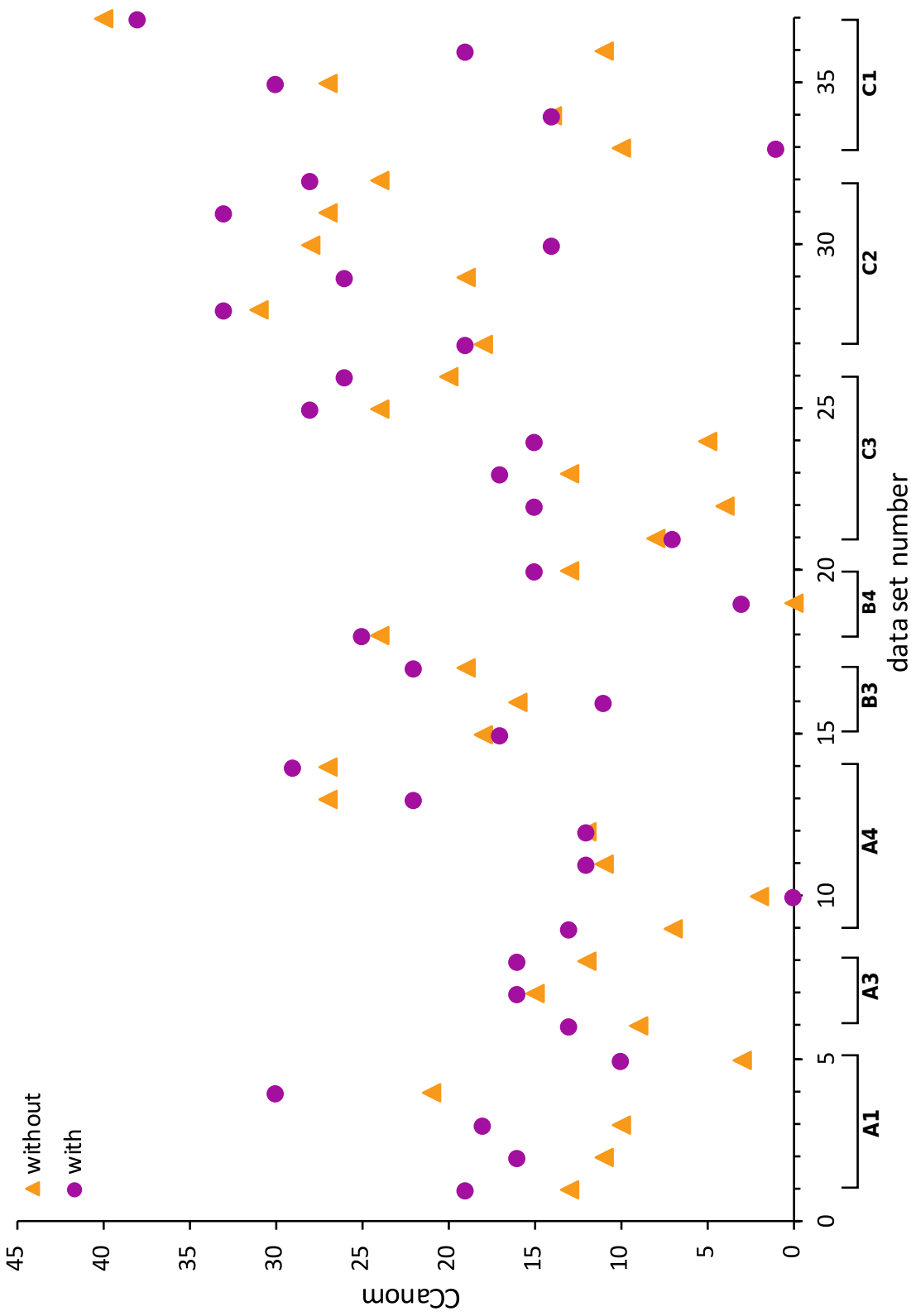


Figure 2.5.: Influence of the strict absorption correction on CC_{anom} during data reduction with XDS. 'with' labels all values which were acquired when the absorption correction was employed, 'without' refers to all values without absorption correction.

all further studies.

Comparison of data quality indicators

ISa One of the strongest indicators of systematic error is the asymptotic signal-to-noise (ISa) value of the unmerged intensity data. The ISa value should be above 30 for a good measurement (Diederichs, 2016). The ISa will decrease when systematic error compromise the measurement. ISa of all single measurements is plotted in Figure 2.6 as calculated by XDS.

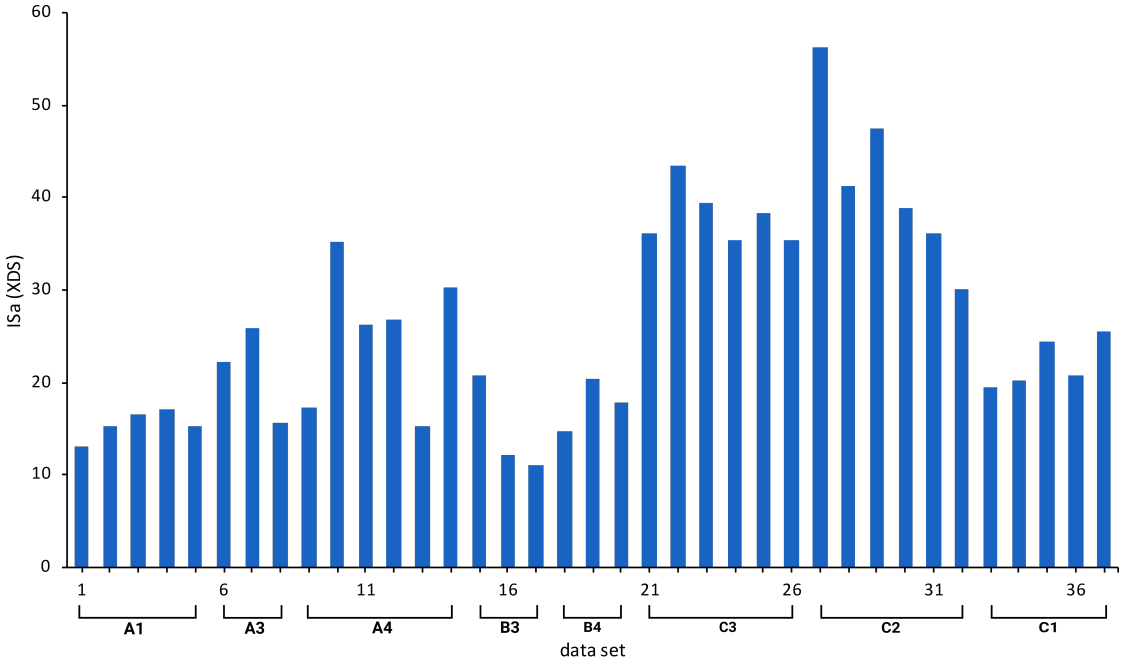


Figure 2.6.: The limit of the asymptotic signal-to-noise ratio (ISa) is plotted for the individual data sets. The ISa was calculated by XDS.

For nearly one third of the individual measurements the ISa is clearly above 30, indicating high data quality. From all measurements conducted at the PetraIII undulator beamline P11 only one single data set (#10) achieved an ISa above 30. On the other hand, data acquired from the crystals C3 and C4 display the highest ISa values. Contrary to expectation, the measurements collected with a Dectris Eiger detector, #30–37, were not generally better than those collected at a Dectris Pilatus detector (#1–28). While the crystals B3–B4 and C3–C2 (#27–29) were all collected at the same beamline, the individual measurements were made at two different times. The measurements #15–20 were collected earlier than the measurements #21–29. The large difference in the ISa

when comparing the two different collection dates might arise due to a difference in beamline specifics leading to different systematic errors.

In conclusion, the best and worst measurements when comparing solely the ISa were collected at the SLS undulator beamline PX10SA. In general, data sets from one crystal display similar ISa values.

Resolution limit Over the years, different limits for the integration or resolution cutoff have been suggested (see Chapter 1). Frames were in general integrated up to the edge of the detector or to the resolution at which $CC_{1/2}$ reached 30%. The different common cutoff criteria are compared in Figure 2.7.

The more conservative resolution limits of $R_{\text{meas}} \leq 0.8$ and mean I/σ(I) ≥ 2 suggest cutoff at higher resolution than $CC_{1/2} \geq 30\%$. For nearly all measurements the difference is ~ 0.1 Å. Several studies suggested that the inclusion of high resolution data beyond the conservative limits improve the phasing result and the model quality (Liu *et al.*, 2011, Wang, 2010). Therefore all frames utilized in this work were integrated to the limit when the quality indicator $CC_{1/2}$ reached 30%. It has been suggested (Karplus and Diederichs, 2015) to integrate all reflections even further, to a limit of $CC_{1/2} \geq 10\%$ even if this data is not used in structure solution or refinement.

Anomalous correlation coefficient CC_{anom} The anomalous correlation coefficient CC_{anom} is one of the most important indicators for the presence of an anomalous signal in unmerged reflections. XDS reports this indicator as 'Anom Corr' for the highest resolution bins. The resolution to which a significant anomalous correlation can be measured is a good indicator for the quality of the integration. One goal of the optimization of data reduction was therefore the maximization of the resolution to which CC_{anom} is greater than 10%. The resolution for this indicator for all measurements of Poly(rA) is displayed in Figure 2.8.

As discussed before, the quality indicator CC_{anom} is a guide to the strength of the anomalous signal. It is therefore one of the most important indicators whether the phasing using the anomalous signal can be successful. The resolution for phasing of nearly all datasets is better than 3 Å and lies for most data sets between 1.5–2.5 Å. This is a surprising result, considering that the XRD experiments were conducted at short wavelength not favorable for an anomalous signal from phosphorus.

Mean I/σ(I) Taking a closer look at the mean I/σ(I) values of the single measurements, the best values were obtained from the crystals C3 and C2. The differences between the

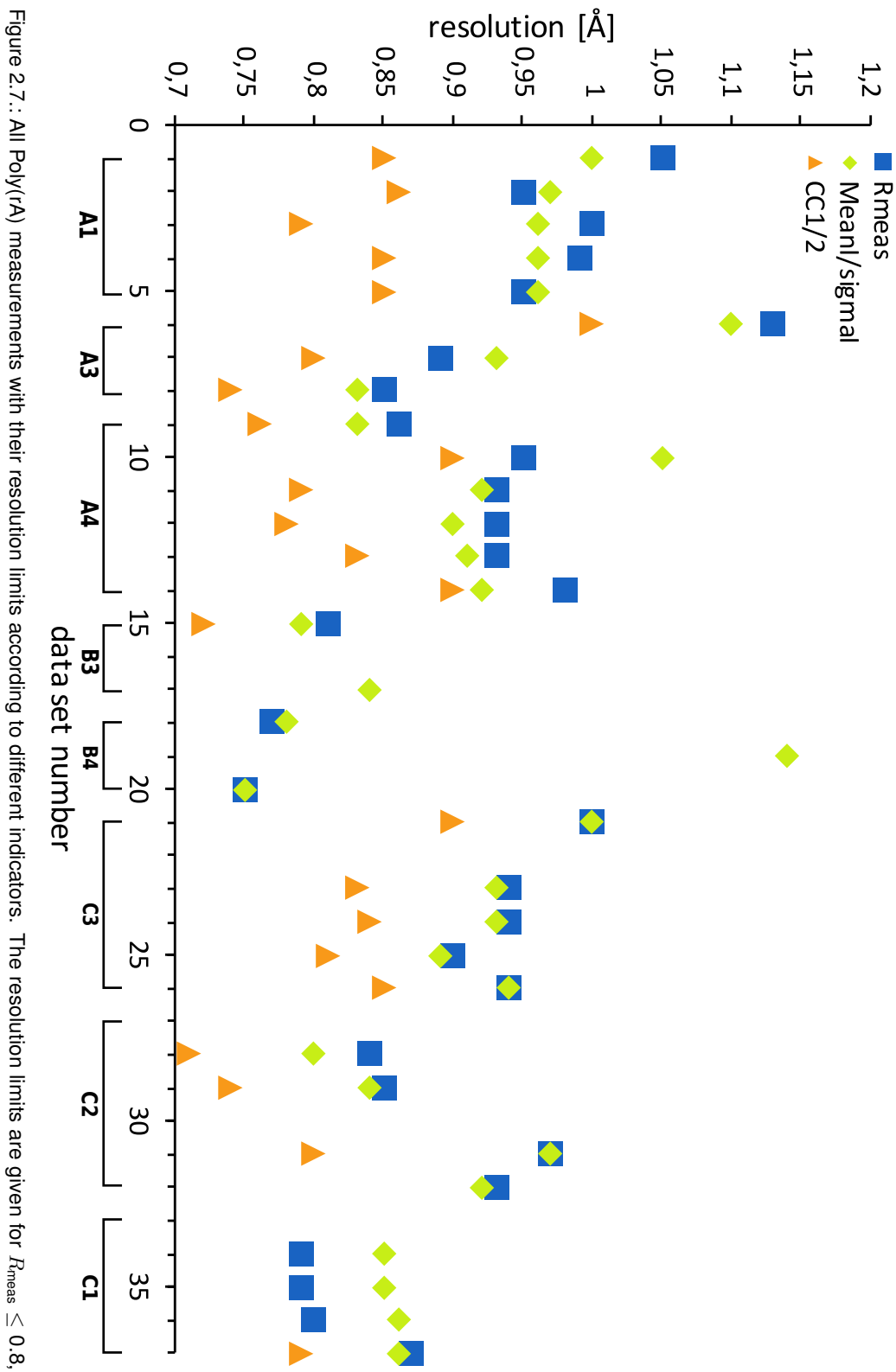


Figure 2.7.: All Poly(rA) measurements with their resolution limits according to different indicators. The resolution limits are given for $R_{meas} \leq 0.8$, $CC_{1/2} \geq 30\%$, and mean I/σ(I) ≥ 2. The indicators were calculated by XDS.

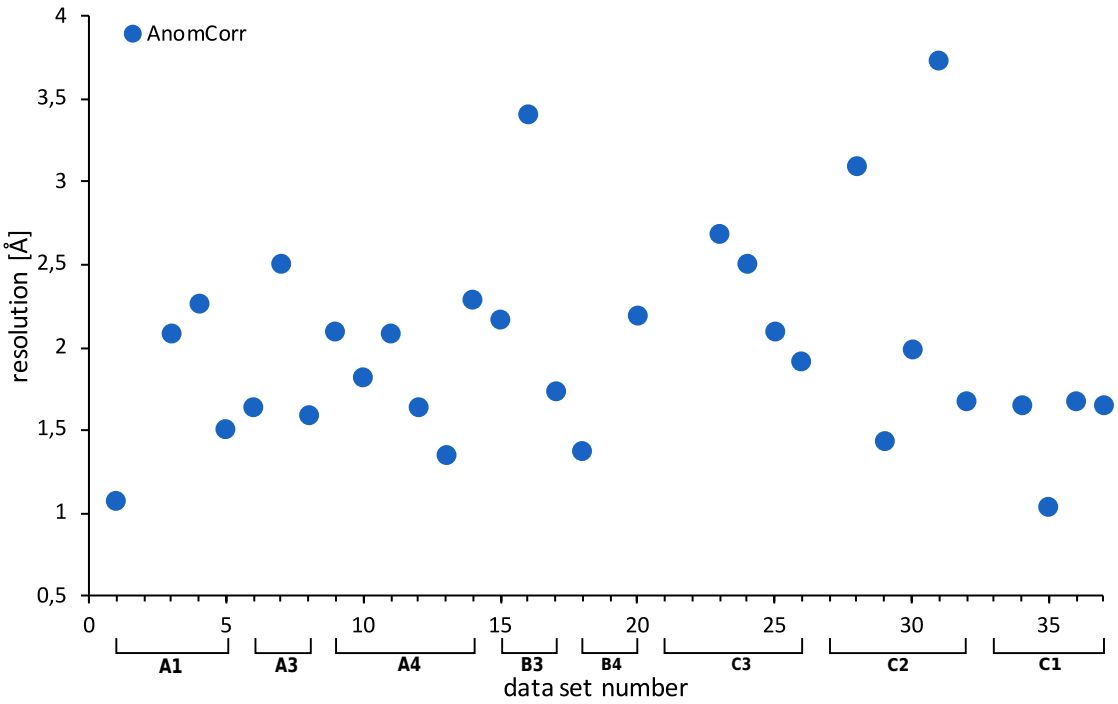


Figure 2.8.: All Poly(rA) measurements with their resolution limit for an anomalous correlation of at least 10%. The anomalous correlation was calculated by XDS.

single data sets are highlighted in Figure 2.9. The values are color-coded by beamline and the individual crystals are marked on the abscissa.

For the measurements from crystal C2 (#27–32) a direct comparison between the two detectors Pilatus 6M and Eiger 16M is possible. Interestingly, only one measurement (#30) from the Eiger 16 M detector shows a significantly higher mean I/σ. All other data sets collected at the Eiger 16 M detector show a mean I/σ comparable to the Pilatus 6M data sets.

2.4.2. Correlation between data quality indicators and anomalous signal strength

In advance of the averaging studies, the correlation between the quality indicators has been evaluated. The indicators mean I/σ(I) and ISa have been consulted for the overall quality of the data sets. Additionally the overall R_{anom} of each data set and the averaged anomalous density were examined. Only data quality indicators directly correlated with the anomalous signal strength should be considered when the further steps for structure solution by anomalous signal are planned.

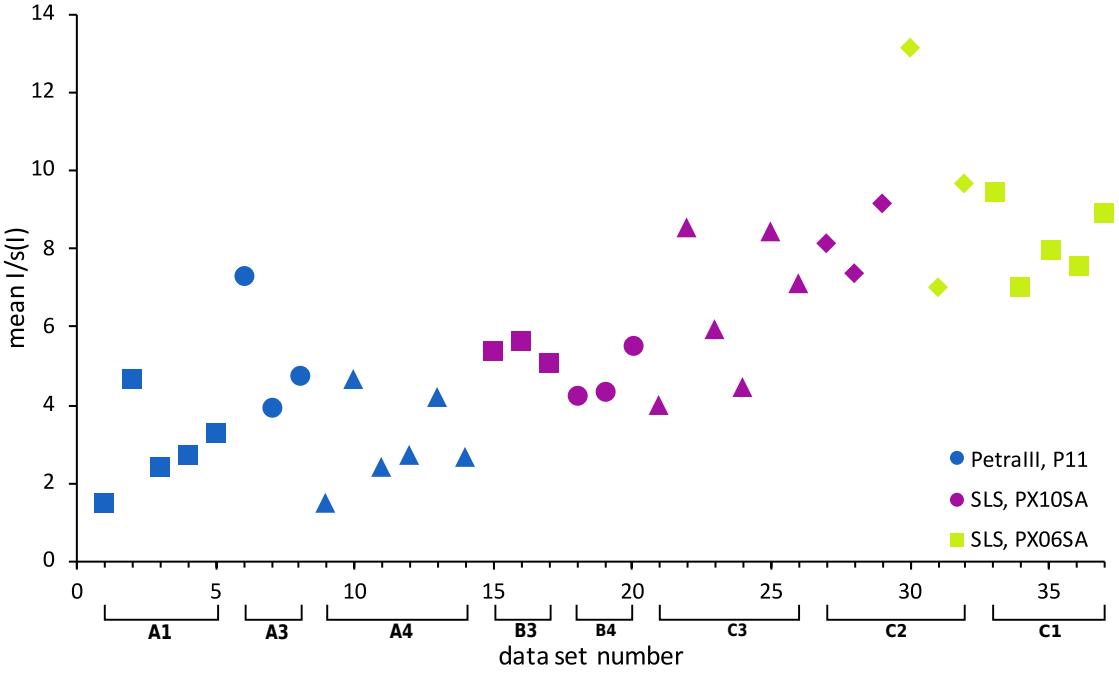


Figure 2.9.: Single data mean I/σ(I) analysis, coded by beamline and crystal.

At first the correlation between the limit of the asymptotic signal-to-noise ratio (ISa) and averaged anomalous density at the positions of the phosphorus atoms was analyzed (see Figure 2.10).

No correlation can be found between the ISa and the averaged anomalous density of the individual data sets. The indicator ISa is not applicable for the evaluation of the strength of the anomalous signal. Interestingly, Assmann *et al.* (2016) reported that the identification of outliers based on the ISa was not a reliable criteria for the rejection of data sets. This indicates that neither the quality of the data set nor the strength of the anomalous signal can be evaluated via the value of ISa.

It furthermore indicates that ISa can only be regarded as indicator for the presence of systematic errors in the processed data but not as deciding factor in consecutive data processing. When considering which data sets can be used for averaging to enhance the anomalous signal or the overall data quality, ISa should not be considered in isolation.

Next, the correlation between the mean I/σ(I) and the strength of the anomalous signal of the phosphorus atoms has been evaluated (see Figure 2.11). The averaged anomalous signal at the position of the phosphorus atoms is stronger when the mean I/σ(I) is higher. The data as presented in Figure 2.11 shows a clear correlation of both values. This relationship is further confirmed in the correlation between the mean I/σ(I)

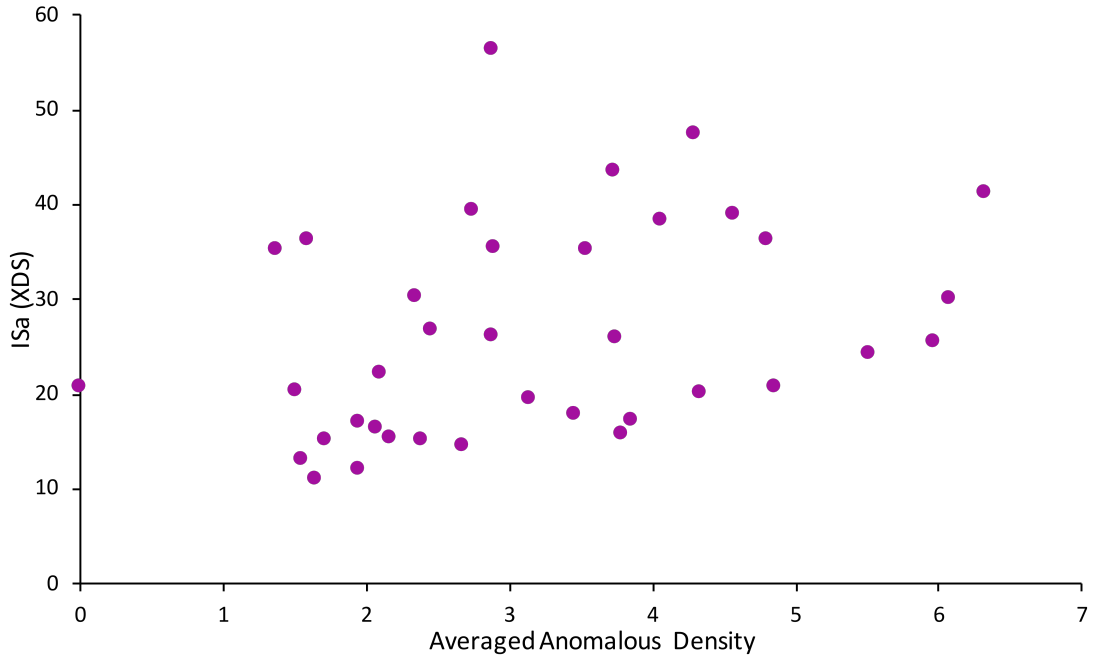


Figure 2.10.: The limit of the asymptotic signal-to-noise ratio (ISa) against the averaged anomalous density is plotted. The ISa was calculated for each data set by XDS and the averaged anomalous density of the phosphorus atom positions was calculated by ANODE.

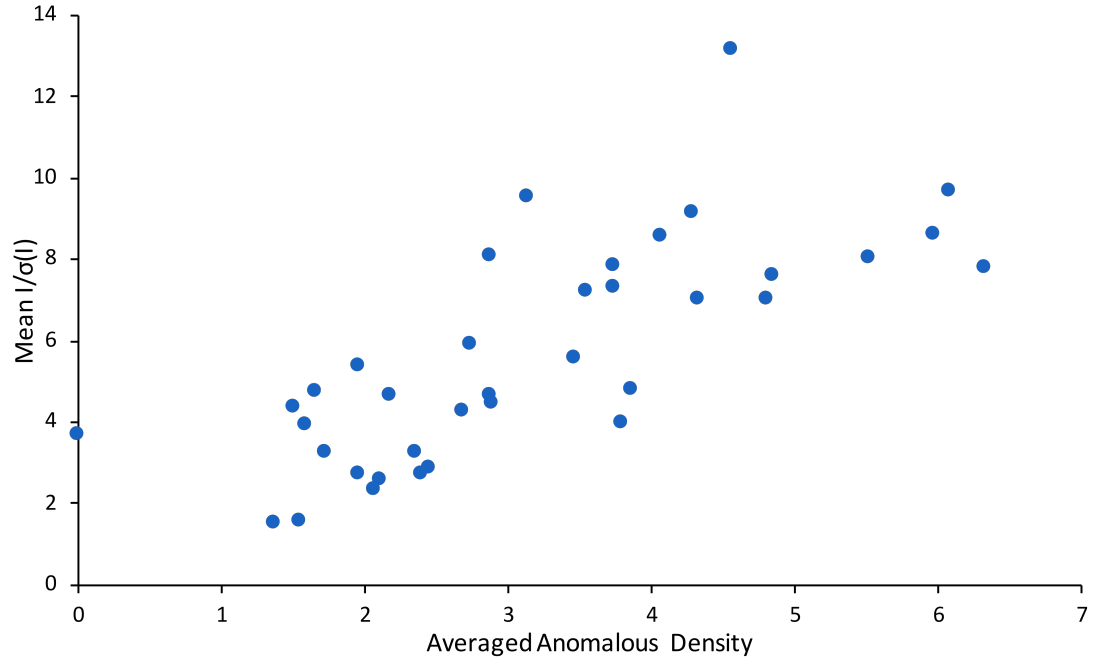


Figure 2.11.: The mean I/σ(I) is plotted against the averaged anomalous density. The mean I/σ(I) was calculated for each data set by XPREP and the averaged anomalous density of the phosphorus atom positions was calculated by ANODE.

and R_{anom} (see Figure A.4). R_{anom} decreases with increasing mean I/σ(I). Based on the direct correlations found, mean I/σ(I) can be an indicator of the anomalous signal strength. Additional correlation plots of the indicators mean I/σ(I), ISa, R_{anom} and the averaged anomalous density are displayed in Appendix Section A.1.4 (Figures A.4 – A.7).

In conclusion, a clear correlation can be found between the data quality indicator mean I/σ(I) and the strength of the anomalous signal. The correlation can be considered as criteria when selecting data sets for averaging studies.

2.4.3. Multi crystal averaging

Selection criteria for multi crystal averaging were evaluated. The main focus is the enhancement of the anomalous signal of the intrinsic scatterer phosphorus. The anomalous signal was evaluated at the atomic position of the phosphorus atoms with the program ANODE. The averaged anomalous density for all phosphorus positions was compared.

The first selection criteria was the crystal itself. All data sets collected from one crystal were averaged and evaluated. The consistency of these data sets is expected to be high and the result can reflect the quality of the crystal.

The second selection criteria was the beamline at which the XRD experiments were conducted. The XRD experiments were conducted at the synchrotron PetraIII undulator beamline P11 (crystals A1, A3, A4) and at two different undulator beamlines at the SLS, PX10SA and PX06SA. The data sets from crystals B3 and B4 as well as the data sets from crystal C3 and C2 (scan 4–6) were collected at PX10SA, but not during the same time slot. Those data sets are grouped as SLS2015 and SLS2016, respectively. Finally the data sets from crystal C2 (scan 1–3) and C1 were collected at PX06SA. The separation of the data sets by beamline also discriminates them by the detector that was used for data collection (see Section 2.3.1).

The final selection criteria was the mean I/σ(I). All data sets are ranked from the highest to the lowest mean I/σ(I) and consecutively merged. Starting with the five highest ranked data sets, more data sets were added until no improvement of the anomalous signal strength was measurable. ISa as criteria for multi crystal averaging was excluded based on the report by Diederichs (2009) and Huang *et al.* (2015) and the results obtained in this work.

Data sets were averaged using the programs PHENIX_scale_and_merge (Adams *et al.*, 2010), XPREP[2] and XSCALE (Kabsch, 2010). These programs were used to

[2] *G.M. Sheldrick*, Bruker AXS Inc., Madison, Wisconsin, USA, 2003.

compare the quality of the averaged files with the development version of SHELXC. The results are presented in Figure 2.12 and Table A.5.

Averaging by crystal Three to six data sets – or scans – were collected per crystal, depending on its size and orientation. The data sets collected from one crystal should display low anisotropy, especially when radiation damage can be excluded (Sygusch and Allaire, 1988).

The anomalous signal strength described by the averaged anomalous density for the data sets averaged by crystal is displayed in Figure 2.12, the first eight values from the left. The averaged data set from crystals A1 and A3 display very little anomalous signal at the position of the phosphorus atoms. Crystals C2 and C1 show a strong anomalous signal. The strength of the anomalous signal is depended on the program for crystal C2. The data sets processed with XPREP performs inferior to those averaged with XSCALE and especially PHENIX. For all other averaged data sets the results from different merging programs are more similar in quality.

In conclusion, except for the outlier computed for crystal C2, the different programs are treating the anomalous signal quite similar. Averaging data sets from one crystal can result in a stronger anomalous signal, but the increase is usually minor.

Averaging by beamline Six to 23 data sets were averaged for the analysis of the beamline and detector influence. The resulting averaged anomalous density is plotted in Figure 2.12, the values named by beamline. The best results were obtained from merging all data collected at the SLS beamlines, a total of 23 data sets. Data sets averaged by PHENIX clearly display a stronger anomalous signal for the data measured at PX06SA and PetraIII. The influence of the program is visible quite distinctly for these data set combinations.

The averaged file for all measurements collected with the Dectris Eiger detector (SLS2016-PX06SA) displays an anomalous signal strength comparable to the measurements from beamline PX10SA (SLS2016-PX10SA and SLS2015-PX10SA). But the quality of the anomalous signal is depended on the program used to average the data sets.

The strongest anomalous signal was obtained when all data sets collected at SLS beamlines were averaged (All SLS). This averaged data set contains the averaged reflections collected in 23 scans from five crystals in total. Here the choice of the program is not as critical, all programs generate files resulting in an averaged anomalous signal of above 9.

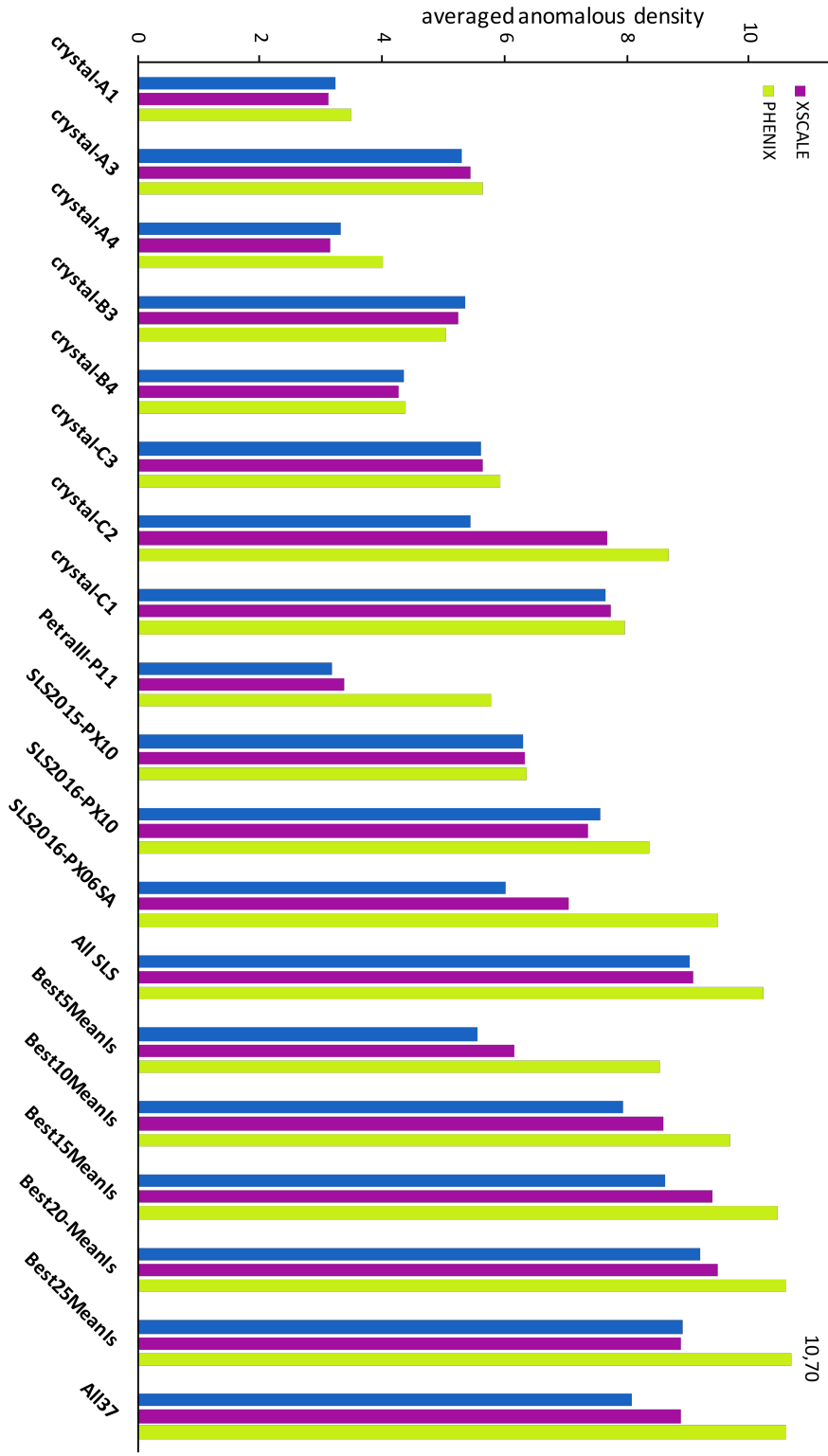


Figure 2.12.: The scans are averaged by crystal, beamline, or best mean I/σ(I). The averaged datasets contain three to all scans. The averaged anomalous density was calculated by the program ANODE. The programs PHENIX_scale_and_merge. XPREP, and XSCALE were used to merge the scans.

Averaging by mean I/σ(I) All data sets were ranked by the quality indicator mean I/σ(I). The data sets were merged consecutively from the highest to the lowest mean I/σ(I) starting with the five best.

A selection of the results are displayed on the right in Figure 2.13. In this evaluation, the merged data sets obtained with PHENIX result in data sets with a stronger anomalous signal. Especially for the data set containing only the best five scans, the averaged anomalous density is above 8 when using PHENIX but below 7 for XPREP and XSCALE.

The best overall result was achieved with the best 25 data sets merged with PHENIX. An averaged anomalous density of the phosphorus atoms of 10.7 was obtained. When more data sets were added, the averaged anomalous density decreased slightly. The programs XSCALE and XPREP presented their highest averaged anomalous density when 20 data sets were merged. Again, the anomalous signal strength decreased when more data sets were added.

The selective exclusion of weak data resulted in a better anomalous signal strength. This result was reported by Akey *et al.* (2014) before. The mean I/σ(I) as best qualifier for the anomalous signal was proposed by Dauter and Adamiak (2001). This study achieved the best overall anomalous signal strength with the mean I/σ(I) as exclusion criteria.

Influence of data redundancy The influence of the number of data sets averaged in the final file is evaluated. A plot of the number of scans per merged file is presented in Figure 2.13.

The number of available measurements of individual intensities stand in direct correlation with the accuracy of the resulting mean intensity (Weiss, 2001). With higher redundancy the Bijvoet pairs and therefore the anomalous signal should be better estimated. This trend is reflected in the present study as well. With a small number of contributing data sets to the merged data sets, a lower averaged anomalous density is recorded. The averaged anomalous density increases with increasing number of averaged data sets. When a certain signal strength is reached, the addition of more data sets does not improve the averaged anomalous density. This has been reported before by (Liu *et al.*, 2012) and is confirmed by this study.

Interestingly, one outlier is present in the evaluation. For the averaged file containing 14 data sets the averaged anomalous density is, independent from the program used, below the expected value given by the trend. This data set contains all measurements collected from PetraIII, P11. Possible reasons for this outlier can be systematic errors specific to the undulator beamline setup or poor quality of the crystals used for XRD.

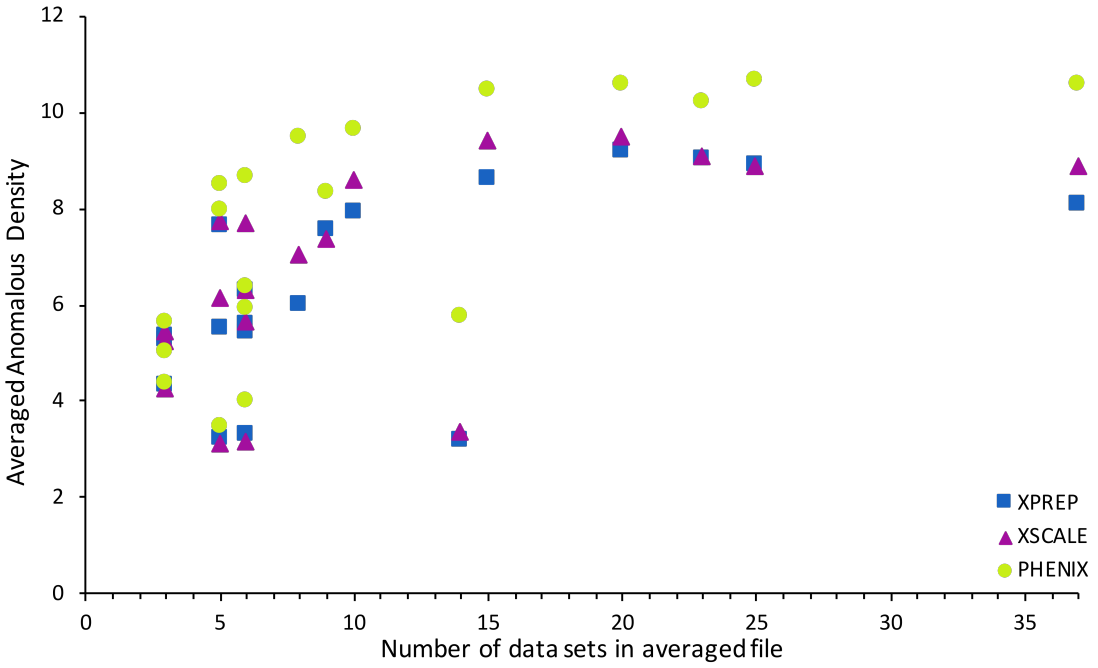


Figure 2.13.: The number of averaged data sets is plotted against the averaged anomalous density. The data sets were selected by crystal, beamline or the highest value of mean I/σ(I).

SHELXC development version The development version of SHELXC capable of multi crystal averaging could not be tested to the full extend in the course of this work. Therefore all preliminary results will be reported at a later time. Based on the down-weighting of outlier data sets, SHELXC will average all available data. The first results obtained for all 37 data sets were promising. The averaged anomalous density calculated with ANODE was in the same range with that of the other programs tested.

2.5. Conclusion and Outlook

In this work, the influence of data processing on data quality and the anomalous signal strength is evaluated.

The strict absorption correction and the optimization of the mean I/σ(I) during data reduction provides data sets of the best quality. The influence of data processing was discussed before (Dauter and Adamiak, 2001, Karplus and Diederichs, 2015, Liu *et al.*, 2012, Wang, 2010). While ISa can be used to detect the influence of systematic errors on the data, it is not the best indicator for the overall quality of a data set.

Furthermore, the program used to average multiple data sets can have a significant

influence on the anomalous signal strength. Since only one set of measurements was used in this study, a general recommendation for one single program cannot not be given, but the results presented here suggest that a selection of merging programs should be tested in borderline cases.

The exclusion of data sets of inferior quality can improve the anomalous signal strength. Hereby the quality indicator mean I/σ(I) can play a significant role and other indicators such as d"/σ(d") or R_{anom} should be tested as well. The concept of multi crystal averaging in SHELXC is based on simple down-weighting outliers. Further studies on the detection of outliers can be conducted. Assmann *et al.* (2016) suggested the identification of rogue data sets based on calculating $CC_{1/2}$ between all data sets. As a result, the data sets could be clustered by their correlation (Diederichs, 2017).

The best averaged data sets displayed an averaged anomalous density of 10.7 for all phosphorus atoms. Although the value was not ideal, P-SAD with the anomalous signal was attempted, but gave no substructure solution with SHELXD. Nonetheless, significant anomalous signal is present and the improvement of multi crystal averaging strategies could lead to a structure solution via the anomalous signal in the future. This could be a compelling argument to use intrinsic anomalous scatterers in macromolecular crystallography, especially for the structure solution of (deoxy)ribonucleic acid structures (Harp *et al.*, 2016).

Overall, the importance of data reduction and processing and the influence on the quality of the anomalous signal strength could be demonstrated. The conclusions drawn from this work were applied for the structure solution of cln5 (see Chapter 4).

3. PDB2INS

PDB2INS

Enabling the refinement of high resolution macromolecular data with the program SHELXL (Sheldrick, 2015) has been an enduring desire for more than two decades now (Sheldrick and Schneider, 1997). With the continued development and availability of more brilliant and better-focused X-ray radiation sources (Duke and Johnson, 2010) as well as more sensitive detector technology (Casanas *et al.*, 2016, Leonarski *et al.*, 2018), the number of macromolecular structures with a resolution of 1.7 Å or better is steadily increasing.

The protein database (PDB) supports *mmcif* and *pdb* file format and all deposited data are freely available. The most common refinement programs adhere to this standard and present the resulting files in these formats as well. Access to SHELXL refinement is limited by the necessity to convert from the most commonly used formats in macromolecular crystallography to the standard in small molecule refinement – *ins* (refinement instructions and atomic coordinates) and *hkl* (reflection data) format. To bridge this gap, the program PDB2INS was developed.

3.1. Background

The PDB archive (wwPDB) is a depository for 3D structural information of macromolecules, such as proteins, nucleic acids and complex assemblies (Berman *et al.*, 2003). Today, the wwPDB consists of four members, the Biological Magnetic Resonance Data Bank (BMRB), the Protein Data Bank Japan (PDBj), the Protein Data Bank Europe (PDBe), and the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB). It has become mandatory for publishing in relevant papers to deposit macromolecular X-ray diffraction (XRD) models and reflection data at the web service provided by the wwPDB (Wlodawer, 2007).

3.1.1. Macromolecular refinement programs

Several refinement programs are available for macromolecular refinement. Some are highly specialized for a specific problem while others can be applied to a broader range of structures. A number of factors influence the choice of refinement program, such as resolution or required parameterization. A more detailed discussion is available by Tronrud (2007) and Shabalin *et al.* (2018).

Some popular refinement programs, such as REFMAC5 (Murshudov *et al.*, 2011) or PHENIX (Adams *et al.*, 2010), are very versatile and offer a broad range of refinement options (see Figure 3.1). The standard repertoire includes least squares, maximum likelihood and phase maximum likelihood target functions with sparse matrix gradient method as minimization algorithm. Mostly coordinate refinement with Konnert-Hendrickson-type restraints and torsion-libration-screw (TLS) parameterization of anisotropy are included for refinement configuration. Special minimization algorithms, refinement target functions or restraint parameterization are implemented as well or are available in other programs. Generalized comparisons of available programs are problematic, since the implementation of functions, parameterizations, and optimizations methods can vary significantly from one program to the next.

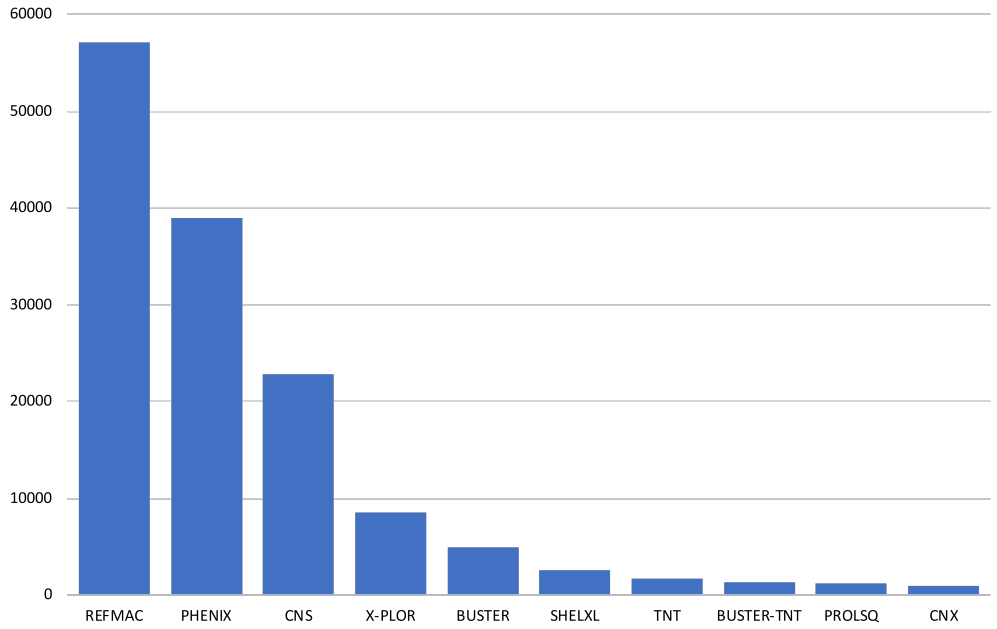


Figure 3.1.: The most popular refinement programs against total numbers of deposited structures in the PDB in March 2019.

Depending on the resolution or stage of refinement, some programs or software suites may offer an advantage. For a refinement with a low resolution data set (3.0 Å or lower) or with a poor starting model, X-PLOR (Brünger, 1992), CNS (Brunger, 2007, Brünger *et al.*, 1998) or PHENIX are more suitable in general. These structures may benefit from simulated-annealing molecular dynamics to increase the convergence radius or a lower parameter count via torsion angle refinement. For structures in the resolution range from 3.0 Å up to 1.4 Å, into which most of the published structures fall, restrained coordinate refinement against maximum likelihood target functions usually yields good results. These are available in most programs including REFMAC5, TNT (Tronrud, 1997, Tronrud *et al.*, 1987), CNS or PHENIX. Atomic resolution data can benefit from a customizable restraint model or the refinement against intensities with a blocked least-squares algorithm as provided by SHELXL.

3.1.2. Advantages of a refinement with SHELXL

Refinement with SHELXL is particularly useful if high resolution data is available. The program allows refinement against neutron and Laue diffraction data (Gruene *et al.*, 2013). When the data-to-parameter ratio is sufficiently high, anisotropic refinement is possible. The refinement of multi-domain twinned data, including non-merohedral twinning, is facilitated. This is beneficial since all data can be used for refinement instead of only data of a single domain.

SHEXL enables the refinement against intensities and allows the refinement of atomic site occupancies. Also complicated disorder can be treated flexibly and refined, including constrained determination of occupancies. For small structures the least-squares estimation of individual standard uncertainties in possible. The calculation of $R_{complete}$ (Luebben and Gruene, 2015) in addition to R_{free} (Brünger, 1992) is implemented, providing access to state of the art structure validation techniques.

One example of the advantages offered by SHELXL refinement of macromolecules was published by Köpfer *et al.* (2014). Various websites and tutorials are available explaining the use of SHELXL with macromolecular data, such as the CCP4 wiki[1], the SHELXL homepage[2] and chapters in the SHELXL manual or in Mueller *et al.* (2006).

[1] K. Schäfer and K. Diederichs, *strucbio.biologie.uni-konstanz.de/ccp4wiki/index.php/SHELXL*, University of Konstanz.

[2] G. M. Sheldrick, *shelx.uni-goettingen.de*, tutorials, talks, open access papers, and overview of all SHELX commands, University of Goettingen.

3.1.3. File transformation to SHELXL formats

The first interface to enable easy access to SHELXL for macromolecular data was SHELXPRO, written by Sheldrick and Schneider (1997). Designed to be self-explanatory and easy-to-use, it was distributed for the first time with the SHELX-97 release of the SHELX program suite. SHELXPRO reads a number of file formats, can generate interatomic distance and bond angle restraints for essential amino acids, can display a variety of plots for analytical purposes, and perform anisotropic scaling (Shakked, 1983). It is also used to create a basic instruction (*ins*) file or to convert from the SHELXL output file back to the PDB format file.

Similar features were later implemented and more conveniently accessed by other programs like COOT (Emsley *et al.*, 2010), XPREP[3], or SHELXL itself.

COOT is capable of reading and writing SHELXL result files to display the model and electron density maps. Other tasks were implemented in XPREP. It is capable of R_{free} flag creation, plot generation, and displaying various data tables. Several features found their way as instructions into SHELXL, such as the generation of *pdb* files[4].

While SHELXPRO itself became partly redundant, some functions – for instance converting to the SHELXL compatible format – remained useful nonetheless. Generating *ins* (instruction) files from PDB files while adding sensible restraint commands and renumbering residues to fit SHELXL nomenclature, for example. To address this requirement, PDB2INS was created as a successor.

PDB2INS was designed to extend the function of converting from macromolecular file formats to the SHELXL compatible formats. The main focus is hereby the creation of the instruction file which is controls the refinement and contains the atomic coordinates. Additional options were added as well as more choice for customizing data processing. The aim is to allow easy access to refinement of macromolecular data with SHELXL for inexperienced users while at the same time grant more directed control for advanced users. In the end, minimal to no additional editing should be necessary before running SHELXL with the created instruction file.

Obtaining the reflection file in SHELXL format from an *mtz* file was possible by the programs MTZ2SCA and SCA2HKL by Grune (2008). The tool CIF2MTZ by Winn *et al.* (2011) could be used to create an *mtz* file. XPREP facilitates the conversion of multiple input formats[5] to *hkl* files as well. For convenience, the transformation from the *mmcif*

[3] G. M. Sheldrick, Bruker AXS Inc., Madison, USA, 1997.

[4] The instruction WPDB was added to SHELXL prompting the program to write a *pdb* file directly.

[5] An overview of all reflection data containing input files for XPREP is given when the program is started without any filename on the command line. The following file formats are read: *sca* (SCALEPACK or

file format used by the PDB to the *hkl* format was introduced into PDB2INS as well. This should limit the need for other conversion programs and allows direct access to all data available in the PDB.

3.1.4. SHELXL file format aspects to consider

Various limitations within the SHELXL instruction format are present and have to be taken into account when converting PDB files to SHELXL instruction files. The *instruction* file contains lines of maximal 80 characters, starting with four letter keywords followed by parameters. The keywords can contain essential information regarding the structure and regulate the refinement which can be fine-tuned with specific parameters.

The *instruction* file can be segmented into three main parts. The first part contains essential keywords which have to appear in a defined succession providing the most basic data such as the cell, wavelength, symmetry, or elements present and their corresponding scattering factors. Also, the type of refinement needs to be specified.

Next, additional information can be provided to guide and improve the refinement. Parameters to allow a refinement of twinned data, keywords regulating the amount of output files and statistics as well as specific refinement restrictions can be defined. Additionally, restraints and constraints can be placed here as well as commands to add riding atoms such as hydrogen.

The final part consists of the atoms present in the asymmetric unit, one line per atom, and starts with the atom name instead of a fixed keyword in contrast to all other lines. The atom name can not exceed four characters and needs to be unique. This stands in stark contrast to macromolecular crystallography guidelines, were atoms of one specific residue are expected to follow a strict repetitive naming scheme. SHELXL allows the combination of a residue number with an atom name leading then to a unique identifier for each atom.

SHELXL restricts the residue number to four characters, therefore limiting the size of macromolecules that can be refined. This presented a particular challenge, since SHELXL was capable to address residues by name and number but not chain identifiers. Accordingly, a program writing an instruction file for SHELXL from a regular PDB file must convert the chain identifiers into a residue number. Often one chain only contains a few hundred amino acids and a new chain could simply be distinguished by adding a multiple of 1000 to the residue number. A new chain would then be identified by the size of the gap in between to consecutive residue numbers in the file. Such a work-around

HKL2000 format), *HKL* (XDS_ASCII.HKL file), *fco* (written by XD), and SHELX file formats.

makes it harder to compare and reference the original file with the files produced after the SHELXL refinement and still limits the size of the molecule that can be processed by the program. In SHELX format one chain can contain up to 9999 atoms when the numbering starts with 0001[6]. Today, only a single character for the chain identifier is allowed and the capitalization of letters is discriminated. In the end, the file is terminated with an instruction specifying the format of the associated reflection file.

3.2. Aim of this work

The small molecule refinement program SHELXL can be used to refine macromolecular XRD data of high resolution. SHELXL offers multiple advantages not implemented in the most commonly used refinement programs in macromolecular crystallography. To facilitate a refinement of suitable data with SHELXL, a program to set up the necessary files in the correct format is needed.

The aim of this work is to create a program, PDB2INS, to set up all refinement files for a SHELXL refinement from *pdb* files. PDB2INS is constructed as simple interface between macromolecular files and a refinement with SHELXL. The conversion is handled in a single program with an optional graphical user interface. Novice and expert users alike can use PDB2INS to create all files needed for SHELXL. PDB2INS allows many options to customize the data and adds instructions and restraints to use SHELXL to its full potential.

3.3. Methods and Implementation

PDB2INS is available as packaged version for Windows, Linux and MacOSX from the homepage of *G. M. Sheldrick*[7]. The source code, command line versions as well as the graphical user interface (GUI) versions are available at *https://github.com/av-luebben*. Some distributions of CCP4i/CCP4i2 also contain the command line version of PDB2INS.

[6] Additionally, negative residue numbers up -999 can be used.
[7] G. M. Sheldrick, *http://shelx.uni-goettingen.de*.

3.3.1. Programs and resources

PDB2INS was written using Python 2 programing language[8]. The integrated development environment (IDE) PyCharm (JetBrains, Prague, Czech Republic) as free, academic edition was employed and the python interpreter PyInstaller[9] was used to generate stand-alone executables. All version control and publication of the code was done using git [10] and the free academic GitHub professional version (GitHub Inc, San Fransisco, USA), respectively. The graphical user interface of PDB2INS was written using TKinter (Tcl/Tk), the standard graphical user interface package of python.

PDB2INS was designed to be called from a command line and can be integrated into automated scripts. The program was tested by selected test users and using an automated script that included downloading files from the PDB website[11] or using a pre-downloaded part of the PDB. Furthermore, SHELXL was called after a successful PDB2INS run to perform a refinement with the created files. The program COOT (Emsley *et al.*, 2010) was used to visually inspect the refinement result and directly modify selected regions.

3.3.2. Data formats

Protein Database formats

The last major changes to the *pdb* file format were made in 2008 resulting in Version 3.20 (Dutta *et al.*, 2009). Since then it became mandatory not only to deposit atomic coordinates but also the corresponding experimental diffraction data. PDB2INS is accepting all *pdb* and *sf-cif* files in this format.

A *pdb* file is divided into nine major sections containing the atom coordinates in addition to information about the source of the macromolecule, programs used during data processing, the authors, and secondary structure features. The file is in a specific format, each line consists of 80 characters with the first six columns indicating the record name.

Each record name has a distinct format for the following columns and is usually terminated with an end-of-line character. All records have to appear in a fixed order and mandatory record types are present in all PDB entries (see Section B.1).

[8] Python is an multi-paradigm, interpreted programming language that is freely available at *www.python.org*, Python Software Foundation, Python Language Reference, Version 2.7.

[9] PyInstaller Development Team, *http://www.pyinstaller.org/*, initiated by Giovanni Bajo.

[10] Junio Hamano, *http://git-scm.com*, initiated by Linus Torvalds.

[11] *www.rcsb.org*, member of wwPDB and EMDataResource, National Science Foundation.

Especially the coordinate section is of interest, containing all atoms and their associated information such as name, B-factor, occupancy and coordinates.

SHELX file formats

SHELXL expects two files, the *instruction* file with the suffix *.ins* and the *structure factor* file with the ending *.hkl* as input. The later file uses a fixed format.

The *hkl* file consists of lines of up to 80 characters with each line containing the data for one hkl index and starts without preamble. While the hkl indices take up four characters each (H, K, and L, right justified), the associated reflection (R) - either intensities or amplitudes - and their estimated standard deviation (S) hold eight characters each. Afterwards, four more characters are reserved for a flag (F), such as R_{free} flags or domain association, resulting in a line structure such as:

$$HHHHKKKKLLLLRRRRRRRRSSSSSSSSFFFF \tag{3.1}$$

The last line in the *ins* file will contain a specification which kind of data are present in the reflection file. The instruction *HKLF 4* stands for reflections in the form of intensities while *HKLF 3* specifies amplitudes, for example.

The *ins* file is more elaborate, each line either starting with a four character instruction followed by additional parameters in free format. A list of possible instructions and a brief summary on their use is given on the homepage of *G. M. Sheldrick*[12]. In the context of this thesis, the *ins* file is described as containing three parts, the general refinement instructions, the structure specific instructions, and the atom instructions.

The general refinement instructions contain crystallographic information, such as experimental setup, cell dimensions, space group in form of symmetry operations, *Z* value and element symbols for scattering factors to be employed. Some instructions have to be given in a specific order and must be given at the start of the file. Also, the refinement options as well as general restraints affecting the whole cell content such as anisotropic refinement, weighting scheme or specifying the output files and their extent are given here.

The structure specific instructions contain mainly geometric restraints for the molecules in the cell, such as bond distances and angles, chiral volumes, and torsion angles. Additionally the creation of hydrogen atoms can be achieved by specific commands written in this section.

[12] G. M. Sheldrick, *http://shelx.uni-goettingen.de*.

The atom instructions are lines beginning with the atom name (up to four characters, the first one must be a letter). This is followed by the scattering factor number, the fractional coordinates, the side occupations factor, and the U factor, with either one isotropic or six anisotropic components. Atom instructions might be intermittent with instructions organizing the molecule into residues or specifying disordered parts.

3.3.3. PDB2INS layout and architecture

PDB2INS can be started from command line under Windows, Linux or MacOSX (see Section 3.3.4). The command line version can be run in interactive mode or by using keywords after the program name at startup. In the first mode, the user is guided through the program by answering a series of questions. Additionally, a GUI version of the program is available for users more comfortable with a graphical interface.

An overview of the main steps in the program from the input file (*pdb*) to the output file (*ins*) is given in Figure 3.2. In the following some key functions of PDB2INS are explained in more detail.

Generation of symmetry operations from the space group symbol

The number of symmetry operations in a unit cell for any space group equals the multiplicity of the general position. The International Tables list for each space group all symmetry operations and specify which function as generators (Aroyo, 2016). Generators are a subset of the symmetry operations which will generate all symmetry elements of the space group by consecutive pairwise combination (Dauter and Jaskolski, 2010). Symmetry operations are tabulated as operation O, followed by the translational component w_g in parenthesis and its location (x_0, y_0, z_0).

$$O(\boldsymbol{w}_g)x_0, y_0, z_0 \tag{3.2}$$

In the International Tables the type of operation in combination with its orientation is tabulated together with the corresponding matrix W. The orientation of a symmetry operation can be derived from the location.

Taking the symmetry operations of the space group, the general positions can be calculated from their generators (Fischer and Koch, 2006). Each generator is described by a (3×3) matrix W – the rotation part – and a corresponding (3×1) column vector w representing the translation vector. With the matrix pairs (W, w) each positional vector p (x, y, z) can be transformed into the symmetry equivalent positional vector p' (x', y', z').

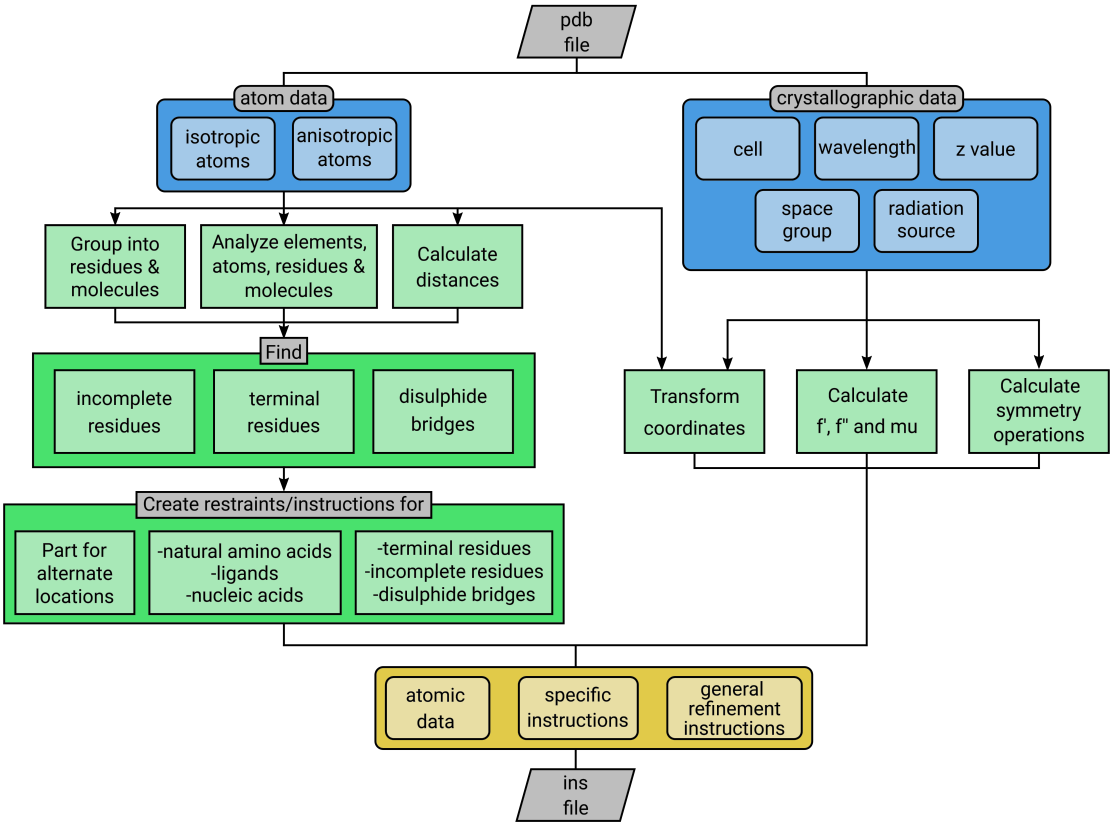


Figure 3.2.: Schematic diagram of the core processes in PDB2INS from the input file (*pdb*) to the output file (*ins*).

It is possible to obtain all symmetry operations for a given space group from the generators. The symmetry operations can be extracted from the translational vector w and the matrix W. w can be calculated from the sum of the translation component, the intrinsic translational part w_g and the location part w_l.

$$\boldsymbol{w} = \boldsymbol{w}_g + \boldsymbol{w}_l = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} \tag{3.3}$$

The location part can be acquired from matrix W, the identity matrix I, and the location.

$$\boldsymbol{w}_l = (\boldsymbol{I} - \boldsymbol{W}) \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} = \begin{pmatrix} x_0' \\ y_0' \\ z_0' \end{pmatrix} \tag{3.4}$$

A new pair of coordinates (W', w') can be derived from two matrix pairs W_n, w_n and

W_m, w_m.

$$\boldsymbol{W'} = \boldsymbol{W}_n * \boldsymbol{W}_m \tag{3.5}$$

$$\boldsymbol{w'} = \boldsymbol{W}_n * \boldsymbol{w}_m + \boldsymbol{w}_n \tag{3.6}$$

To obtain all possible symmetry operations from the generators, all matrix pairs have to be multiplied with each other until no new result is obtained.

PDB2INS contains the module SPAGSYDATA, that can generate all possible coordinate triplets from the space group symbol. It contains a dictionary of all non-centrosymmetric space groups as symbol and their associated generators as lists of matrix pairs (W, w). From this database all symmetry operations are derived by matrix multiplication until no new matrix pair is generated (Equations 3.5 and 3.6).

Each resulting matrix pair W', w' can be translated into positional vector p'.

$$\boldsymbol{p'} = \boldsymbol{W} * \boldsymbol{p} + \boldsymbol{w} = \begin{pmatrix} W_{11}x + W_{21}y + W_{31}z + w_1 \\ W_{12}x + W_{22}y + W_{32}z + w_2 \\ W_{13}x + W_{23}y + W_{33}z + w_3 \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \tag{3.7}$$

These positional vectors are given in the form of SYMM instructions in the *ins* file.

$$\text{SYMM} \quad x', y', z' \tag{3.8}$$

For each new symmetry operation a new line with a separate symmetry instruction (SYMM) is created. SHELXL does not read a space group symbol and needs the SYMM instructions as shorthand for all symmetry operations, facilitating refinement of structures in arbitrary settings.

This module is freely available and can be used independently from PDB2INS[13].

Lattice type as SHELXL command

In addition to the symmetry operation SHELXL needs the lattice type to complete a space group assignment. The lattice type is saved in the *ins* file as an integer from -7 to 7. The SHELXL code is given in Table 3.1. The code must be negative for non-centrosymmetric space groups.

[13] *https://github.com/av-luebben/spagsydata*.

Table 3.1.: Lattice type as input in SHELXL files.

SHELXL	lattice type
1	primitive (P)
2	body-centered (I)
3	rhombohedral (obverse on hexagonal axis)
4	face-centered (F)
5	A-base-centered
6	B-base-centered
7	C-base-centered

Wavelength specific considerations

SHELXL recognizes the wavelength of the most common in-house X-ray sources. The dispersion and absorption coefficients of the natural elements up to element 98 for the wavelength produced by the elements Cr, Cu, Mo, Ag and In are available. For other X-ray sources, such as synchrotron sources, SHELXL requires an instruction specifying these parameters. The instruction is given via the keyword "DISP", followed by the element prefixed by a "$" and afterwards defining f', f" and μ.

$$\text{DISP} \quad E \quad f' \quad f'' \quad \mu \tag{3.9}$$

PDB2INS calculates f', f'' and μ for all wavelengths not known by SHELXL from Kissel scattering factor tables (Roy *et al.*, 1993). If w_1 is the wavelength the data was collected at, w_b and w_a are the tabulated wavelengths, the difference is given by

$$\Delta = w_b - w_a. \tag{3.10}$$

f' can be calculated from the tabulated values f'_b and f'_a

$$f' = |f| \cdot (f'_b - f'_a) + f'_a \tag{3.11}$$

with

$$f = \frac{(w_a - w_1)}{\Delta}. \tag{3.12}$$

The calculation of f'' and μ is achieved analogously:

$$f'' = |f| \cdot (f''_b - f''_a) + f''_a, \tag{3.13}$$

$$\mu = |f| \cdot (\mu_b - \mu_a) + \mu_a. \tag{3.14}$$

Refinement commands

In the first part of the instruction file a number of commands regulating the refinement and parameters are given. They largely remain the same for all data sets and can be customized by the user during later stages of the refinement. The first three lines of the file contain the titel (TITL), wavelength, cell (CELL) and standard deviation of the cell as well as the Z value (ZERR). The title specifies from which file this *ins* file was generated, the wavelength, cell and Z value are transferred as given in the *pdb* file. The cell standard deviations are given as 0.1% from the cell length, zero for angles constrained by space group symmetry and 0.05° for all other angles. These are just general estimates and do not reflect realistic standard deviations. Adding general estimates for the cell standard deviation is necessary since no standard deviations are included in the *pdb* file.

Next, the space group is stated as remark (REM), a line which is not interpreted by SHELXL but can contain explanations or additional annotations for the user. This tool is used on multiple occasions throughout the file. The space group is then represented by one lattice statement (LATT) followed by lines defining symmetry operations (SYMM) as described in Section 3.3.3. Should the data set contain neutron diffraction data, the NEUT instruction is inserted after the symmetry operations. One line defining all elements present in the model via the SFAC command follows and their quantity in the same order on a separate line (UNIT). Should the wavelength not be that of an in-house source recognized by SHELXL, the instructions establishing the absorption and dispersion coefficients for present elements (DISP) is inserted in between these two instructions.

To change the applied standard deviations for the geometric restraints the instruction DEFS is given. Default values of 0.02 for DFIX, SADI and SAME, 0.1 for CHIV and FLAT, 0.01 for DELU, and 0.04 for SIMU are used and not changed, this can be done manually by the user if desired. The next command defines the least-squares solving method employed by SHELXL, by default PDB2INS defines the conjugate-gradient least squares (CGLS) algorithm with 20 cycles. When the second parameter is given as a negative number, it uses the fraction of the reflections reserved for R_{free} validation. This parameter is set to '-1' so that the R_{free} set as given in the *hkl* file is used. In later

stages of the refinement the CGLS command might be changed to L.S., for a full-matrix least-squares refinement. The BLOC command can be added optionally at this point to facilitate a least-squares refinement using only parts of the asymmetric unit per cycle.

Next, the included resolution range is defined (SHEL) and the grid for Fourier synthesis is specified (FMAP). Then follow some commands specifying the content and format of several SHELXL output files. The extent of a list of positive Fourier peaks that should be written to the output coordinate file is defined by the instruction PLAN. The reflection output file (*fcf*) format is defined by the command LIST and can be read by macromolecular standard programs like COOT (Emsley *et al.*, 2010). The refined coordinates are written to a *pdb* file, also including anisotropic parameters when present, when the command WPDB is given.

Restraints in general

A number of globally applied restraints follow the refinement instruction in the first part of the *ins* file. Several restraints on atomic displacement parameters (sADPs) can be applied when using SHELXL, especially during an anisotropic refinement. An anisotropic refinement requires six parameters per atom for the displacement instead of just one as is the case in an isotropic refinement. It is advisable to refine anisotropically only when the resolution is high enough and the data-to-parameter ratio is sufficiently high.

The restraint SIMU makes anisotropic ADP of two atoms more equal, which can be useful for disordered parts of the model when the atoms overlap. The rigid bond restraint DELU can be used to restrain the motion of two bonded atoms to be more equal along the bond. Another powerful restraint is RIGU, which assures that the relative motion of two bonded atoms is perpendicular to the bond joining them. This reduces the number of independent parameters per atom. The restraint can also be applied to 1,3 distances or globally on all atoms, further improving the data-parameter-ratio. The constraint XNPD defines a minimum value for all atoms' displacement parameters, enforcing positive definite values.

Furthermore, a number of specific restraints can be generated for the given content of the *pdb* file:

- Restraints for all natural amino acids present.
- Instructions for the generation of hydrogen atoms (HFIX) for all natural amino acids are given with 'REM' as prefix. Removing the prefix will add the hydrogen in the next refinement cycle as riding atoms.

- Restraints of the most common ligands are available and added when present. This only includes distance and angle distance restraints but not HFIX instructions.

- For all cysteine residues the distance between the sulfur atoms is calculated. Should two sulfur atoms be closer than 3 Å, restraints for a disulfide bridge are added.

- All amino acids are checked for their completeness. To prevent error messages in SHELXL when restraints are employed, appropriate 'HFIX 0' instruction are added for incomplete amino acid side chains.

- All terminal residues of a polypeptide chain are identified by distance calculations. Once terminal residues are recognized, the naming of the *C*-terminal oxygen residues is reviewed. Other naming conventions apply for terminal residues than for the rest of the peptide chain and the oxygen atoms are renamed if necessary. Otherwise, only the appropriate restraints are added, including the correct HFIX command for the *N*-terminal end.

- All residues showing disorder are sorted into parts. This method makes sure the disorder is handled correctly. The user can later refine the occupation of this residues by assigning free variables to the disordered parts. Until then the occupations as given in the *pdb* file are transferred.

- Bond, angle and chiral volume restraints for proteogenic amino acids are added.

Restraints for the most common ligands were created using the GRADE server[14] and are added automatically where appropriate. Some of the more specific restraints, e.g. for the terminal restraints, are explained in the following in more detail.

Terminal restraints

To generate the correct geometry and riding hydrogen atoms restraints, the terminal residues have to be handled separately. The *N* terminus needs a different hydrogen atom command, since it is consisting of an amine group. The *C* terminus is a carboxyl group and the geometry is restraint with distance restraints.

[14] *http://grade.globalphasing.org*, Global Phasing Ltd.

N-terminus:	HFIX	33				
C-terminus:	DFIX	1.249	C	OT1	C	OT2
	DANG	2.379	CA	OT1	CA	OT2
	DANG	2.194	OT1	OT2		

(3.15)

Terminal residues are identified by the calculation of distances to other residues (see Algorithm 1).

Generation of fractional coordinates

PDB2INS generates the atom instruction with positions in fractional coordinate space. Cartesian coordinates, as given in *pdb* files, are transformed as described before by Grosse-Kunstleve and Adams (2002) and Lübben *et al.* (2015).

The angles α, β, and γ in degrees are transformed to radians α', β', and γ':

$$\alpha' = \frac{\alpha}{180^\circ} \cdot \pi, \quad \beta' = \frac{\beta}{180^\circ} \cdot \pi, \quad \gamma' = \frac{\gamma}{180^\circ} \cdot \pi \tag{3.16}$$

All conversions can be achieved using the matrix M_{cf}, transforming from Cartesian to fractional space.

$$\boldsymbol{M}_{cf} = \boldsymbol{M}_{fc}^{-1} = \begin{pmatrix} \frac{1}{a} & -\frac{\cos(\gamma')}{a \cdot \sin(\gamma')} & bc\frac{\cos(\alpha')\cos(\gamma') - \cos(\beta')}{a \cdot V \cdot \sin(\gamma')} \\ 0 & \frac{1}{b \cdot \sin(\gamma')} & \frac{\cos(\beta')\cos(\gamma') - \cos(\alpha')}{b \cdot V \cdot \sin(\gamma')} \\ 0 & 0 & \frac{\sin(\gamma')}{c \cdot V} \end{pmatrix} \tag{3.17}$$

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \boldsymbol{M}_{cf} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{3.18}$$

where u, v, w are the coordinates in fractional space, x, y, z are the coordinates in Cartesian space, and V is the volume of the unit cell.

$$V = \sqrt{1 - \cos(\alpha'^2) - \cos(\beta'^2) - \cos(\gamma'^2) + 2 \cdot \cos(\alpha')\cos(\beta')\cos(\gamma')} \tag{3.19}$$

Generation of displacement parameter U_{iso} from atomic temperature factor

All *pdb* files contain the atom coordinates together with a temperature factor for the atoms, called B-factor. When the diagonal elements of the matrix are all equal to each

Algorithm 1: Find terminal residues in a macromolecule.

```
list: naturalAA and nonAA ;
// list of all natural amino acids or all else, respectively.
for each residue number in molecule do
    if residue has smallest residue number and is naturalAA then
        append residue to ListOfN − TerminalResidues
    end
    if residue has highest residue number and is naturalAA then
        append residue to ListOfC − TerminalResidues
    end
end
for atoms, residue in ListOfN − TerminalResidues do
    if atomname of atom is N then
        for atom in listofnonaminoacidresidues do
            calculate distance between atom and N;
            if distance ≤ 3.0 then
                N is not terminal residue;
                remove residue from ListOfN − TerminalResidues
            end
        end
    end
end
for each residue in ListOfC − TerminalResidues do
    if atomname of atom is O then
        // Terminal oxygen atoms can also be named OT1, OT2 or OTX
        for atom in listofnonaminoacidresidues do
            calculate distance between atom and O;
            if distance ≤ 3.0 then
                O is not terminal residue;
                remove residue from ListOfO − TerminalResidues
            end
        end
    end
end
```

other and the off-diagonal elements are equal to zero, the B-factor is isotropic. Only one number is needed to describe the isotropic B factor.

$$\boldsymbol{B}_{iso} = \begin{pmatrix} B_{11} & 0 & 0 \\ 0 & B_{11} & 0 \\ 0 & 0 & B_{11} \end{pmatrix} \tag{3.20}$$

SHELXL requires the atomic displacement parameter U_{iso} instead. PDB2INS calculates the displacement parameter via a simple conversion.

$$U_{\text{iso}} = \frac{B_{\text{iso}}}{8 \cdot \pi^2} \tag{3.21}$$

Transformation of anisotropic ADP

All anisotropic ADPs are converted into fractional coordinates as established (Grosse-Kunstleve and Adams, 2002, Lübben *et al.*, 2015). The anisotropic ADP in Cartesian space is represented by a matrix $U_{ij,\text{cart}}$.

$$\boldsymbol{U}_{ij,\text{cart}} = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{pmatrix} \tag{3.22}$$

The ADP can be transformed into the matrix representation in fractional coordinates $U_{ij,\text{frac}}$ using

$$\boldsymbol{U}^*_{ij} = \boldsymbol{M}_{cf} \cdot \boldsymbol{U}_{i,j} \cdot \boldsymbol{M}^{\mathsf{T}}_{cf} \tag{3.23}$$

and

$$\boldsymbol{U}_{ij,\text{frac}} = \boldsymbol{N}^{-1} \cdot \boldsymbol{U}^*_{ij} \cdot \left(\boldsymbol{N}^{-1}\right)^{\mathsf{T}}. \tag{3.24}$$

The conversion matrix M_{cf} is defined in Equation 3.17 and the scaling factor N is given by:

$$\boldsymbol{N} = \begin{pmatrix} \frac{1}{a} & 0 & 0 \\ 0 & \frac{1}{b} & 0 \\ 0 & 0 & \frac{1}{c} \end{pmatrix}. \tag{3.25}$$

Handling of residues

SHELXL finds most application in the refinement of small molecule structures. Many small molecule crystallographers use SHELXL in combination with the graphical user interface SHELXle (Hübschle *et al.*, 2011). In small molecule crystallography most structures do not need to be grouped into smaller units like molecules or residues for the refinement.

Nonetheless, SHELXL provides useful features for macromolecular crystallographers in that respect. While one atom per line is defined by only the atom name, coordinates, occupancy and U values, it is possible to group the atoms with specific instruction lines. Atoms can be grouped into residues and residues can be further grouped into chains. Before each separate entity, a line defining it as residue (RESI) can be entered, specifying a residue name (three characters), chain ID, and residue number. For example, the residue tyrosine in chain A with the residue number 21 can be defined by the following line specified before the corresponding atom lines:

$$\text{RESI} \quad \text{TYR} \quad \text{A} : 21 \tag{3.26}$$

It is allowed to use a residue name multiple times, but the residue number must be unique in each chain.

Once the data is structured into residues, restraints can be globally applied to all residues carrying the same name. This feature is especially useful when handling macromolecular data where a chain containing a repeating sequence of amino acids can have a standard set of restraints for all natural amino acids. For example, the following command restraints the C_α-C_β bond in all alanine residues:

$$\text{DFIX\_ALA} \quad 1.521 \quad \text{CA} \quad \text{CB} \tag{3.27}$$

Furthermore, all atoms can be explicitly named by referring to them in the form

$$\text{atom name\_chainID} : \text{residue number}, \tag{3.28}$$

e.g. OT1_A:125 refers to atom 'OT1' in residue 125 in chain A.

The restraints are implemented in PDB2INS as they were published by Engh and Huber (1991). Additionally, PDB2INS contains a library of restraints for the most common ligands in the PDB. These restraints were mainly generated using the GRADE server.

Another hurdle is the residue name accepted by SHELXL. The residue name can be three characters long, as it is common in macromolecular crystallography, but must

contain at least one letter. Most common residues names contain a letter, such as amino acids or nucleobases, and do not pose a problem. But since a large number of ligands and small molecules can be present in a crystal, some of these have a PDB code containing only digits. These ligand PDB codes are naturally used as residue name for identification purposes but can not be processed by SHELXL. PDB2INS recognizes these ligands and offers the option to rename them.

Treatment of hydrogen atoms

Should neutron data be present, a specific instruction (NEUT) must be included in the first part of the *ins* file. This instruction prompts SHELXL to employ neutron scattering factors and absorption coefficients and adjusts distance restraints as well as a more sensible treatment of hydrogen atoms. Furthermore, all hydrogen atoms already present in the *pdb* file are kept and transferred to the new file by PDB2INS. This option is also available for XRD data, but not done automatically. Iin that case it is recommended to constrain all hydrogen with commands (HFIX) as riding atoms. Specific treatment of hydrogen atoms in a refinement against neutron data is described in Gruene *et al.* (2013).

PDB2HKL

A separate module PDB2HKL was written, since the handling of reflection data represents a challenge in macromolecular crystallography. The most common reflection data format in macromolecular crystallography is the crystallographic file format *cif* or *mmcif*. This should not be confused with the small molecular *cif*, since no consistent file format standard is applied. A number of other file formats are in use such as *mtz*, *sca* or *HKL*. In small molecular crystallography the *hkl* format is predominant and is used for SHELXL refinements. There are a number of scripts and programs one can employ to convert between different file formats. Most of them are applied automatically as part of pipelines or with graphical user interfaces such as CCP4i2 or PHENIX. Others are stand-alone programs such as CIF2MTZ (Winn *et al.*, 2011), MTZ2HKL (Grune, 2008), or XPREP. Since SHELXL only reads *hkl* format reflection files and one objective was to create an easy access to macromolecular refinement with SHELXL, the utility program PDB2HKL was written. PDB2HKL can take the crystallographic reflection data *mmcif* file as it is provided by the PDB and create an *hkl* file suitable for SHELXL refinement.

PDB2HKL reads the '*name*-sf.cif' file and searches for keyword sets. Keyword sets as listed in Table 3.2 are recognized. All keywords of a given keyword set must be present

in the reflection data file to be accepted as a complete data set. Should more than one set of reflections be present, the program selects intensities over amplitudes and unmerged over merged data. The program will write feedback to the console, stating which data was found and transferred into the *hkl* file.

Table 3.2.: Keywords recognized by PBD2INS module PDB2HKL from reflection files in *mmcif* format. A set of keywords has to be complete to be accepted.

reflection type	keyword set
unmerged intensities + standard uncertainties	_refln.pdbx_I_plus, _refln.pdbx_I_minus, _refln.pdbx_I_plus_sigma, _refln.pdbx_I_minus_sigma
merged intensities + standard uncertainties	_refln.I_meas_au, _refln.I_meas_sigma_au
	_refln.I_meas, _refln.I_meas_sigma
	_refln.F_squared_meas, _refln.F_squared_sigma
	_refln.intensity_meas, _refln.intensity_sigma
unmerged amplitudes + standard uncertainties	_refln.pdbx_F_plus, _refln.pdbx_F_minus, _refln.pdbx_F_plus_sigma, _refln.pdbx_F_minus_sigma
merged amplitudes + standard uncertainties	_refln.F_meas_au, _refln.F_meas_sigma_au
	_refln.F_meas, _refln.F_meas_sigma

Since the standard format of a reflection in the *hkl* file only allows seven characters for the amplitude/intensity of the reflection, it can not exceed 9999999. Is the amplitude/intensity greater than this value, the whole dataset is scaled by a factor setting it to 9999999. Reflections where either h, k, l, reflection intensity/amplitude or estimated standard deviation (e.s.d.) are undefined (indicated by a '?' character) are omitted. The file will be terminated with a line containing all zero items.

PDB2HKL was integrated into PDB2INS and is executed as a first step. The generation of an *hkl* file is optional, but the information whether amplitudes or intensities are present in the data file is mandatory for the *ins* file. PDB2HKL will extract this information when used and forward it to the main program.

3.3.4. Versions of PDB2INS

All versions are available as standalone executables for Linux, Windows and MacOSX. The source code is published under the GNU lesser general public license and is freely

available.

Command line use

It is possible to use PDB2INS is as an interactive command line program. The user is guided through the creation of input files for a SHELXL refinement in a step-by-step manner. When starting the program without any additional commands, the start screen is visible as depicted below.

```
INFO: +++ starting PDB2INS. +++

    ##############################################################################
    #                                PDB2INS                                     #
    #                 by Anna V. Luebben (Darwin\_x86\_64@2018-01-04)            #
    ##############################################################################

    Reads a PDB file and generates an .ins file for SHELXL.
    The PDB file is assumed to conform to the Protein Data Bank notes
    'Atomic Coordinate and Bibliographic Entry Format Description Version
    3.30'. For remarks and problems please contact aluebbe@gwdg.de.

    Usage:
        pdb2ins <filename|@pdbcode> [options]

    <filename|@pdbcode> Exact name of a file in PDB format or
                        a legal pdb code prefixed by '@'.

    Enter '--help' for complete options list.
    Enter 'q' or 'exit' to exit the program.


Create .hkl file from structure factor file (cif) or PDB code? (y or n) [N]:
```

At first, an *hkl* file containing reflection data can be created from an *-sf.cif* file. The format of the reflection is a necessary information for the creation of the *ins* file. For this reason the program asks for a reflection data file first and extracts the format.

Successively, the program will create the *hkl* file and read the *pdb* file. For all missing information or supplementary options the user is prompted for input. Additionally, all relevant changes or information is given to the user using the keywords **info**, **attention** or **error**.

PDB2INS will confirm that all requested files were written before terminating. The user can end the program at any point by simply using the command 'q' or 'exit'.

Several options can be passed to the program via command line arguments (see Table 3.3). An overview of all options will be displayed if the program is started with the argument 'pdb2ins - -help'. These options allow the user to control nearly all decisions

in the program while starting it. This provides a powerful tool to use the program as part of a script or pipeline and skip all interaction.

Table 3.3.: Options available during startup of PDB2INS prefixed by '-' are tabulated.

option	description
w	followed by the wavelength in Ångstroms to enter a wavelength manually.
h	followed by a value to enter hkl format in SHELX syntax.
c	followed by the six coordinates (a,b,c,alpha,beta,gamma) in Ångstroms and Degrees to enter the cell (No spaces allowed).
s	followed by the space group (No spaces allowed).
i	skip user input interrupts.
a	use anisotropic displacement data if available.
b	create an *hkl* file from a PDB structure factor file (*cif*) or PDB code.
d	followed by a filename to specify structure factor file in *cif* format.
e	all hydrogen atoms are transferred from the *pdb* file to the *ins* file.
o	followed by a filename to specify the output filename.
r	if a PDB code is given with '@', this option fetches the *pdb* file from the PDB_REDO server.
z	followed by a number specifying the Z value (number of molecules per cell).

PDB2INS can be used in a customized, automated way. The option '-i' allows the user to skip all questions when running the program and use default answers instead. In combination with the other options these can be modified to suite the users needs.

Graphical User Interface

TKinter was used to create the GUI (see Figure 3.3). The top part of the GUI is used to capture user input as well as select options for the file conversion. The bottom part displays a text window either showing output from the program or an explanation for the buttons of the top half.

At the top left corner the file name can be typed in or selected from the file structure. In case a PDB file should be downloaded from the RCSB PDB website, the four letter PDB code must be given prefixed with '@'. On the right a 'Load' button is located. Once the file field is filled in, the Load button should be pressed to display the important information from the file. When a PDB code was inserted, the file is downloaded beforehand. Should

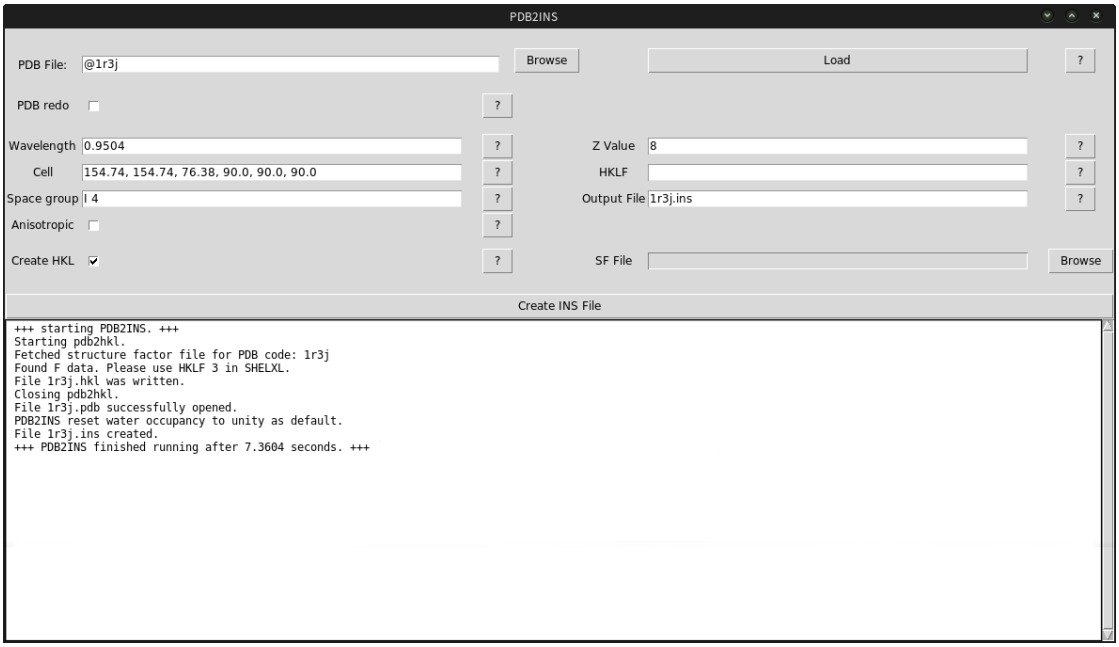


Figure 3.3.: Depiction of the PDB2INS graphical user interface developed to streamline the use of SHELXL with macromolecular crystallography data.

one press a question mark button next to one of the fields or buttons in the top half, the corresponding explanation will appear in the text window.

Most options of the command line program are available in the GUI, for others default answers are used. For example hydrogen atoms are automatically transferred to the *ins* file for neutron diffraction data but not for XRD data. The program will not start until all essential information is completed in the fields above the start button. Is information missing, the corresponding field is highlighted with a red background color. A successful run of the program can be monitored in the text window on the bottom half, were command line output is displayed. Also errors while running the program will be displayed in the text window.

This graphical user interface is freely available[15].

3.3.5. Refinement with SHELXL

Refinement with SHELXL can be started from the command line with the command 'shelxl *name*', where name is the file name base (without extension) of the required files. It is recommended to perform the refinement against intensities (F-squared) in conjugate gradient (CGLS) mode. Results of the refinement can be inspected using an interactive

[15] *http://github.com/av-luebben/PDB2INSGUI*.

graphics display program such as COOT. It is advisable to compare the models with the respective electron density map after each refinement iteration.

Optionally, disordered side chains or sections can be modeled by using the PART instruction. Disorder can be modeled as two-fold or multi-fold and the occupation of each part can be refined using free variables (FVAR). Free variables in SHELXL are a powerful concept to refine e.g. the occupancy of a single atom or a group of atoms in a disorder. For example, in combination with the PART instruction, the occupancy of a disordered side chain in two parts can be refined for each part individually with the restriction that the sum of the individual occupancies must be one.

At first, it is recommended to refine all parameters isotropically to quasi convergence and then continue with an anisotropic refinement if the resolution of the data is sufficient (Sheldrick, 1996). An anisotropic refinement is not recommended for structures with an overall resolution lower than 1.5 Å. The introduction of anisotropic displacement parameters should lead to a drop in both R_1 (R_{cryst}) and R_{free}. If a drop of R_{free} is not observed, the refinement of anisotropic ADP is not sensible and should be abandoned.

The addition of hydrogen atoms can be achieved by adding appropriate commands to the instruction file (HFIX). This commands were generated automatically by PDB2INS during the file creation, but are not used by default. To use these instructions, the preceding 'REM' before the command must be removed.

As final step, a cycle of full-matrix least squares (L.S.) refinement, optionally blocked (BLOC), can be performed to obtain standard uncertainties of all refined parameters.

3.4. Results and discussion

PDB2INS successfully generates input files for a refinement of macromolecules with SHELXL. Multiple ways to access PDB2INS are available to all users. The program can be used with a graphical interface (PDB2INSgui) or in an interactive mode via command line. Also, the command line version can be used without user input in an automated way, for example to implement PDB2INS in scripts.

PDB2INS recognizes a variety of specific PDB keywords and takes appropriate action to translate them into SHELXL format without loss of information. The need to manually edited the instruction file for SHELXL is reduced to a minimum or, in most cases, made obsolete.

PDB2INS automatically introduces restraints developed in recent years in a sensible way. Furthermore, with PDB2INS the direct access to SHELXL is possible using a PDB code. No other programs are necessary to generate all files essential to refine with

SHELXL when using PDB standard format files.

PDB2INS offers several advantages that have not been available before. In comparison to SHELXPRO, the user has more control over the information the instruction file can contain. Also, more information is extracted from the input files by default. Some additional restraints for ligands are directly added to the instruction file for SHELXL. Additionally, the conversion of structure factor *cif* files is now available directly from within PDB2INS. Another improvement is the availability of a graphical user interface for users not comfortable with command line programs.

3.4.1. Test of PDB2INS against protein database files

PDB2INS was tested in an automated way against data available from the PDB. Only XRD structures deposited after 2007 and containing reflection data as well as coordinates were selected. All data with a resolution of 1.7 Å or better where chosen as a test set, amounting to 23 974 structures in total.

From all files tested, 95.98% were converted to the SHELXL input files without error and could be refined successfully (see Figure 3.4). A successful refinement is assumed when SHELXL starts with the refinement and finishes it after ten conjugate-gradient least-squares cycles. The remaining 4.02% of files could not be processed fully. The points of failure were investigated and are presented in this section and in Section B.2.

At several points during the test, information about problems and error messages was collected. PDB2INS recognizes the most common complications and writes appropriate log messages. This messages can be utilized by the user to edited the input files, if necessary, and run PDB2INS again.

SHELXL also displays feedback when a refinement with the applied files is not possible. Usually, this feedback can be used to correct the input files as well. Some complications occur more frequently than others, the most common ones are listed in Table 3.4.

The test of nearly 24 000 *pdb* files showed that the majority of files (23 010 files, 95.68% total) could be refined with SHELXL without any problems. From the remaining data sets (964 files, 4.02% total), 19.7% (190 files, 0.79% total) manifested an error in PDB2INS and 80.3% (774 files , 3.23% total) in SHELXL. This does not mean that the complication itself was caused by the program displaying the error, but that it was simply detected while using it. An overview of the most commonly reported complications is displayed in Figure 3.5, grouped by the reporting program.

Table 3.4.: Most common complications while using PDB2INS and SHELXL in succession.

program[a]	error	description
PDB2INS	incomplete structure factor file	The inquiry file *-sf.cif* did not contain a complete set of reflection keywords. PDB2INS only converts a reflection file if a keyword set is complete. Accepted keyword sets are listed in Table 3.2.
	more than one model	It is legal to deposit more than one model in one *pdb* file. PDB2INS cannot handle those files and will terminate without writing a new file. Removing all but one model from the input file will enable PDB2INS to convert the file.
	problem while handling insertion codes	PDB2INS is capable of renumbering residues when an insertion code is used. This is necessary since no specific method of handling insertion codes is available. When this renumbering is affected by complications, the residue that could not be renumbered is included in the error message.
SHELXL	bad resi	SHELXL terminated without refinement when one of the residues has a residue name containing only digits. This is allowed in *pdb* files. PDB2INS will prompt the user to change this residue names when run in interactive mode. However in an automated mode, PDB2INS only writes a warning but continues without renaming.
	reflection *hkl* has wrong format	One reflection does not adhere to the *hkl* format as specified in Section 3.3.2. The *hkl* file can be edited manually to fix this reflection. The next version of PDB2INS will reformat those reflections appropriately.
	unknown element	The *pdb* file contains an element name in the SFAC instruction line that does not correspond to the first 98 elements of the periodic table. Most commonly, one of the elements is specified as 'X', an unknown element, which naturally does not correspond to any scattering factors in SHELXL.
	no match for *atom* in *restraint*	One or more atoms do not have the appropriate restraints specifying them. The user can check if the mentioned residues are named correctly.

[a] The program reporting the error is listed here. This does not necessarily imply that the error lies within the program itself.

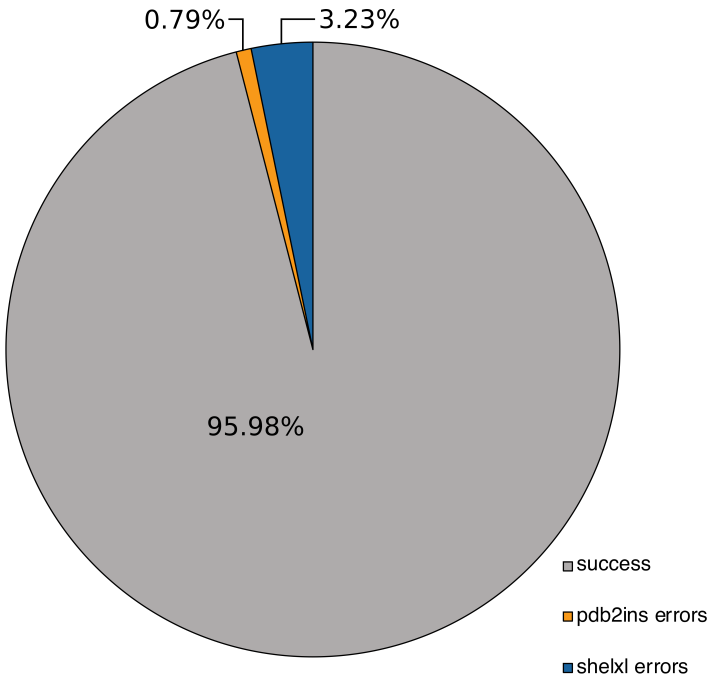


Figure 3.4.: Analysis of PDB2INS test against all XRD data to a resolution of 1.7 Å deposited with reflection data in the PDB since 2008.

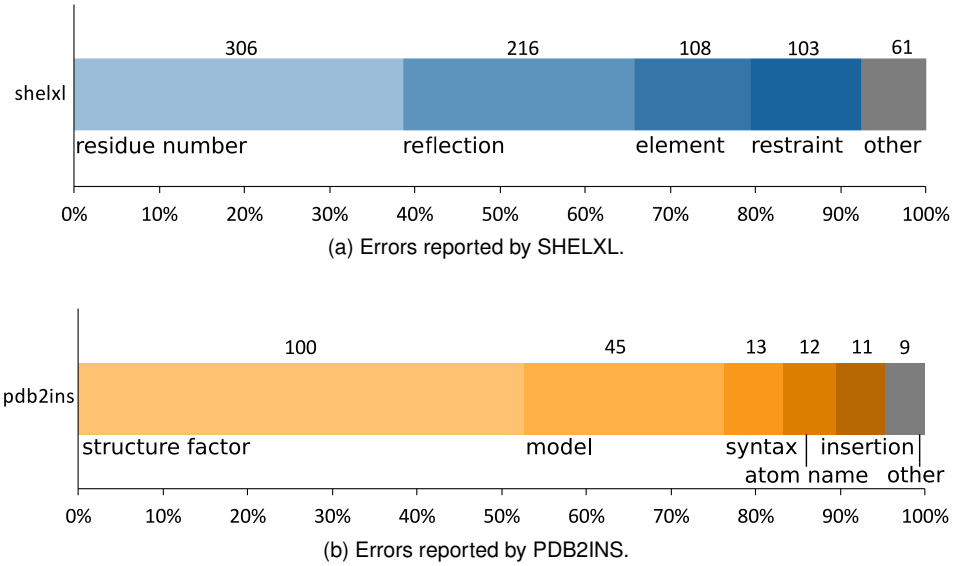


Figure 3.5.: Overview of the test results of PDB2INS against a selected part of the PDB database. Complications reported by SHELXL (a) and PDB2INS (b) are displayed with number of occurrence (above) and percentage (below).

The most common problem was an issue regarding the residue name. To this problem appertain 31.7% (306 files, 1.28% total) of all failures to refine a dataset within SHELXL. The affected data sets contain one or more residues with a residue name consisting of only digits. These residues do not adhere to the naming convention in SHELXL but are legal in the PDB format. PDB2INS provides the option to rename these residues while using the interactive mode. The user is automatically informed of the expected conflict and can immediately rename the residue.

The second most common problem was regarding an issue with the format of reflections. This problem is reported by SHELXL and occurs in 22.4% (216 files, 0.90% total) of all failures during refinement. This complication is caused by the conversion of the reflection from *-sf.cif* to *hkl* format in PDB2INS. Since only a single or a few reflections are affected by this format error, it is possible to manually edit the *hkl* file.

Furthermore, 11.2% (108, 0.45%) of errors are caused by the inclusion of an unknown element in the structure, often named element X. Since this element does not exist in the periodic table of elements, SHELXL cannot assign scattering factors to the element. This requires the users intervention to assign valid elements to the atoms in question. All other issues are listed in more detail in Section B.2.

In conclusion, PDB2INS can facilitate an easy, direct access to a refinement of macromolecular structures with SHELXL for nearly all of the suitable data sets. The test was designed to bring bugs in PDB2INS to attention although it also provides evidence of the success. Further developments in PDB2INS will improve upon this results.

3.5. Outlook

3.5.1. Possible developments in SHELXL

SHELXL allows the modeling of diffuse solvent regions using a variation of *Babinet's principle* (Moews and Kretsinger, 1975). This is implemented in a single instruction (SWAT) and can improve the agreement of the very low angle data. The method warrants improvement and cannot compete with implementations in other refinement programs. The development of a better, more suitable solvent model in SHELXL is desirable.

3.5.2. Further prospects of PDB2INS

Currently the program PROSMART (Nicholls *et al.*, 2012) can be used to generate external restraints (Murshudov *et al.*, 2011), which are then converted to SHELXL format (Gruene *et al.*, 2013). It is conceivable to extent PDB2INS with an option to derive

restraints for a SHELXL refinement directly from a homologue structure. Restraints from a homologue could be generated in the form of 1,4 distance restraints and applied to the new structure. A search for the longest common substring in the amino acid sequence can be implemented into PDB2INS. The so established sequence identity can be the basis of the restraints.

Furthermore, the progress of the PDB and their file format system have to be taken into account. While most refinement programs still use *pdb* as their standard output file format, the PDB might change the available file format for direct download. Before long, the *pdb* format will be replaced by a *pdb-cif* format for all coordinates and associated information. Support for this file format can be implemented into PDB2INS as well.

A further possible addition to PDB2INS can be an interface to the GRADE server for generating restraints. Today, only a database of the most common ligands in the PDB is available in PDB2INS. In addition to the pre-computed restraints stored within PDB2INS, users could be given the option to generate appropriate restraints on-the-fly. The user can use the output from PDB2INS directly and be sure that all molecules present are restraint appropriately. However, right now the server is available for non-interactive scripts only upon request.

4. Ceroid Lipofuscinosis Neuronal Protein 5

Ceroid-Lipofuscinosis Neuronal Protein 5

Neuronal ceroid lipofuscinoses (NCLs), also known as Batten disease, is a group of rare and severe autosomal-recessively inherited diseases involving neurodegeneration – a progressive loss of neurons. The disease group is associated with retinopathy, cognitive decline, myoclonic epilepsy and progressive cerebellar atrophy. The NCLs belong to the superordinate group of lysosomal storage diseases. The inherited conditions are characterized by the lysosomal accumulation of auto-fluorescent lipofuscin-like lipopigments in neuronal and extra-neuronal tissues (Goebel, 1997). Earlier classified among the amaurotic family idiocies, the term neuronal ceroid lipofuscinosis was first used by Zeman and Dyken (1969). NCL became delimited to inherited storage disorders with progressive decline of mental, motor, and visual functions (Dyken, 1989).

NCL associated proteins are heterogenous in their function and cellular location. Mutations in these proteins cause common clinical and pathological features. While the function of many NCL proteins is established, the protein structure of cln5 remained unknown and the function poorly understood. The ceroid-lipofuscinosis neuronal protein 5 (cln5) is the focus of this work[1].

4.1. Neuronal Ceroid Lipofuscinoses

4.1.1. A brief history

O. C. Stengel, a physician in Norway, first described a probable NCL in a case of four siblings. The clinical features are compatible with the disease today classified as CLN3 disease (Stengel, 1826)[2]. The children displayed blindness and progressive dementia with juvenile onset of the disorder. Later, the English neurologist F. E. Batten described

[1] In this work the following protein and gene nomenclature will be applied: Human gene symbols are in uppercase letters and italicized. Disease designations are the same as the gene symbol but not italicized. Protein symbols are in lowercase letters of the disease symbol unless named after their protein product. The mouse gene is italicized but only the first letter is uppercase. For example: human gene: *CLN5*; disease: CLN5; protein: cln5; mouse gene: *Cln5*.

[2] The original article 'Beretning om et mækeligt Sygdomstilfælde hos fire Sødskende I Nærheden af Röraas' was translated to 'Account of a singular illness among four siblings in the vicinity of Røraas' and mentioned first in (Armstrong *et al.*, 1982).

two cases within one family showing the neuropathology of cerebral degeneration with ocular macular changes (Batten, 1903, 1909). Shortly after, F. Spielmeyer (1905) and H. Vogt (1906) reported a similar disorder, which became known as Batten disease or Batten-Spielmeyer-Vogt disease.

With a late-infantile onset, a similar disease was reported by J. Janský (1908) and M. Bielschowsky (1913) and referred to as Janský-Bielschowsky disease (earliest mention of CLN2). An adult-onset form with similar pathological characteristics was described by H. Kufs in the 1920's and 1930's (Anderson *et al.*, 2013, Kufs, 1925). In contrast to the earlier described juvenile and late-onset cases, Kufs disease, or CLN4, did not progress with loss of vision. A disease with early-onset, known as Haltia-Santavuori disease or classic infantile, was firstly described by M. Haltia and P. Santavuori during the 1970's (Haltia *et al.*, 1973).

An overview of the NCLs associated genes and their classification is displayed in Table 4.1. Further background about NCL pathogenesis is provided in Section C.1.1.

4.1.2. Classification

While earliest cases were associated by similar pathology it became clear that the diseases display a different age of onset and were classified thereby. Although known by their eponyms, the expression 'Batten disease' was often used for the juvenile-onset disease as well as for the whole association of diseases. As causative genes were unknown, the classification was achieved by distinctive ultrastructural patterns in addition to age of onset (Zeman and Dyken, 1969). Later discovered forms of NCL were referred to by their country of origin, e.g. CLN5 and CLN7 were referred to as Finnish variant and Turkish variant, respectively (Santavuori *et al.*, 1982, Williams *et al.*, 1999).

At first, NCLs had been associated with the so-termed 'amaurotic family idiocies' due to the resemblance of clinical features with Tay-Sachs disease, the prototype of this disease family. Zeman and Dyken (1969) revised this by designating them as inherited storage disorders, clinically characterized by progressive decline of mental, motor, and visual functions. Thereby a new group was created, distinctly separated and termed neuronal ceroid lipofuscinoses in 1969 (Zeman and Dyken, 1969).

At length the nature of the storage material accumulated had been discussed and was a contributing factor in distinguishing Tay-Sachs-type from NCL diseases (Terry and Korey, 1960). Long described as auto-fluorescent lipopigments, ceroid or lipofuscin (Zeman and Donahue, 1963, Zeman and Dyken, 1969), the major component of the accumulated storage bodies was identified as subunit c of mitochondrial adenosine triphosphate synthase (ATPase) for most animal and human NCL forms (Palmer *et al.*, 1986). The flu-

Table 4.1.: NCL genes and their associated proteins (inspired by (Mole *et al.*, 2012)).

gene	linked disease[a]	protein	sub-cellular location	function/description[a]
CLN1	INCL (LINCL, JNCL, ANCL)	Palmitoyl protein thioesterase 1 (PPT 1)	Lysosomal matrix, lipid rafts, presynaptic areas in neurons	soluble protein, palmitoyl thioesterase involved in lysosomal degradation of S-fatty acylated proteins.
CLN2	classic LINCL	Tripeptidyl peptidase 1 (TPP 1)	Lysosomal matrix and ER	soluble protein, serine protease, removes *N*-terminal tripeptides at acidic pH.
CLN3	JNCL	CLN3 TM protein	Late endosomal/ lysosomal membrane	TM protein, unknown function.
CLN4	ANCL, Kuf's disease, Parry type	Cysteine string protein α (CSPα)	cytosolic, synaptic vesicles in neurons, secretory granules	soluble protein Hsc10 co-chaperone.
CLN5	Finnish vLINCL	cln protein 5	lysosomal matrix	soluble protein, unknown function.
CLN6	vLINCL earlyJNCL (ANCL)	cln protein 6	ER membrane	TM protein, unknown function.
CLN7	Turkish vLINCL	major facilitator superfamily domain-containing protein 8 (MFSD8)	lysosomal membrane	TM protein, member of the major facilitator superfamily (MFS) of secondary active permeases, endolysosomal transporter.
CLN8	progressive EPMR, northern epilepsy, LINCL	cln protein 8	ER-Golgi intermediate compartment membrane	TM protein, unknown function, related to lipid homeostasis.
CLN10	congenital classic NCL, LINCL, ANCL	Cathepsin D (CTSD)	lysosomal matrix, extracellular	soluble protein, aspartyl endopeptidase.
CLN11	ANCL	Progranulin (PRGN)	extracellular	secretory protein, multi-domain protein to be proteolytically cleaved to granulins A-G.
CLN12	JNCL	P-type ATPase (ATP12A2)	lysosomal membrane, multi-vesicular bodies	TM protein, unknown function, possible shuttle across cell membranes.
CLN13	Adults Kufs type B	Cathepsin F	lysosomal matrix	soluble protein, cysteine protease.
CLN14	INCL	Potassium channel tetramerization domain-containing protein 7 (KCTD7)	cytosolic	soluble protein, unknown function (probable TM protein voltage-gated potassium channel complex).

[a] INCL = infantile NCL, LINCL = late infantile NCL, vLINCL = variant late infantile NCL, JNCL = juvenile NCL, ANCL = adult NCL, EPMR = progressive epilepsy with mental retardation, cln = ceroid-lipofuscinosis neuronal, TM = trans-membrane.

orescence of the storage bodies is an aggregate property of non-fluorescent compounds (Palmer *et al.*, 1993, 2002). The ATPase is a transmembrane protein responsible for the synthesis of adenosine triphosphate (ATP). Adjacent sphingolipid activator proteins (ssaposins) A and D have been found as main components in infantile NCL and autosomal dominant form of adult NCL (Nijssen *et al.*, 2002, Tyynelä *et al.*, 1993). Saposins facilitate the catabolism of glyco-sphingolipods in lysosomes (Munford *et al.*, 1995). Thus, another subdivision of the NCL diseases into those accumulating subunit c of the mitochondrial ATPase and those storing saposins A and D, is possible.

Developments in the field of genetics allowed to link the CLN3 disease to chromosome 16 (Eiberg *et al.*, 2008). In 1995 the gene responsible for the infantile Haltia-Santavuori disease (CLN1) was identified by a positional candidate gene approach (Vesa *et al.*, 1995). In the same year, the genes underlying classic late infantile Janský-Bielschowsky disease and juvenile Spielmeyer-Sjörgen disease were identified as *CLN2* and *CLN3*, respectively (Lerner *et al.*, 1995, Sleat, 1997). Due to the fast advancements in the identification of NCL genes, a new classification based thereupon evolved (Goebel *et al.*, 1999, Haltia, 2003, Hofman and Peltonen, 2002, Mole, 2004, Wisniewski *et al.*, 2001). So far, nine genetic forms of NCL have been identified (*CLN1–8*, *CLN10*) and more show evidence of accumulation of ceroid lipofuscin (*CLN11–14*).

As more forms were discovered, additional features distinguished the NCLs, for instance specific clinical features, rate of progression, underlying cell biology, and biochemistry. A proposed general definition is "A progressive degenerative disease of the brain and, in most cases, the retina, in association with intracellular storage of material that is morphologically characterized as ceroid lipofuscin or similar." (Mole *et al.*, 2012) With advances in gene classification and analysis, many experts suggest a primary classification by gene and secondary classification by clinical features and age of onset (Williams *et al.*, 2011). Diagnosis based on clinical or phenotypical features remains difficult, since different mutations within one gene may result in varying phenotypes and age of onset. Therefore, a more extensive taxonomy has been recommended as a clinical nomenclature consisting of seven axes (Williams and Mole, 2012):

1. affected gene (CLN gene symbol)
2. mutation diagnosis
3. biochemical phenotype
4. clinical phenotype
5. ultrastructural features
6. functionality
7. other remarks.

4.1.3. Interaction and common pathways

Most of the NCLs show common pathways, but full understanding of the interactions and mechanisms remains elusive. A common factor in the severe neurodegenerative diseases of some NCLs is the disturbed lipid metabolism. Lyly *et al.* (2008) proposed that cln1 interacts with the F_1-complex of the mitochondrial ATPase. In *Cln1* knockout mice (*Cln1*$^{-/-}$) studies, neurons show entropic α and β subunits of ATPase enriched in plasma membrane (Getty and Pearce, 2011). It has been reported earlier that infantile NCL patient's fibroblasts show reduced basal ATPase activity and deficient regulation of the enzyme (Das *et al.*, 1999).

Components of ATPase localize to the plasma membrane of many different cell types and may act as receptors for multiple ligands as well as participate in regulation of lipid metabolism and cellular proliferation (Arakaki *et al.*, 2003, Kim *et al.*, 2004, Martinez *et al.*, 2003). *Cln1*$^{-/-}$ mice have been reported to show reduced levels of total cholesterol, apolipoprotein A1 (ApoA1), and attenuated phospholipid transfer protein (PLTS) activity by serum lipid analysis (Lyly *et al.*, 2008). Reportedly PLTS is involved in the transfer process of cholesterol to high density lipoprotein by the ATP-binding cassette transporter 1 (ABCA1)-mediated pathway (Oram *et al.*, 2003). Furthermore, in *Cln1*$^{-/-}$ mice glial cells and neurons an increased uptake of radio-labelled ApoA1 was observed (Lyly *et al.*, 2008).

The interaction of the protein cln1, palmitoyl-protein thioesterase 1 (PPT1), and the protein cln5 with the mitochondrial F_1-ATPase in *in vitro* assays was reported by Lyly *et al.* (2008, 2009). Mitochondrial F_1-ATPase was described to act as a receptor for ApoA1 (Martinez *et al.*, 2003). Abnormal molecular phospholipid species were found in the cerebral cortex of CLN1 and CLN3 patients (Käkelä *et al.*, 2003).

Other NCLs have been described to play a role in disturbed lipid metabolism as well. The cln8 protein reportedly influences ceramide synthesis, lipid regulation, and protein translocation in the endoplasmic reticulum (ER) (Haddad *et al.*, 2012, Kuronen *et al.*, 2012, Winter and Ponting, 2002). Cathepsin D (CTSD), the cln10 protein, and cln6 deficient mice exhibited increased amounts of complex glycosphingolipids in neurons and glial cells (Jabs *et al.*, 2008). Haidar *et al.* (2006) described that CTSD influences ABCA1-mediated efflux and cholesterol trafficking.

4.2. CLN5

CLN5 disease is linked to the *CLN5* gene located on chromosome 13, which encodes a polypeptide of 407 amino acids (Savukoski *et al.*, 1998). Early reports of CLN5 described it as late infantile NCL, but cases from early juvenile to adult onset have been reported since. The encoded protein has an *N*-terminal signaling sequence that is cleaved after localization to the ER. Furthermore, the lysosomal protein has eight putative *N*-glycosylation sites that are utilized *in vivo* (Moharir *et al.*, 2013). Further background on the pathogenesis and clinical features of CLN5 are provided in Sections C.1.2 and C.1.3, respectively.

4.2.1. Protein and modifications

The *CLN5* gene was assigned to chromosome 13, 13p21–q32 (13q21.1-32)[3] by linkage analysis in Finnish families with common ancestors (Savukoski *et al.*, 1994, Varilo *et al.*, 1996).

The gene contains four exons decoding the 407 amino acid long polypeptide, cln5 protein (cln5) (see Figure 4.1). The protein shows no sequence homology with any other protein (Isosomppi *et al.*, 2002, Savukoski *et al.*, 1998). Cln5 is prevalent in vertebrae and shows strong sequence conservation (see Figure 4.1).

Several potential methionine start sites are encoded in exon 1, holding four in-frame initiation methionines. Initiator methionines are located at positions 1, 30, 50, and 62, producing cln5 polypeptides with expected molecular masses of 40.3, 41.5, 43.4 and 46.3 kDa, respectively (Isosomppi *et al.*, 2002, Savukoski *et al.*, 1998). The longest polypeptide encoded from the first methionine was reported to represent a membrane bound form of the protein (Vesa *et al.*, 2002). Schmiedt *et al.* (2010) reported that proteolytic cleavage of cln5 produces polypeptides of identical size, independent from the used initiation methionine.

According to *ab initio* prediction methods, the isoform 1 of cln5, starting with methionine 1, contains a signal peptide covering amino acids 1–42. Cln5 was additionally reported to have a signal peptide which was cleaved at Ile96 (Schmiedt *et al.*, 2010) or Val93 (Jules *et al.*, 2017).

At first, cln5 was reported to be a transmembrane protein (Savukoski *et al.*, 1998) or to exist in either in soluble and transmembrane forms (Vesa *et al.*, 2002). The protein sequence contains two hydrophobic stretches at amino acids 76–91 and 353–373

[3]Genomic RefSeqGene NG_009064.1 (17594 bp) Transcript RefSeq NM_006493.3, NCBI RefSeq NP_006484, EC nomenclature MIM# 256731.

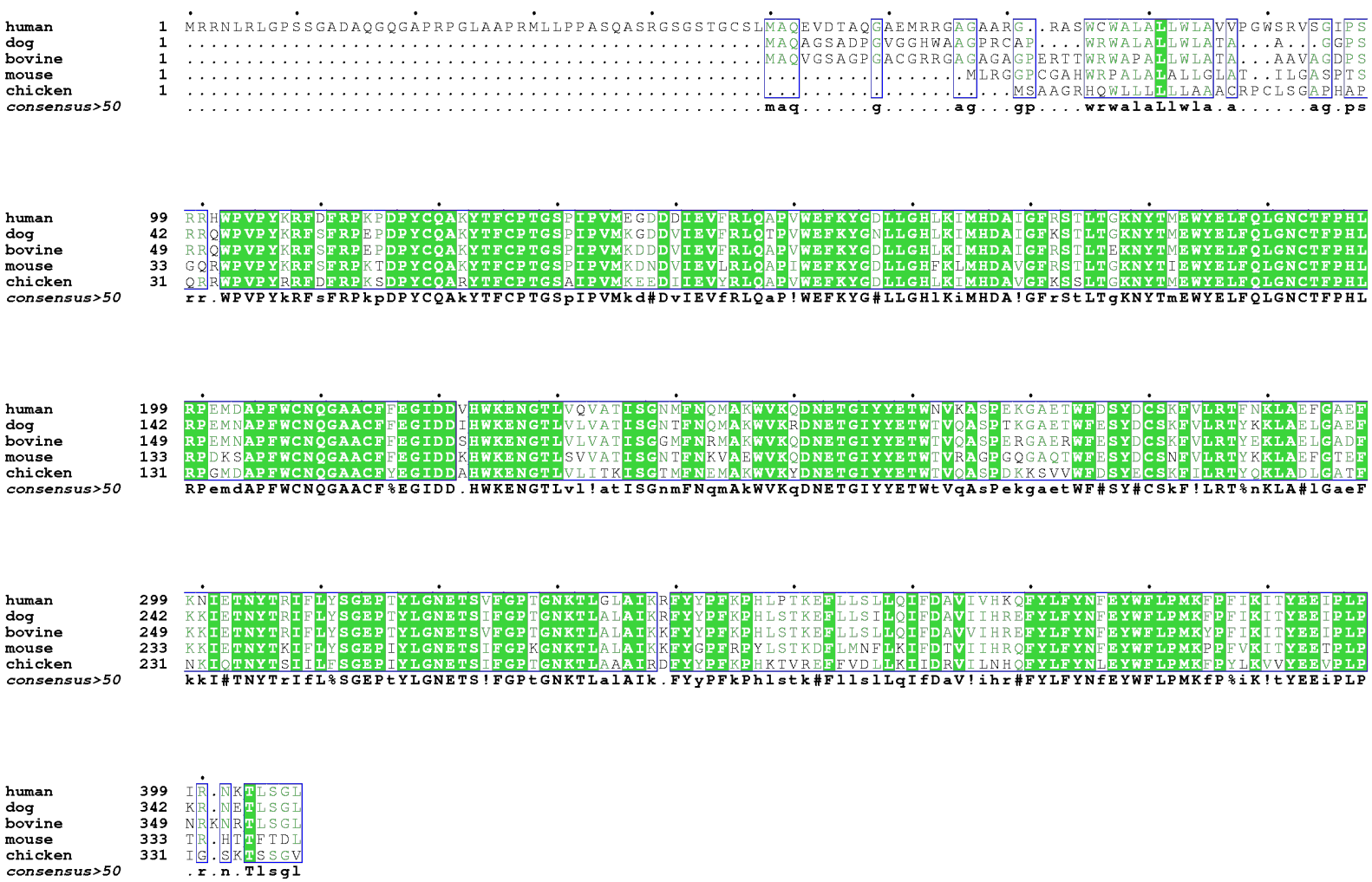


Figure 4.1.: The high sequence preservation in vertebrae is depicted. High consensus is marked as green box, low consensus as green letter. Consensus defined as mean sequence identity $\geq$ 50% and mean sequence similarity $\geq$ 80%. The figure was created with MultAlin (ESPript3.0 (Corpet, 1988).

predicted to be transmembrane domains (Savukoski *et al.*, 1998). However, prediction programs SOSUI (Hirokawa *et al.*, 1998) and TMHMM (Krogh *et al.*, 2001) did not identify any potential transmembrane helices. Today, cln5 is revealed to be a soluble glycoprotein (Schmiedt *et al.*, 2010). Also, soluble proteins with diverse apparent molecular weights ranging from 39 kDa to 47 kDa were reported from *CLN5* cDNA expression (Isosomppi *et al.*, 2002, Vesa and Peltonen, 2002).

Furthermore, cln5 has eight putative *N*-glycosylation sites with Asn-X-Thr/Ser consensus sequence, for both high-mannose and complex sugar modifications (Isosomppi *et al.*, 2002, Vesa and Peltonen, 2002). *N*-glycosylation is covalently linked to asparagine residues and has a pentasaccharide core consisting of two *N*-acetyl-D-glucosamine (NAG) and three D-mannose (MAN) residues. After the pentasaccharide follow either MAN residues in the high-mannose type or a mixture of sugar residues, e.g. galactose, fucose, or others are used to modify the protein in the complex-type modification (Berg *et al.*, 2012). Glycosylation sites are located at amino acid positions 179, 192, 227, 252, 304, 320, 330, and 401 (Moharir *et al.*, 2013). Glycosylated proteins expressed from BHK-21 and COS-1 cells revealed apparent molecular weights ranging from 60–75 kDa in sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE) .

Endo-β-*N*-acetylglycosaminidase H (EndoH) treatment resulted in cln5 forms of 40 kDa or 47 kDa and PNGaseF treatment yielded cln5 of 38 kDa or 45 kDa (Isosomppi *et al.*, 2002). EndoH is an endoglycosidase that cleaves high-mannose glycosylation from glycoproteins adjacent to the linkage with asparagine between two NAG residues (Maley *et al.*, 1989). PNGaseF is an endoglycosidase cleaving the protein-sugar bond between asparagine and NAG of most *N*-linked glycoproteins (Tarentino *et al.*, 1985). Site directed mutagenesis studies revealed that *N*-glycosylation displays functional differences affecting folding, trafficking, or lysosomal function in cln5 (Isosomppi *et al.*, 2002). Mutating Asn179, Asn252, Asn304, and Asn320 affects the folding of the protein and it is retained in the ER without glycosylation. The glycosylation at Asn401 is reported to play a role in endosome/lysosome trafficking. The *N*-glycosylation sites Asn192 and Asn227 are reportedly involved in lysosomal function (Qureshi *et al.*, 2018). Also, cln5 contains other potential modification sites including myristoylation and phosphorylation sites, which have not yet been experimentally analyzed (Savukoski *et al.*, 1998).

4.2.2. Protein localization

Cln5 mainly localizes to the endosomal-lysosomal compartment as shown by studies with confocal immunofluorescence microscopy in transiently transfected BHK-21 cells

(Isosomppi *et al.*, 2002). The polypeptide was reported to localize to the ER and has been found in the axons of neuronal cells (Holmberg *et al.*, 2004). Studies on embryonic and postnatal mouse brains showed abundant mRNA expression in developing cerebral cortex, cerebellum and in the ganglionic eminence (Holmberg *et al.*, 2004). In the adult mouse brain the most intense signal was reported in the Purkinje cell layer of the cerebellum, in the cerebral cortex, as well as in the hippocampal principal cell layers. Cln5 expression could also be found in a variety of tissues such as aorta, kidney, lung and pancreas (Savukoski *et al.*, 1998). Holmberg *et al.* (2004) reported that sub-cellular cln5 localizes predominantly to the neuronal soma.

Most soluble lysosomal proteins contain specific carbohydrate modifications, especially mannose-6-phosphate (Man-6-P) glycosylation, for recognition by the mannose-6-phosphate receptors (sMPRs). The glycosylation directs the targeting to the lysosomal compartment (Pohlmann *et al.*, 1995). Cln5 contains Man-6-P residues on high mannose-type oligosaccharides linked to Asn320, Asn330, and Asn401. Thereby it has been suggested that cln5 is trafficked by the MPR-dependent pathway (Kollmann *et al.*, 2005, Sleat *et al.*, 2006). This supports a soluble cln5 variant (Sleat *et al.*, 2006). Other studies suggested cln5 may be transported on Man-6-P independent trafficking route (Schmiedt *et al.*, 2010), and other routes (Kollmann *et al.*, 2005, Sleat *et al.*, 2009).

Furthermore, cln5 may undergo a retrograde transport step from the Golgi to the ER (Schmiedt *et al.*, 2010). It has been suggested earlier that cln5 undergoes cleavage prior to transport to the lysosome (Holmberg *et al.*, 2004, Isosomppi *et al.*, 2002, Vesa and Peltonen, 2002).

4.2.3. Proposed functions of cln5

Schmiedt *et al.* (2012) described the developmental regulation of the *Cln5* gene in mice. A role of cln5 in regulating microglial function is suggested by the highest expression observed in microglial cells within the brain combined with very early activation of microglia in *Cln5*$^{-/-}$ mice. Indicating the involvement of cln5 in lipid metabolism, alterations in serum lipid profiles, defective lipid transport and hypomyelination were observed in *Cln5*$^{-/-}$ mice (Schmiedt *et al.*, 2012).

Mamo *et al.* (2012) reported cln5 is essential for the recruitment of the multi-modular protein assembly retromer, a protein complex responsible for the sorting and recycling of lysosomal receptors (Bonifacino and Hurley, 2008). In HeLa cells over-expressing HA-tagged cln5 and myc-tagged sortilin, a lysosomal enzyme transporter, co-immunoprecipitate. The stability of sortilin and cation-independent mannose-6-phosphate receptor (CI-MPR) diminishes with cln5 depletion. Also CTSD levels increase upon cln5 depletion,

which was interpreted as plausible by Markmann *et al.* (2015) since alternative lysosomal targeting pathways for CTSD exist. The loss in function of CI-MPR transporter originates in defects in the recruitment of the retromer due to less loaded Rab7, which is required for this purpose (Mamo *et al.*, 2012). Mamo *et al.* (2012) suggested “cln5 is part of an endosomal switch that determines whether the lysosomal sorting receptors are recycled to the Golgi compartment or degraded in lysosomes.“ This was discussed as beeing inconsistent with cln5 properties as soluble lysosomal protein (Cárcel-Trullols *et al.*, 2015).

Very modest reduction in vacuolar protein sorting 26 (VPS26) and VPS35 was additionally reported. These proteins are part of the vacuolar protein sorting heterotrimer with cargo binding function in the retromer (Mamo *et al.*, 2012).

The retromer has been linked to the pathogenesis of late-onset Alzheimer’s disease (Muhammad *et al.*, 2008). Recently, a study reported that the mutant Asn320Ser of cln5 protein might mediate Alzheimer’s disease associated toxicity by affecting retromer function in microglia (Qureshi *et al.*, 2018). HeLa cells and mouse N2a cells expressing the mutated CLN5 variant displayed a decrease of full-length amyloid precursor protein (APP). This was interpreted as a possible defect in retromer-dependent trafficking.

Furthermore, cln5 may function as modulator of dihydroceramide synthase (CerS), of cell growth, and apoptosis. One study in fibroblasts from cln5 deficient patients found that sphingolipids downstream of CerS, ceramide and dihydroceramide, were diminished (Haddad *et al.*, 2012). It was proposed that cln5 and cln8 are functionally related, as cln8 corrects the cln5 defect and altogether modulates CerS 1 and 2 (Haddad *et al.*, 2012). No proof like, e.g. co-immunoprecipitation of cln5 and CerS or cln5 and cln8, neither in normal fibroblasts nor fibroblasts of cln5 deficient patients, was given.

Huber and Mathavarajah (2018) proposed that cln5 is a glycoside hydrolase based on studies of *dictyostelium discoideum* cln5 with a glycoside hydrolase activity assay (see chapter 4.2.5). Furthermore, the study names 61 proteins cln5 shows interaction with based on *dictyostelium* cln5-green fluorescent protein (GFP) immunoprecipitation and subsequent LC-MS/MS analysis.

Still, the exact function of cln5 remains unknown. The high relative expression level of cln5 in central nervous system neurons and microglial cells as well as defective myelination in the brain of *Cln5*$^{-/-}$ mice and late infantile CLN5 disease patients (Schmiedt *et al.*, 2012) points towards a distinct importance of cln5 function in the brain.

Interactions with other NCL proteins

The protein cln5 locates to the lysosomes, although it may also function elsewhere in the cell (Holmberg *et al.*, 2004, Schmiedt *et al.*, 2010). First reports of interactions with other NCL proteins by Sleat (1997) described altered tripeptidyl-peptidase 1 (TPP1) (cln2) levels in cln5 patients.

Based on co-immunoprecipitation and *in vitro* binding assays, Vesa and Peltonen (2002) reported direct interactions of cln5 with cln2 and cln3 proteins. While interactions of cln5 and cln3 persist with the 'Fin major' and 'European' variant mutations of cln5, they are abolished with cln2 for these mutations consistently (see Chapter 4.2.4). These results are suggesting the cln2 interaction domain is located *N*-terminal from amino acid 224, were the cln5 European mutation is located.

Lyly *et al.* (2009) could confirm interactions between the proteins cln1, cln2, and cln3 with cln5 via co-immunoprecipitation assays. The same study reported additional interactions between cln5 and cln6 as well as cln8. Over-expression of PPT1 (cln1) rescues the lysosomal trafficking defect of cln5 ($vLINCL_{FinMajor}$) in HeLa cells and human neuroblastoma (SH-SY5Y) cells. Normally retained in the ER and Golgi, although this mutant cln5 protein is ER exit-competent, it now co-localizes with PPT1 to the lysosomes (Lyly *et al.*, 2009). This suggests the interaction occurs in the ER, where the interactions with the ER resident NCL proteins cln6 and cln8 would occur. The same study reported that cln5 interacts with subunits of F_1-ATPase as does PPT1 (Lyly *et al.*, 2008).

Recently a *CLN7* defect was reported to lead to depletion of mature cln5 due to increased proteolytic degradation by cysteine proteases (Danyukova *et al.*, 2018). Also an Alzheimer's disease associated *CLN5* mutation was reported to hinder the maturation of proCathepsin D to mature CTSD (Qureshi *et al.*, 2018).

The interaction of cln5 with other proteins associated with NCLs suggests a common pathway in neuronal cell metabolism. Nonetheless the manner and function of the interactions is not yet understood.

4.2.4. Mutations

Thirty-seven different mutations and nine sequence variations in *CLN5* are identified in patients of Finnish and non-Finnish origin as of September 2018 recorded in the NCL mutation database[4] (Bessa *et al.*, 2006, Cismondi *et al.*, 2008, Holmberg *et al.*, 2000, Lebrun *et al.*, 2009, Pineda-Trujillo *et al.*, 2005, Savukoski *et al.*, 1998, Sleat *et al.*, 2009, Xin *et al.*, 2010). These consist of nineteen missense mutations, twelve small deletions

[4]University College London, *https://www.ucl.ac.uk/ncl-disease/mutation-and-patient-database*.

or insertions causing frame-shift, one large deletion, and nine nonsense mutations. Seven cDNA point mutations are known that do not cause an amino acid change. Tables 4.2 and 4.3 provide an overview of all disease causing mutations while mutations with no effect on the amino acid sequence are not listed (known by the identifiers cln5.x with x = 019, 023, 028, 025, 026, 037). Also heterozygote mutations are not included in the tables.

Table 4.2.: NCL mutation database excerpt for known pathogenic *CLN5* mutations involving a frame shift (insertion or deletion) or premature stop codon.

identifier	cDNA exchange	amino acid change	NCL phenotype
cln5.001	c.1175_1176delAT	p.(Tyr392X)	late infantile
cln5.002	c.225G>A	p.(Trp75X)	late infatile
cln5.005	c.669dupC	p.(Trp224LeufsX30)	juvenile
cln5.007	c.565C>T	p.(Gln189X)	late infantile
cln5.014	c.1072_1073delTT	p.(Leu358AlafsX4)	juvenile
cln5.016	c.1026C>A	p.(Tyr342X)	late infantile
cln5.017	c.1054G>T	p.(Glu352X)	late infantile
cln5.019	c.291dupC	p.(Ser98LeufsX13)	late infantile
cln5.020	c.1103_1106delAACA	p.(Lys368SerfsX15)	late infantile
cln5.027	c.527_528insA	p.(Gly177TrpfsX10)	late infantile
cln5.033	c.919delA	p.(Arg307GlufsX29)	juvenile
cln5.035	c.1038delT	p.(Phe361LeufsX4)	juvenile
cln5.048	c.741_747delinsTT	p.(Trp247CysfsX5)	late infantile

The most common mutation is the 'Fin major' mutation, a 2 base pair deletion in exon 4 (cln5.001, c.1175delAT) resulting in the change p.Tyr392X and a truncated polypeptide of 391 amino acids. This is the only mutation exclusively identified in families of Finnish descent. All other mutations have been found in families of diverse descent in many different countries including Finland, UK, Afghanistan, Argentina, Canada, China, Colombia, the Czech Republic, Egypt, Italy, the Netherlands, Portugal, Pakistan, Sweden, and USA (Bessa *et al.*, 2006, Cannelli *et al.*, 2007, Cismondi *et al.*, 2008, Holmberg *et al.*, 2004, Kousi *et al.*, 2009, Lebrun *et al.*, 2009, Pineda-Trujillo *et al.*, 2005, Sleat *et al.*, 2009, Xin *et al.*, 2010). Today, mutations in the *CLN5* gene are reported in families of diverse origin such as Asian, Hispanic and Arabian origin, rendering the naming as Finnish variant obsolete.

Second most common mutation is "Fin minor', where a G to A transversion in exon 1 (cln5.002, c.225G>A) causes the substitution p.Tyr75X, resulting in a polypeptide of only 74 amino acids. It was also found in a Swedish family with an additional C insertion at amino acid 223 (cln5.005, c.669insC) causing a frameshift and premature stop codon at position 253 (predicted) (Holmberg *et al.*, 2000).

Table 4.3.: NCL mutation database excerpt for known pathogenic *CLN5* mutations involving a point mutation resulting in a changed residue.

identifier	cDNA exchange	amino acid change	NCL phenotype
cln5.003	c.835G>A	p.(Asp279Asn)	late infantile
cln5.006	c.335G>A	p.(Arg112His)	juvenile
cln5.011	c.772T>G	p.(Tyr258Asp)	juvenile
cln5.012	c.4C>T	p.(Arg2Cys)	late infantile
cln5.018	c.1137G>T	p.(Trp379Cys)	late infantile
cln5.021	c.377G>A	p.(Cys126Tyr)	adult
cln5.022	c.1121A>G	p.(Tyr374Cys)	adult
cln5.029	c.575A>G	p.(Asn192Ser)	juvenile
cln5.030	c.620G>C	p.(Trp207Ser)	juvenile
cln5.044	c.613C>T	p.(Pro205Ser)	late infantile

Several polymorphisms, genetic variants at a locus within the population, are known so far. One mutation in exon 4, (c.1103A>G, cln5.004), leading to a mutation of the polypeptide of p.Lys368Arg, has a carrier frequency of 20% in the Finnish population and about 10% in USA (Kousi *et al.*, 2012, Savukoski *et al.*, 1998, Xin *et al.*, 2010). Also displaying a carrier frequency of 10% in the USA is the mutation cln5.012 in exon 1 (c.4G>T, p.(Arg2Cys)). The mutation cln5.008 (c.335G>C, p.(Arg112Pro)) might also be a polymorphism as it was only found on allele carrying c.835G>A, p.(Asp279Asn) (cln5.003).

4.2.5. Protein structure prediction

Cln5 was reported as a soluble protein in studies using Triton X-114 solubility assays on transfected COS-1 cells. These were consistent with predictions using TMHMM2.0 and SOSUI software (Holmberg *et al.*, 2004, Schmiedt *et al.*, 2010).

Earlier, cln5 was reported to have two transmembrane domains predicted with BCM transmembrane prediction program (Savukoski *et al.*, 1998) or to exist in either soluble

and transmembrane forms depending on the starting methionine used during expression (Vesa and Peltonen, 2002).

While comparing *dictoystelium* cln5 and *human* cln5 sequences, Huber and Mathavarajah (2018) proposed that cln5 may function as glycoside hydrolase (see Figure 4.2). Based on the results of the online web server RaptorX, a three-dimensional structure model was predicted to have a putative binding site for *N*-acetylglucosamine thiazoline (NGT). This substrate also binds to β-hexosaminidase, a glycoside hydrolase linked to Tay-Sachs and Sandhoff disease (Mark *et al.*, 2003).

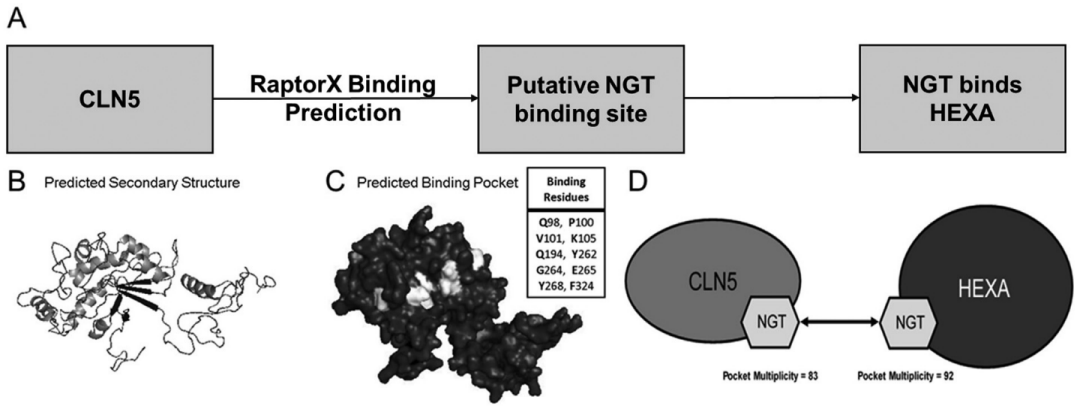


Figure 4.2.: Secondary structure prediction and consequently reported NGT binding pocket for cln5, picture taken from Huber and Mathavarajah (2018).

4.3. Aim of this work

The function and interaction of NCLs proteins is currently the focus of intensive research. The crystal structure and function of most of the proteins involved is already known, e.g. TPP1, PPT1, CLN3 TM protein or CTSD. Nonetheless, the role of the proteins in neurodegeneration and the interaction with other proteins remains elusive. The protein associated with the CLN5 gene has not been characterized functionally before and no conclusive structural information was available. Interactions of cln5 with the NCL proteins TPP1, PPT1, or CLN3 TM protein have been reported (Lyly *et al.*, 2009, Sleat, 1997, Vesa and Peltonen, 2002). In addition, various studies reported an influence of cln5 in regulation of microglial function (Schmiedt *et al.*, 2012), lipid metabolism, or lysosomal sorting (Bonifacino and Hurley, 2008). The exact role of cln5 has not been unambiguously identified.

The aim of this work is the study of cln5 structure and interactions. Initial analytical studies of the protein by CD spectroscopy in solution are presented. Moreover, the

analysis of the protein sequence for secondary and tertiary structure prediction with various web services is discussed. Additionally, interaction studies are introduced, giving insight to possible interactions of cln5 with other NCL proteins and proteins associated with cell apoptosis.

In this work the crystal structure of cln5 is presented for the first time. Furthermore, a structural relationship of cln5 with proteins of the NlcP/P60 superfamily (Xu *et al.*, 2011) can be established based on the study of the crystal structure. The structural similarity of cln5 to the protein PPPDE1 is a valuable clue to function, as the central 3D fold, including a reactive center, remains conserved. PPPDE1 was reported to function as deSUMOylating protein (Shin *et al.*, 2012) and deubiquitinating protein (Iyer *et al.*, 2004, Xie *et al.*, 2017). The implications of this structural relationship for cln5 are discussed. The structure can be the crucial factor to determine the function of cln5 and can help to finally understand the role of cln5 in neurodegeneration. Based on this structure, more focused studies of the function of cln5 are possible and are carried out at the moment. The here presented work can be the key to finally assigning a function to cln5 and to place it in the greater context of NCL diseases and apoptosis.

4.4. Materials and Methods

4.4.1. Protein structure prediction

At the beginning of the crystallization studies web services for secondary structure prediction were used to better understand cln5. Some prior knowledge of a possible structure or structure elements can help in the choice of crystallization methods and screens. The web based services of the structure prediction servers I-TASSER (Yang *et al.*, 2015, Yang and Zhang, 2015), BLAST (Altschul *et al.*, 1990), Swiss Model (Waterhouse *et al.*, 2018), HHPred (Soding *et al.*, 2005) and RaptorX (Wang *et al.*, 2011) were employed with the protein sequence of cln5.

Two main approaches to protein structure prediction find wide application: template based modeling (TBM) (homology modeling) and *ab initio* prediction.

Homology modeling is possible when likely homologues of the protein exist and is based on database searches of known structures for sequence similarity with the query sequence. The best match is picked as template, establishing the basis to build a model. For different parts of the protein different templates can be used and refined. This procedure is called TBM (Karplus *et al.*, 1998, Šali and Blundell, 1993). Sequence similarity can be measured via sequence identity, amino acids matching exactly in both

sequences, or in similarity, amino acids having the analogous properties, such as polarity or functional groups. Profiles of the sequence can be created, describing for example the probability of mutation of each amino acid, and used to compare two sequences. This can be implemented by using hidden Markov models (sHMMs).

Ab initio modeling is used when no structural homologues can be identified and employs knowledge based energy functions (Klepeis *et al.*, 2004). A model must be built from scratch by understanding the physiochemical principles of how proteins fold in nature. The modeling algorithms have three key concepts: energy function design, conformational search and model selection.

I-TASSER uses a hierarchical approach, identifying structural templates from the PDB using LOMETS, a meta-threading server for target-to-templates assignment. Functional insights of the target are derived using the database BioLiP for function prediction (Yang *et al.*, 2015, Yang and Zhang, 2015).

BLAST compares protein or nucleotide sequences to sequence databases and reports statistically significant matches (Altschul *et al.*, 1990). Statistical significant results can be used to infer functional or evolutionary relationships.

Swiss Model Homology web service builds protein homology structures relying on ProMod3, extracting structural information from templates (Waterhouse *et al.*, 2018).

HHPred is a server for protein homology detection, structure and function prediction using HMMs. Various databases are searched for local or global alignment based on the query sequence and can produce 3D structural model from the search result (Soding *et al.*, 2005).

RaptorX is a template-based prediction server for proteins without sequence homologues in the protein database (PDB). It predicts secondary and tertiary structures and from these structures contacts, solvent accessibility, disordered regions and binding sites (Wang *et al.*, 2011).

Furthermore, preliminary circular dichroism measurements were used to determine secondary structure elements and protein integrity (see Section 4.4.3).

4.4.2. Protein preparation

All protein samples were obtained from the laboratory of Prof. R. Steinfeld, Department of Paediatrics and Paediatric Neurology, UMG Georg-August University Göttingen, provided by Dr. R. Krätzner, K. Schreiber and M. Ziegenbein. To facilitate purification, human cln5 was modified with a short linker and a HIS_6-tag (RSHHHHHH) at the *C*-terminus. Transfection of the modified CLN5 and selection of stable HEK293 cell lines were performed as described before by Steinfeld *et al.* (2004).

Additionally, kifunensine was added to inhibit α-mannosidase, resulting in hyper-mannosylation of the glycosylation sites (cln5-k). Afterwards, cln5-k was treated with EndoH to cleave the hybrid oligosaccharides from *N*-linked glycosylation sites leaving one *N*-acetyl-D-glucosamine (cln5-e). In a separate cell culture HEK293 cells, in addition to kifunensine, were treated with selenomethionine (SeMet) (MSE in 3-letter amino acid nomenclature) replacing all methionine residues (cln5-k-Se). Recombinant protein samples were purified from the cell-culture supernatant, which was kept frozen until purification.

After thawing, the supernatant was filtered through a membrane (pore size 0.22 μm Corning PES filter) under water-jet vacuum. After addition of 20 mM K_2HPO_4 pH 7.5, 0.5 M NaCl and 40 mM imidazole the crude solution was loaded onto a 5 mL HisTrap^{TM} FF Ni affinity column (flow rate 1 mL/min, GE Healthcare). Protein elution with a step-wise gradient resulted in two peaks (see Figure C.1). The fractions of the two peaks were separately pooled and concentrated in centrifugal filters (Amicon Ultra, size exclusion 10 000) at 4 °C, 3000x rpm. The protein was exchanged into 10 mM K_2HPO_4 pH 7.2, 50 mM NaCl and concentrated to a final concentration of approximately 10 mg/mL. The final concentration was determined via bicinchoninic acid (BCA) assay from a theoretical extinction coefficient, 99320 $M^{-1}cm^{-1}$, based on the protein sequence. Purified protein was kept on ice and protein purity and integrity was confirmed via SDS-PAGE according to Laemmli (Laemmli, 1970) on aliquots from before and after purification and buffer exchange. Successful purification resulted in a band in SDS-PAGE of 60 kDa, 55 kDa, and 37 kDa for cln5, cln5-k/cln5-k-Se, and cln5-e, respectively (see Section C.2).

Interaction of cln5 with other NCL proteins

All interaction studies were performed in the lab of Prof. R. Steinfeld at the UMG Göttingen. The cln5 over-expression in stable HEK293 cells showed an influence on cln10 (CTSD) and LC3-I/II (studies performed by A. Wolf[5]).

Here the influence of cln5 depletion on NCL proteins and autophagy markers, LC3-II and p62 is evaluated. LC3 in the lipidated form, LC3-II, has been shown to be an autophagosomal marker acting in autophagosome formation. LC3-I represents the cytosylic form and LC3-II is the membrane-bound form of LC3 (Kabeya, 2000). It is therefore used to study autophagy in neurodegenerative disorders (Tanida *et al.*, 2004). The autophagy substrate p62 can bind LC-3 and is degraded by autophagy (Eskelinen and Saftig, 2009).

[5]Department of Paediatrics and Paediatric Neurology, UMG Georg-August University Göttingen.

The cell lysate from stable HeLa *CLN5*$^{-/-}$ cells was separated by 10%, 12% or 18% SDS polyacrylamide gel electrophoresis (40 µg protein per lane) and transferred by electroblotting to nitrocellulose membranes (Protran, Schleicher and Schüll, Germany) according to standard protocols. The blot was blocked with phosphate-buffered saline (PBS) containing 5% milk powder and incubated with rabbit or mouse anti-antibodies (1:1000 in blocking solution). The blot was then washed twice in PBS containing 0.05% Tween 20, incubated with a horseradish peroxidase-conjugated anti-rabbit or anti-mouse IgG (1:10000 in blocking solution; Dianova, Hamburg, Germany) and washed as before. Signals were visualized using the enhanced chemiluminescence (ECL) technique. Detection of the control protein β-actin was performed in the same manner after removal of the anti-bodies with 200 mM glycin/HCl pH2.2, 3.5 mM SDS), 1% Tween 20. All western blot studies were repeated twice.

Investigation of cln5 processing by CathepsinD

CTSD, the protein of CLN10 disease, is an aspartyl endopeptidase with numerous substrates in the lysosome (Benes *et al.*, 2008). Additionally to protein degradation, CTSD has been linked with apoptosis and lipid homeostasis (Getty and Pearce, 2011). The role of CTSD in the NCL pathways is not clear and interactions with other NCL proteins are possible.

Here, the potential processing of cln5 by CTSD is studied.

Frozen CTSD stored in phosphate-buffered saline was taken up in 100 mM sodium acetate buffer pH 3.8 to achieve a concentration of 1 µg/µL for activity studies and 3 µg/µL for all other studies. CTSD was incubated at 37 °C for ten minutes with 50 mM sodium fumerate pH 3.0 or 100 mM sodium acetate buffer pH 3.8 to obtain activate CTSD.

For kinetic studies, 0.5 µg of cellular lysates were incubated with 1.3, 2, 3.3, 6.7, 10, 13.3, 23.3, 33.3, and 45 µM of substrate peptide MOCAc-GKPIIFFRLK(Dnp)-R-NH_2 (Yasuda *et al.*, 1999). Reaction velocities were measured at defined time intervals with a microplate reader (Synergy HT [BIO-TEK Instruments]). Linear phase velocities were plotted against the substrate concentrations; K_m and V_{max} were calculated by nonlinear regression using the equation

$$V = \frac{V_{\text{max}} \cdot [S]}{K_{\text{m}} + [S]} \tag{4.1}$$

where V is linear phase velocity, V_{max} is maximal enzyme velocity, $[S]$ is the substrate concentration, and K_{m} is the Michaelis-Menten constant.

For interaction studies 50 µL reaction solutions were prepared containing 500 ng,

1.25 μg or 2.5 μg active CTSD with 2.5 μg cln5, resulting in ratios of 1:5, 1:2 and 1:1 CTSD:cln5 concentrations in PBS buffer pH 4.5. At time intervals of 0 min, 10 min, 30 min, and 1 h sample of the reaction solution were withdrawn and added to pre-prepared SDS-PAGE solutions to stop any reaction. The reaction solutions were separated via SDS-PAGE and analyzed by western blot.

4.4.3. Circular dichroism

Circular dichroism (CD) can be used to estimate the secondary structure of proteins or peptides (Greenfield, 2006). It is characterized by the unequal absorption of left-handed and right-handed circularly-polarized light. If the light is polarized by passing through suitable prisms or filters, its electric field will oscillate sinusoidal in a single plane.

CD spectra were measured on a *Jasco* J-810A spectropolarimeter with temperature control *Jasco* PTC423S. The sample chamber was flushed with a nitrogen stream. Measurement parameters were employed as listed in Table 4.4.

Table 4.4.: CD measurement parameters.

data mode	CD and absorption
absorption band width	1.0 nm
response	1.0 s
data pitch	0.1 nm
scanning speed	50 nm
cuvette length	0.1 cm
sensitivity	high

The CD spectra were adjusted by blank measurement to correct for solvent influence. Five measurements were measured in a wavelength range from 300 to 190 nm. The resulting spectra were smoothed with the means movement method (convolution width 25). With the formula 4.2, the measured ellipticity (Θ_{abs} [mdeg]) was converted to mean residue ellipticity Θ [10^{-3} deg cm^2 dmol^{-1}].

$$\Theta = \left(\frac{\Theta_{\text{abs}}\,[\text{mdeg}] \cdot 0.001}{c\,[\text{mol/L}] \cdot l\,[\text{cm}] \cdot n \cdot 10} \right) \tag{4.2}$$

With: c [mol/L] = concentration of protein; l [cm] = length of cuvette; n = number of amino acids in the peptide.

4.4.4. Crystallization

A large variety of crystallization conditions were screened via sitting drop method on 96 well plates (Axygen, Corning Incorporated, Corning, USA). Reservoir solutions of 100 µL were pipetted employing a TECAN Genesis Robot. The drops of 0.2 µL consisting of 1:1 (v:v) protein and reservoir solution were set by a TTP Labtech Mosquito. Hanging drop experiments were set up in VDXm 24 well plates (Hampton Research, Aliso Vieji, USA) containing 1 mL reservoir solution and 2-4 µL protein solution with varying well-to-reservoir ratio. Commercially available crystallization screens were used for initial screening (Hampton Research: Index Screen HT, Crystal Screen HT, PEG/Ion HT, Natrix HT; Rigaku Reagents: Wizard™HT, Wizard™3+4; Qiagen®: MPD Suite, PEGs Suite, JCSG+; Jena Bioscience: XP Screen, Pentaerythritol, Pi-Minimal HTS, Pact++). The XP screen contains the additive Anderson-Evans polyoxotungstate $[TeW_6O_{24}]^{6-}$ (TEW), allegedly not only promoting protein crystallization but also diffraction quality (Bijelic and Rompel, 2017). This screen was employed to facilitate anomalous phasing using the intrinsic tellurium without additional HA soaking experiments. In addition, variation of temperature (4, 10 and 20 °C), protein-to-reservoir ratios (1:1, 2:1, 1:2 (v:v)), and protein concentration (4 to 12 mg/mL) were implemented. Furthermore, under-oil crystallization, streak seeding, micro-seed seeding with serial dilution and micro-seed matrix screening (MMS) were carried out (D'Arcy *et al.*, 2014). Co-crystallization with Polyvalan Crystallophore No 1 (Tb-Xo4) (Molecular Dimensions, Suffolk, UK) was attempted by addition of 10 mM Tb-Xo4 to the protein solution before crystallization. The compound allows convenient phasing using the anomalous signal of the lanthanide terbium (Engilberge *et al.*, 2017). Initial findings were reproduced on 96 and 24 well plates using both sitting and hanging drop methods as well as seeding methods. In-drop validation of the latest crystals was performed using an Jansi UVEX UV microscope.

4.4.5. Crystals

cln5 native

One single crystal could be obtained from 20% (w/v) polyethylene glycol (PEG) 3350, 0.2 M potassium chloride at 20 °C after eight months. The crystal was transferred in a step-wise manner into a crystallization solution containing additionally up to 12% glycerol as cryoprotectant and flash cooled in liquid nitrogen for storage. All attempts to obtain more crystals from this condition employing grid screening and seeding methods were unsuccessful.

cln5-kifunensine

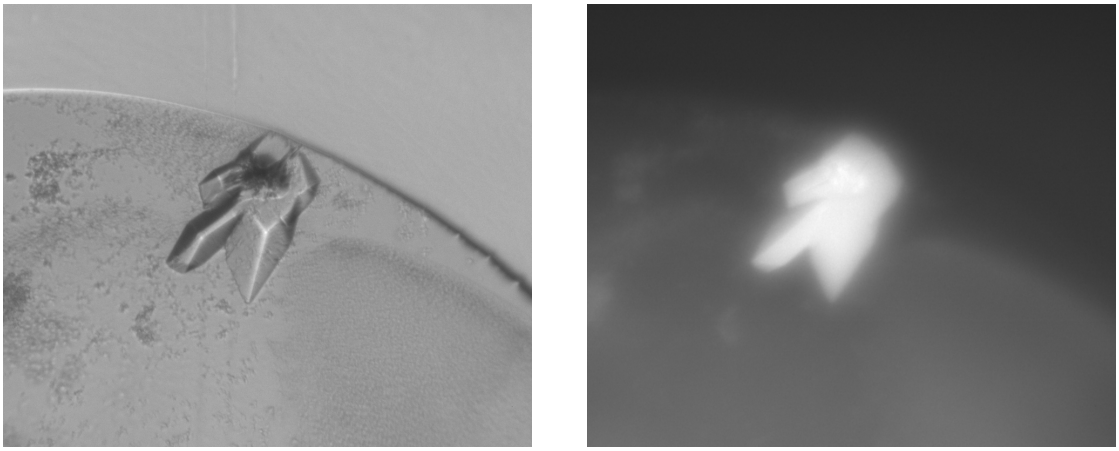
First screenings resulted in a single crystal from 30% (v/v) pentaerythritol ethoxylate (PE) 15/4, 50 mM BisTris pH 6.5, 50 mM ammonium sulfate (condition A) after seven weeks at 20 °C. Attempts at reproducing the crystallization were successful with reservoir solutions from 25-31% (v/v) PE 15/4, 50 mM BisTris pH 6.1-6.5, 50 mM ammonium sulfate, resulting in multiple crystals of different quality and sizes (Table 4.5). Furthermore, individual crystals could be obtained, but not reproduced, from the following conditions:

- 20% PEG 3350 (w/v), 200 mM magnesium nitrate at 20 °C after six months (condition B).
- 20% PEG 6000 (w/v), 100 mM HEPES pH 7.0, 200 mM magnesium chloride at 20 °C after seven months (condition C).
- 20% PEG monomethylether (PEG MME) 5000 (v/v), 0.1 mM BisTris pH 6.5 at 20 °C after four months (condition D).

In general the crystal size was 20-30 μm with some exceptions showing a size in one direction of up to 70-80 μm. Crystals displayed various shapes, from diamond-shaped to needles, plates and ingot-like. The precipitant from condition A, PE 15/4, was reported to be sufficient for cryo-protection, starting with a concentration of 25% (v/v) (Gulick *et al.*, 2002). Crystals from condition B were transferred into cryoprotectant solution containing 13% glycerol (v/v) in addition to 20% PEG 3350 (w/v). Crystallization solutions of 20% PEG 6000 and 20% PEG MME were prepared with an additional 20% glycerol for cryo-protection. Transfer to the final cryoprotectant solution was performed in a step-wise manner. All crystals were stored in liquid nitrogen until measurement.

All attempts to grown crystals of better quality or larger size using various seeding techniques in addition to optimization were unsuccessful except for one single crystal (see Figure 4.3b). Random sparse matrix MMS yielded one crystal from 20% PEG 1000 (v/v), 200 mM calcium acetate, 100 mM imidazole pH 8.0 with seeds originating from condition A.

Heavy atom soaking experiments Multiple soaking experiments were conducted to facilitate anomalous phasing for structure solution. Soaking solutions were prepared by recreating the crystallization condition with the heavy atom compound added. Four crystals were soaked in 100 μM Polyvalan for 2–5 min. Five crystals each were transferred to 500 mM sodium iodide or sodium bromide, respectively, and soaked for time ranges from 30 sec to 7 min. Also, 20 μM 4-bromopyrazole soaks were performed with



(a) picture from light microscope.

(b) picture from UV microscope, fluorescence 0.1 s.

Figure 4.3.: Crystal of cln5-kifunensine after 135 days obtained from optimization of condition A.

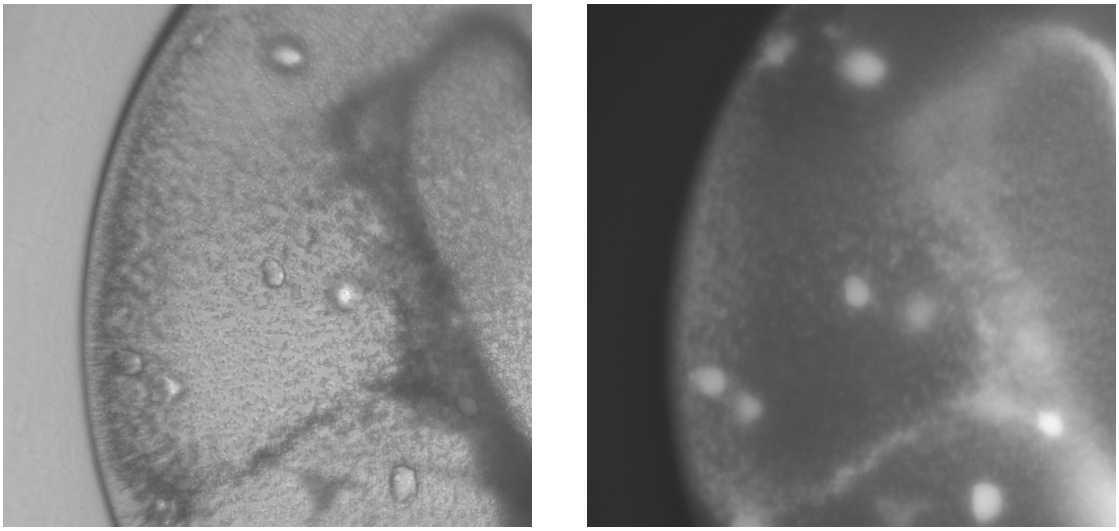
two crystals for 30 min and 1:30 h. Finally, cln5-k-Se crystals were soaked in 1 mM methylmercury chloride (MMCl), 1 mM *p*-chloromercuribenzene sulphonate (PCMBS), 10 µM ethylmercurithiosalicylate (EMTS, Thiomersal) and 10 mM trimethyllead acetate (TMLA) for 1 to 5 d. All crystals were transferred to cryoprotectant, flash-cooled and stored in liquid nitrogen until measurement.

cln5-kifunensine-EndoH

No crystals could be obtained from the cln5 protein variant that was treated with kifunensine and subsequently de-glycosylated with EndoH (cln5-e).

cln5-kifunensine-Se-methionine

Due to the low amount of the SeMet-labeled protein available, the protein was used in sparse matrix screens and in grid screen conditions derived from previous successes with the native kifunensine variant (conditions A – D). After 12 weeks crystals were observed in conditions ranging from 22.3 – 25.0% PEG 3350 (w/v), 0.179 – 0.193 M magnesium nitrate (grid screen from condition B) (see Figure 4.4b). These crystals were transferred in a step-wise manner to a cryoprotectant solution containing either 5% glycerol in addition to the crystallization solution or an increased concentration of the precipitant, 35% PEG 3350 (w/v). The crystals were flash-frozen and stored in liquid nitrogen until measurement.



(a) picture from light microscope. (b) picture from UV microscope, fluorescence 0.1 s.

Figure 4.4.: Crystals of cln5-kifunensine-SeMet after 85 days in condition B.

4.4.6. Data collection and processing

cln5-kifunensine

X-ray diffraction (XRD) data sets from all crystals were collected on a Dectris Pilatus 6M detector at the Swiss Light Source (SLS) undulator beamline PXII (X10SA) using monochromatic radiation. Native data collection wavelength was 1.00 Å for non-derivate crystals. Single-wavelength anomalous diffraction (SAD) data from derivate crystals were collected at the peak (peak of f") or inflection (negative peak of f') wavelength of the anomalous signal obtained from the plot of the anomalous scattering factors against the wavelength with an SLS in-house beamline software tool. An on-beamline fluorescence scan was conducted to verify the presence of anomalous signal before each measurement.

Overall only 5% of all crystals selected for XRD diffraction showed any signal beyond 4 Å (see Table 4.5 and Figure 4.6a). None of the soaking experiments yielded any significant anomalous diffraction. The data sets displayed high anisotropy upon merging. Four data sets were collected from crystal cln5k-79. Two of those were collected at low energy wavelength to facilitate S-SAD phasing. No sufficient anomalous signal could be collected.

Table 4.5.: List of best diffracting crystals obtained from cln5-kifunensine.

crystal number	cryst. cond.	space group	unit cell [Å]	diffraction limit[a] [Å]	ISa[b]	notes
cln5k-6	A	80 (97)	a=129, c=52	3.25	16.34	4-bromopyrazole soak
cln5k-26	A	80 (97)	a=128, c=50	4.60	10.34	NaI soak
cln5k-31	A	(97)	a=129, c=52	3.85	8.80	NaBr soak
cln5k-79	A	16	a=65, b=68, c=82	2.75	15.97	four data sets were collected
cln5k-83	A	97	a=128, c=148	3.60	15.74	TbXo4 soak

[a] diffraction limit given by $I/\sigma(I) \geq 2$.

[b] calculated by XDS.

cln5-kifunensine-SeMet

XRD data sets from all crystals were collected on a Dectris Pilatus 6M detector at the Swiss Light Source (SLS) undulator beamline PXII (X10SA) using monochromatic radiation. An on-beamline fluorescence scan was performed to verify the presence of Se anomalous signal (see Section C.5) and data were collected at the inflection wavelength (Figure 4.5).

Four data sets were collected each from the two best diffracting crystals (cln5k-Se-8 and cln5k-Se-11) and processed individually. Data collection statistics for all individual scans are summarized in Tables C.2 and C.3 (Appendix Section C.5) and for the merged data set, from which the structure was solved, are given in Table 4.6. Due to erroneous calculations of expected radiation damage while planing the experiment on beamline[6], the initial data set of each crystal was only measured with a 90 degree rotation about phi. Fortunately, the protein crystallized in the high symmetry space group P 3_221 and only 1/12 coverage of R* was required for a full set of unique reflections. Data were reduced and integrated using the XDS software suite by Kabsch (2010) as described in Section 1.3.

[6]The on-beamline software GUI DA+ and its the extension for MAD data collection is capable of calculating the expected radiation exposure. From this exposure a maximum radiation dose is calculated and experimental parameters are suggested.

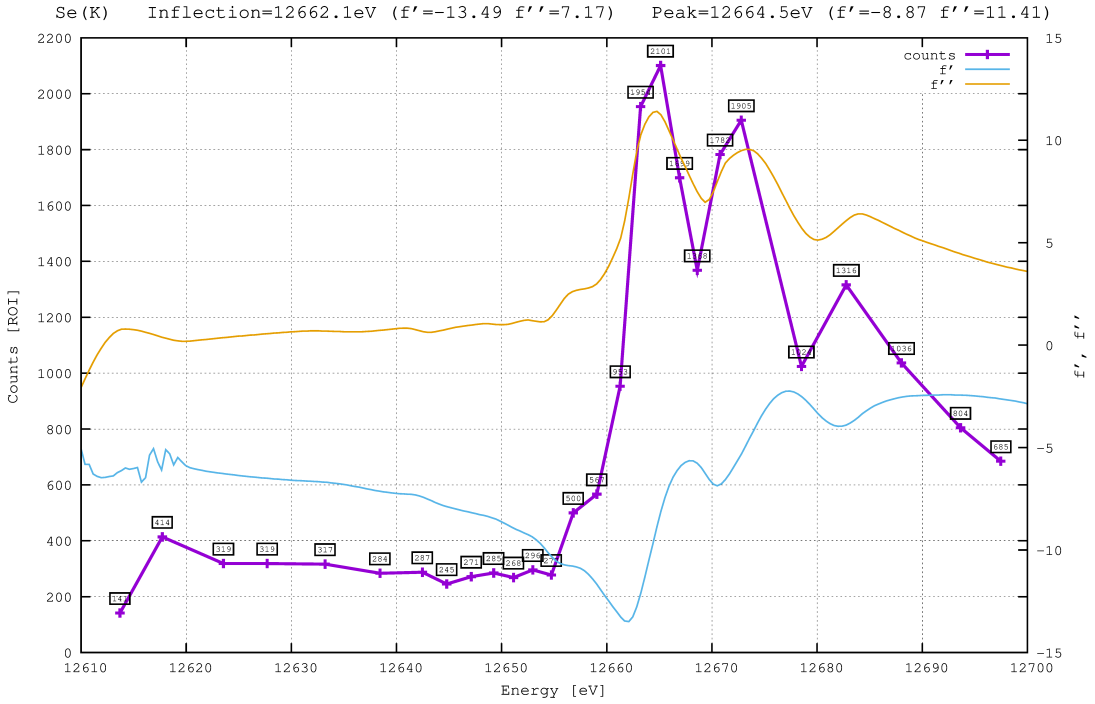
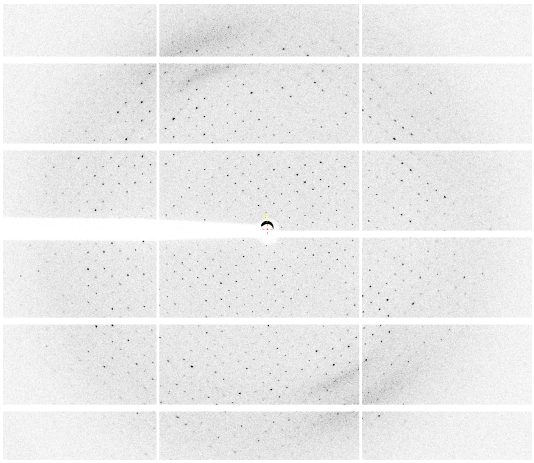
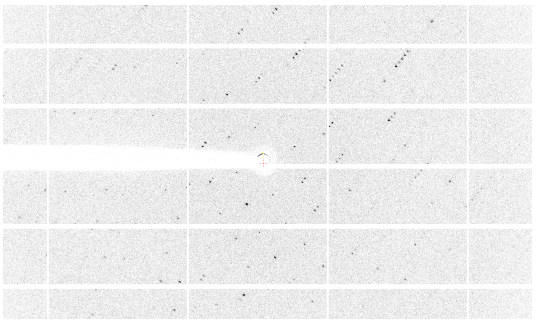


Figure 4.5.: On-beamline fluorescence analysis to identify the inflection point of the selenium signal. Only the region of interest as determined by the fluorescence scan, displayed in Figure C.5, is depicted. The anomalous scattering factors f' and f" are shown in blue and yellow, respectively.



(a) Diffraction image of a cln5-k crystal with reflections visible up to a resolution of 2.6 Å.



(b) Diffraction image of a cln5k-SeMet crystal with reflections visible up to a resolution of 3.3 Å.

Figure 4.6.: Diffraction images of cln5 crystals measured at the PSI SLS beamline PII X10SA at 100 K. Data collection with Pilatus 6M detector at 1.00 Å for the crystal in Figure 4.6a and Se inflection wavelength for the crystal in Figure 4.6b.

Table 4.6.: Statistics for merged data set used for data solution and refinement.

data statistics	cln5k-Se-merged
X-ray source	PSI SLS PII-X10SA
detector	Pilatus 6M (25 Hz)
wavelength [Å]	0.97898
space group	$P3_221$ (SG 154)
unit cell [Å]	a = 58.42
	c = 179.05
resolution range[a] [Å]	48.59–2.70 (2.79–2.70)
number of observations	540369
number of unique reflections[a]	10352 (1013)
redundancy[a]	52.2 (19.0)
completeness[a] all [%]	99.94 (100)
completeness[a] anomalous [%]	100 (100)
mean I/σ(I)[a]	19.08 (2.35)
$CC_{1/2}$ [a]	99.9 (83.9)
CC*[a]	100 (95.5)
R_{merge}[a]	0.2343 (1.255)
R_{pim}[a]	0.0305 (0.2873)

[a] highest resolution shell in parenthesis.

The crystals showed nearly no anisotropy, the resolution limits along the reciprocal space axis displayed a maximum difference of 0.069 Å (along a* 2.666 Å, along b* 2.666 Å, along c* 2.597 Å). Neither L-test nor the multivariate Z-score show signs of twinning in the CCP4 suite program TRUNCATE (Winn *et al.*, 2011) or PHENIX_xtriage (Adams *et al.*, 2010).

Enhancing the anomalous signal Further processing was accomplished using PHENIX_scale_and_merge (Adams *et al.*, 2010), XPREP[7] and XSCALE (Kabsch, 2010). Additionally, an experimental version of SHELXC (Sheldrick, 2010) was used to combine all data sets. This experimental version is capable of merging multiple data sets applying a weight to each scan to achieve the best possible anomalous signal and overall data set.

[7] *G.M. Sheldrick*, Bruker AXS Inc., Madison, Wisconsin, USA, 2003.

A more detailed account of the techniques used is available in Section 1.4 and Chapter 2.

XSCALE was employed making use of the option for zero dose extrapolation assessing and correcting for radiation damage. The output file provides a wealth of information to evaluate the quality of the data set combination. Parameters such as ISa for the whole data set and $CC_{1/2}$, mean I/σ(I), CC_{anom}, R-value or anomalous density by resolution were reviewed. Special care had to be taken to assure all data sets were integrated without influence of indexing ambiguity, since XSCALE cannot recognize or resolve the ambiguity.

PHENIX_scale_and_merge accepts a reference data set or can use the best data set as reference. The later was used for all files. The graphical interface of Phenix displayed a single table as result for the merged file. As suggested by the developers, the criteria $CC_{\text{cumulative}}$ at the lowest listed resolution in the statistics was used as initial quality indicator (see Graph C.6).

XPREP displays multiple quality indicators for merged and unmerged data sets, including $CC_{1/2}$ graphs, all common R-factors, and general data statistics (e.g. number of reflections, mean I/σ(I)). Futhermore d"/σ(d") statistics for anomalous data were evaluated.

Multiple combinations of the eight scans from two crystals were tested to obtain a merged data set with the best anomalous signal. The CFOM (combined figure of merit) value of a successful scan of the data set with SHELXD (Schneider and Sheldrick, 2002) and the averaged anomalous signal calculated from ANODE (Thorn and Sheldrick, 2011) were used to evaluate the quality of a merged data set.

Data quality

All data integration was repeated to optimize the resulting data quality and anomalous signal as discussed in Chapter 2.

All diffraction images were integrated to reach a $CC_{1/2}$ of 10%, a cutoff was employed in consecutive steps of data merging, solution and integration according to quality indicators. Typical resolution limits for all collected data sets is listed in Table C.4. The data statistics are presented for all single data sets to a $CC_{1/2}$ of 50% generated with PHENIX. ISa for the data sets as calculated by XDS and the results are displayed in Figure 4.7.

Since the collected data were rather weak and the crystals diffracted only to a resolution of 2.5–2.9 Å, special care was taken in further processing of the integrated data. As only a total of eight data sets from two crystals were available, multiple combinations

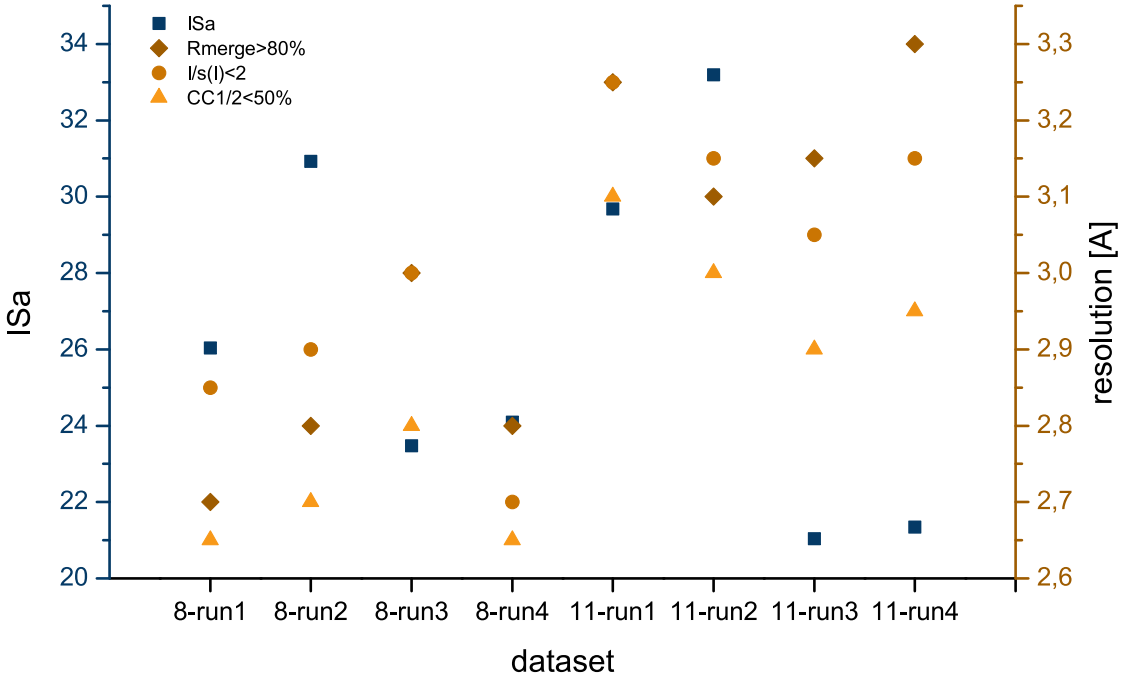


Figure 4.7.: The data quality of all individual data sets was evaluated after integration with XDS. The resolution limits for the commonly used indicators are listed as well as the ISa limit.

were scaled and merged individually resulting in 13 reflection files of different size and quality. For scaling and merging the programs XSCALE, PHENIX_scale_and_merge and XPREP were used in addition to an experimental version of SHELXC. The resulting reflections files could comprise from two to all eight scans and were evaluated based on different criteria.

Since the structure contains seven selenium atoms in SeMet residues per molecule, the anomalous signal was used for SAD phasing. Additionally, the strength of the anomalous data could serve as quality indicator for the merged files and give some idea whether the structure could be solved using the resulting merged data set. Evaluating all merged files originating from different merging and scaling programs with the same quality indicators in the same program increases the comparability. The anomalous signal was therefore examined with SHELXD and ANODE.

Evaluation of the merged data sets

SHELXD is a commonly used program for structure solution. The program is capable of locating the anomalous scatterers from SAD, MAD, SIR or SIRAS data to yield a substructure solution based on direct methods. A likely solution is indicated by the best

CFOM value achieved in all tries and should be above 35. CFOM is the combination of CC_{weak} and CC_{all}, which should be above 15 and 25, respectively. The best CFOM in a fixed number of tries can serve as a quality indicator when comparing the merged data sets (see Figure 4.8).

Structure solution should be possible with all data set combinations presented in Figure 4.8 but the overall quality of the substructure solution can differ considerably. Using two scans (scan 1+2) from cln5k-Se-8 or cln5k-Se-11 results in the lowest CFOM values and the value increases if more scans (scan 1+2+3 or all scans) are used instead. It is clearly visible that the combination of multiple scans from one crystal increases the CFOM independently from the merging program used.

The best overall CFOM was achieved with seven out of the eight data sets using the experimental version of SHELXC. Leaving out cln5k-Se-8 scan3 improved the overall CFOM when using the programs XSCALE and SHELXC. XPREP and PHENIX produced their best result from all eight data sets, but the CFOM was not the best achieved overall when compared with other programs.

ANODE was used to calculate the anomalous density at the positions of the specified heavy atom. With the known substructure and approximate phases the program calculates the averaged density for each heavy atom position. A log file gives the anomalous density for each heavy atom position as well as the strongest anomalous peaks which can be used as quality indicator. The strength of the anomalous signal specific for selenium atoms can be evaluated (see Figure 4.9).

In general it is assumed that the averaged anomalous signal for the anomalous scatterer has to be above 10 to allow a substructure solution. Using only the first two scans of cln5k-Se-11 would fail to produce enough anomalous signal strength to solve the structure. When adding more scans to the merged file, the averaged anomalous density increases for data sets from both cln5k-Se-8 and cln5k-Se-11. This effect was observed in SHELXD as well.

Interestingly, the best overall result and individual results were obtained using XSCALE. The best overall averaged anomalous density was obtained from using all eight data sets – closely followed by leaving out only cln5k-Se-8 scan3. The increase in signal compared from all scans of cln5k-Se-8 to including scans from cln5k-Se-11 is surprisingly small. The same effect is also visible in the SHELXD statistics.

With both SHELXD and ANODE the use of all eight data sets or seven of all excluding cln5k-Se-8 scan3 resulted in the best anomalous signal. While using the scans from cln5k-Se-11 resulted only in a small increase in the anomalous signal strength in total, they were used nonetheless to improve the overall data accuracy and quality.

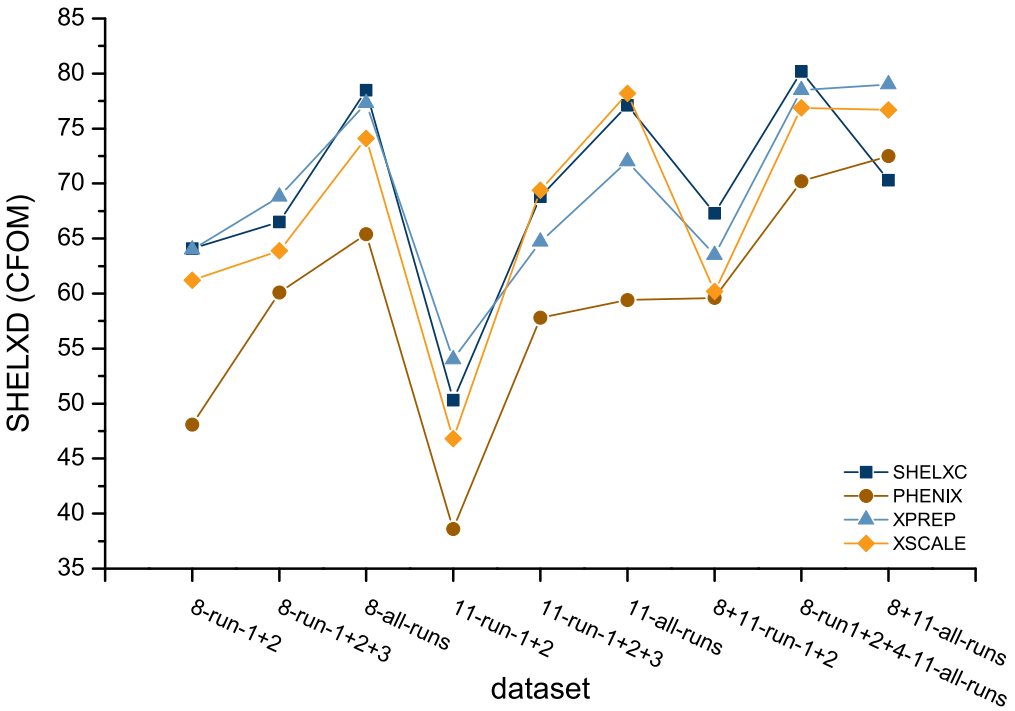


Figure 4.8.: SHELXD results for selected combinations of cln5-k-Se data sets. The results obtained from the files merged by the programs PHENIX, SHELXC, XPREP, and XSCALE are presented. Lines serve as visual guides only.

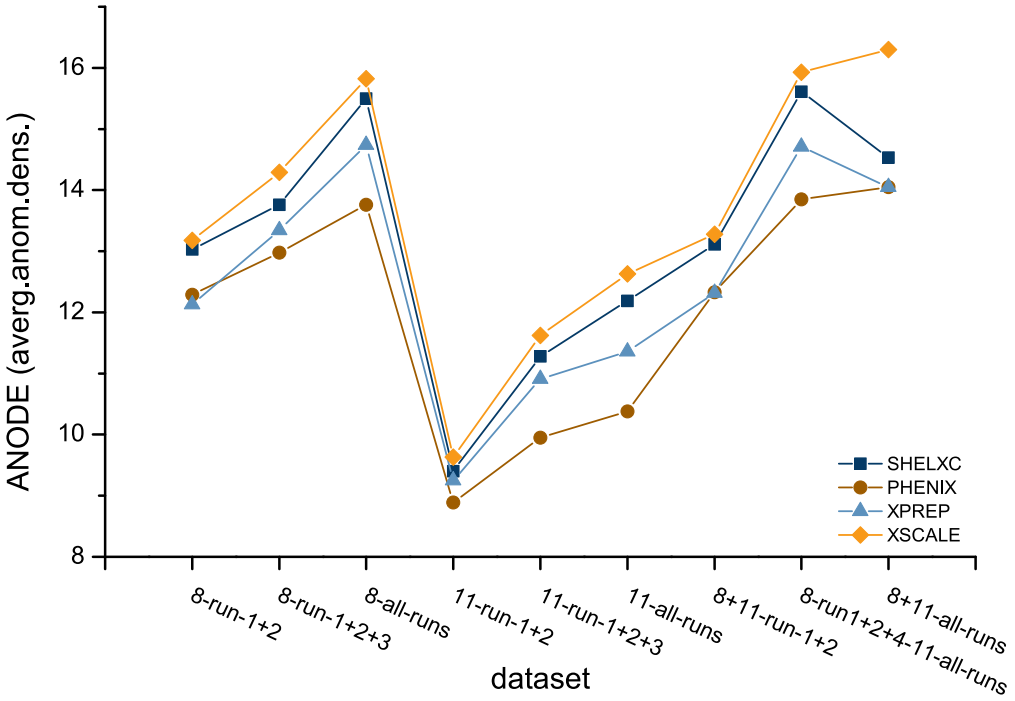


Figure 4.9.: ANODE results for selected combinations of cln5-k-Se data sets. The averaged anomalous density for the selenium atoms obtained from the files merged by the programs PHENIX, SHELXC, XPREP, and XSCALE are presented. Lines serve as visual guides only.

Considering all data quality indicators, the final merged data set used for structure solution and refinement was created using XSCALE using all scans from cln5k-Se-8 and cln5k-Se-11, except cln5k-Se-8 scan3.

4.4.7. Structure solution and refinement

All data sets were evaluated for their quality and from the best data sets an averaged file with the best anomalous signal was created (see Chapter 4.4.6). The merged data set could be used to find a solution using SHELXC/D/E and the initial Cα-trace was extended using ARP/wARP (Langer *et al.*, 2008). The solution was improved by iterative manual model revision in COOT (Emsley *et al.*, 2010) and refinement with REFMAC5 (Vagin *et al.*, 2004) using the CCP4i2 GUI (Potterton *et al.*, 2018). The manual model revision was performed by inspecting the agreement of the model with electron density maps: a map with model bias correction ($2mF_0$-DF_c) and a difference electron density map (mF_0-DF_c).

Additionally, the SHELXC experimental version automatically recognized and resolved the indexing ambiguity, creating a merged file from all scans of one crystal. This file was sufficient to solve the selenium substructure in SHELXD and a good alpha-traced solution in SHELXE was found, using the initial solution as a reference file.

Glycosylation was present, as visible in the difference electron density map (see Figure 4.17). Therefore the Carbohydrate module in COOT was used to model high-mannose sugar modification (Agirre *et al.*, 2017). The sugar moieties were only modeled into the visible difference electron density and when justified by the fit and B-factors. Privateer (Agirre *et al.*, 2015) was used to generate restraints for NAG and β-D-mannose (BMA) and conformation validation.

A model could be built for the amino acid residues 99 to 399 of the cln5 sequence and high-mannose sugar modifications at four positions (Asn179, Asn192, Asn252, and Asn304). Two gaps in the chain from amino acids 151–160 and 347–352 remained in the final model. The overall B-factor was above 50. All residues refining to a B-factor over 100 were removed from the model.

Model validation was performed using the PDBredo (Joosten *et al.*, 2014), MolProbity (Chen *et al.*, 2010) and FitMunk (Porebski *et al.*, 2016) services. The last refinement steps were conducted using PHENIX_refine (Adams *et al.*, 2010), which yielded a structure model with a final R/R_{free} of 22.1/24.7, respectively. The final refinement results were deposited with the PDB under assertion code 6r99.

4.4.8. Model quality

An overview of the model and refinement statistics are listed in Table 4.7.

Table 4.7.: Refinement statistics for cln5k-Se-merged data set.

Refinement statistics	Cln5K-Se-merged
resolution range[a]	38.56–2.70 (2.79–2.70)
reflections used in refinement[a]	10352 (1013)
reflections used for R-free[a]	522 (45)
R-factor[a]	0.2207 (0.2983)
R_{free}[a]	0.2472 (0.3741)
CC(work)	49.2
CC(free)	52.4
r.m.s. deviations from ideal geometry:	
bonds [Å]	0.004
bond angles [°]	0.74
No of non-hydrogen atoms	2441
No of residues	284
No of sugar moieties	7
No of water molecules	35
Ramachandran plot [%]:	
residues in preferred regions	94.62
residues in allowed regions	4.66
residues in disallowed regions	0.72
rotamer outliers [%]	0
clashscore (Molprobity)	4.30
average B-factor [Å^2]:	
overall	57.05
macromolecule	56.93
ligands	61.21
solvent	53.54

[a] highest resolution shell in parenthesis.

The initial refinement was performed using REFMAC5 via the CCP4i2 GUI. During model validation the R/R-free gap increased to 5 and no improvement in the amino acid geometry was achievable. The manual adjustment of the weighting scheme did not result

in the recommended root mean square (r.m.s.) deviation of bonds around 0.02 Å but remained significantly higher. Therefore PHENIX_refine was employed for the final steps of refinement during model validation. The R/R_{free} gap decreased significantly owing to lower R_{free} values. Both the MolPobity score and the clashscore decreased from 2.24 to 1.58 and from 7.08 to 4.3, respectively, indicating a better overall model quality. Furthermore, the r.m.s. bond deviation decreased and the protein geometry improved resulting in more favored rotamers and more residues in Ramachandran favored regions.

The overall data resulted in a good model considering the resolution limit of 2.7 Å. The limited resolution is reflected in multiple close contacts remaining and the lack of reasonably modeled solvent molecules. Only a few water molecules with close distance to other residues are kept in the final model and some regions of un-modeled electron density remain.

The final model contains two regions where side chains of the amino acids were not modeled, because no electron density was visible to support any sensible conformation. This manifested in unusually high B-factors which were physically unreasonable. A depiction of the structure colored by the B-factor of the main chain atoms is presented in Chapter C.7. It is conceivable that these regions of the protein are more flexible and exist in multiple conformations. In a structure based on XRD data to a higher resolution these might be visible and could be modeled.

Overall, the final model quality statistics are in the expected ranges for a structure solved at a resolution of 2.7 Å. Compared to other models produced from a similar resolution – the cln5 structure model is of good quality.

SHELXL refinement

While the resolution of the acquired data does not justify a SHELXL (Sheldrick, 2015) least-squares refinement, the program can help refining the occupation of the selenium atoms in the structure. Since the modification with SeMet was achieved by starvation and changing of the available medium, it can not be expected that all methionine residues are completely exchanged. Therefore PDB2INS (see Chapter 3) was used to convert the *pdb* file to the input format *ins* required by SHELXL and the structure factor file was produced from the initial merged file with XPREP. Since the R_{free} flags were not present in the original merged file, a new set was generated in XPREP. While restraints for all common amino acids and the most frequent ligands are automatically added to the instruction file by PDB2INS, the restraints for the residue SeMet, MSE, were initially

missing. Restraints for MSE were generated by GRADE server[8] and added manually.

First, the refinement of the Se occupancy was attempted by assigning each individual Se atom to a separate part. Since the same site could be occupied by either selenium or sulfur, one atom of both types was placed at each site. The sum of occupancies of both atoms of the site was restrained to be one, but the distribution of occupancy between the two element types was refined freely with one free variable for each residue. The occupancy of all positions was not refined at once. The positions were added step-wise with ten refinement cycles (conjugate gradient least squares) in between, starting at the position with the lowest expected occupancy reported by ANODE. This strategy lead to no meaningful result, evident e.g in residue 202 which initially refined to an occupancy for selenium of around 73% but rose to (over) 100% when all selenium positions were refined at the same time (see Table C.5).

Next, SHELXL was instructed only to refine the occupation of all selenium atoms with one free variable shared by all residues, leading to an overall occupancy of 82%. Standard deviations were calculated with one cycle of least squares refinement while allowing no shifts in the atom's positional parameters. Occupation and displacement parameters were refined separate from the coordinates.

Phenix_refine and REFMAC5 are capable of refining the occupancy of selenium sites by simply reducing the occupancy of the site. The resulting occupancies are listed in Table 4.8.

Table 4.8.: Occupancies of selenium atoms as refined by PHENIX_refine and REFMAC5.

SeMet residue	135	165	182	202	240	244	383	mean
PHENIX_refine	0.73	0.77	0.75	0.68	0.88	0.78	0.82	0.77
REFMAC5	0.95	0.96	0.99	0.97	0.97	0.98	1.00	0.97

The resulting mean occupancy from PHENIX_refine with 77% is close to the result obtained from SHELXL. In contrast, REFMAC5 refined the occupancies to an average of 97%, contrary to general expectations. In conclusion, the Met residues were mostly replaced with SeMet, in all probability to about 80%. This is the ratio of exchange reported before (Barton *et al.*, 2006).

4.4.9. Molecular replacement

Structure solution via molecular replacement was attempted using the data sets cln5k-79 (single scans and merged data sets) and the model refined against the final merged

[8] *http://grade.globalphasing.org*, Global Phasing Ltd.

data set cln5k-Se-merged. The full model was used as well as a model containing only the Cα main chain. The programs MolREP (Vagin and Teplyakov, 2010), Buccaneer (Cowtan, 2006), SIMBAD (Simpkin *et al.*, 2018), and PHASER_MR as implemented in PHENIX (McCoy, 2007) and CCP4i2 were tested. Furthermore, Arcimboldo_Shredder (Sammito *et al.*, 2014) was employed.

No program yielded any acceptable solutions. The XRD data of the cln5k crystals all show anisotropy and difficulties during space group assignment. A closer look at the quality of these data sets might result in a structure solution in the future.

4.4.10. Structure similarity studies

Protein structure comparison can expose distant evolutionary relationships not detected by sequence identity. The three-dimensional fold is believed to have a major impact on the stability, protein behavior, and the ability of a protein to bind ligands or other proteins. Therefore, similarity analysis of a protein structure is a valuable tool in understanding the proteins function or role.

Most common structure comparison programs and web services use different techniques such as geometrical positions resemblance or secondary structure elements to find local alignment. The qualification of structure similarity is often based on functions very specific for the given server, as reviewed in Hasegawa and Holm (2009). With the final model of the structure of cln5 different structure similarity search servers were employed to find a homologous structure. The web services Dali (Holm *et al.*, 2008, Holm and Laakso, 2016) and PDBeFold (Krissinel and Henrick, 2004) were used.

PDBeFold parameters

Analyzing structure similarity with PDBeFold, the following parameters were used to quantify the result. PDBeFOLD reports the quality of a result by the parameters Q-score, P-value, P-score, and Z-score (Krissinel and Henrick, 2004). The following definitions were taken from the official PDBeFold website[9].

The Q-score represents the quality function of the C_α alignment, maximized by the secondary structure matching (SSM) alignment algorithm. The Q-score reaches 1 for perfect alignment and drops with increasing r.m.s. deviation or decreasing alignment length.

The negative logarithm of the P-value is the P-score. The P-value measures the probability of achieving the same or better *quality of match* at chance. A P-score lower

[9] *http://www.ebi.ac.uk/msd-srv/ssm/*

than 3 indicates a statistical insignificant match. Z-score is statistical significant in terms of Gaussian statistics, the higher the Z-score the higher the statistical significance.

Furthermore some general information on the matched secondary structure are presented. N_{res} and N_{align} state the number of residues and length of alignment as number of matched residues, respectively. $\%_{seq}$ represents a quality characteristic of Cα-alignment as sequence identity in percent. N_{SSE} and $\%_{SSE}$ describe the number of residues and percent of matched SSEs, respectively.

Dali parameters

The web server Dali evaluates structure similarity by a Z-score. Results yielding Z-scores lower than 2 are considered spurious and should not be considered any further. Matches with a Z-score above 2 show significant similarities. Furthermore, matches with an Z-score above an empirical cutoff, depending in the size of the protein, or a sequence identity above 20% are regarded as strong. The empirical cutoff is given by $N/10 - 4$, where N is the number of residues in the query structure.

4.5. Results and discussion

4.5.1. Interaction Studies

Influence of cln5 deficiency on other NCL proteins

Studies of cln5 deficiency on other NCL proteins and lysosomal proteins with a role in autophagy were conducted. The influence on the expression of other NCLs and autophagy proteins in knockdown *CLN5* HeLa cells was analyzed (see Figure 4.10 and Table 4.9). The results for cln11, cln12, cln13 and cln14 are not shown since the antibody did not display adequate specificity. No visible or measurable staining could be achieved with the cln2 antibody, therefore the results were excluded as well.

Table 4.9.: Analysis of influence of *CLN5* knock-down on NCL and autophagy proteins by western blot.

	cln5	cln1	cln3	cln4	cln6	cln10	p62	LCI/II
control	1	1	1	1	1	1	1	1
	0.07	0.99	0.68	1.25	0.85	0.94	2.99	1.77
CLN5$^{-/-}$	-	0.34	0.45	0.35	0.01	0.27	1.07	1.30
	-	0.45	0.87	0.45	1.84	0.76	0.98	0.52

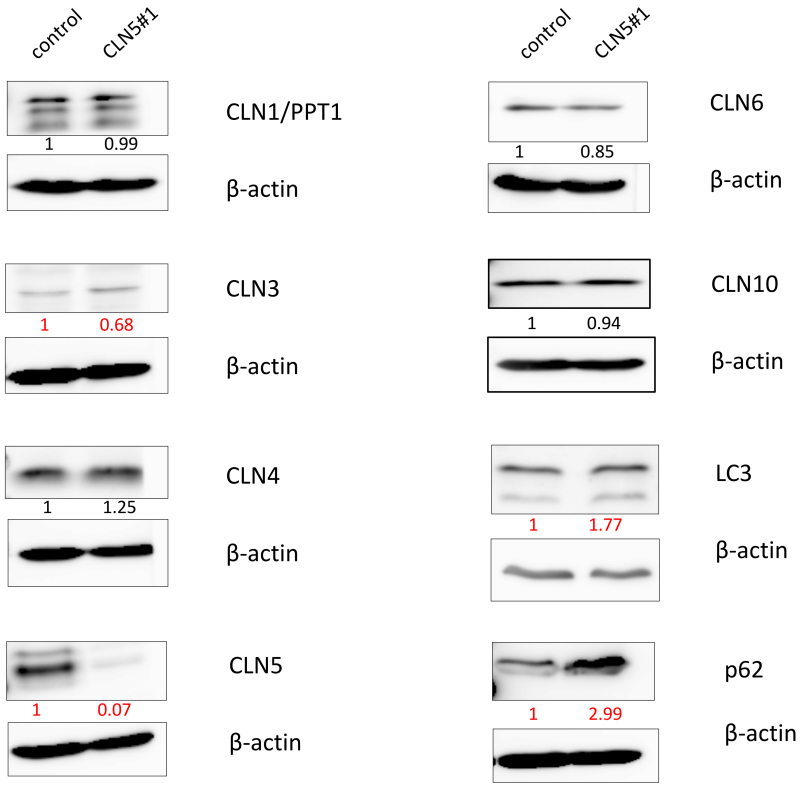


Figure 4.10.: Interaction studies by western blot on un-transfected cells expressing cln5 (shRNA, control) and *CLN5* knockdown Hela cells (*CLN5*#1 shRNA).

The deficiency of cln5 displayed an influence on other NCL proteins and autophagy associated proteins p62 and LCI/II. Overall the results did not display a clear trend, since large variations are present in the results. For example, the first measurement showed an up-regulation of the proteins p62 and LCI/II. This result could not be confirmed by the subsequent measurements, where p62 showed no change in comparison to the control and LCI/II even showed a decrease.

The NCL proteins cln1, cln3, cln6, and cln10 displayed no change in expression level compared to the control.

Next, an increase of cln10 with cln5 depletion as expected from the results by Markmann *et al.* (2015) could not be verified. Also an influence of cln5 deficiency on cln3 could be reported here. The amount of cln3 was depleted slightly with cln5 deficiency. Vesa *et al.* (2002) reported that cln5 and cln3 show interaction via co-immunoprecipitation assay. Furthermore, the proposed interaction between cln5 and cln6, as suggested by Lyly *et al.* (2009) could not be conclusively confirmed. Further studies will be necessary

to determine the influence of cln5 deficiency on NCLs proteins or proteins linked to autophagy.

Investigation of cln5 processing by CathepsinD (cln10)

The interaction between cln5 and active CTSD was researched by SDS-PAGE studies. To determine whether cln5 is processed by CTSD, cln5 was incubated with different ratios of active CTSD for various time periods. To simulate lysosomal acidity, the studies were performed in 100 mM sodium acetate puffer pH 4.5. Firstly, CTSD activity was verified at lysosomal pH (see Table C.1).

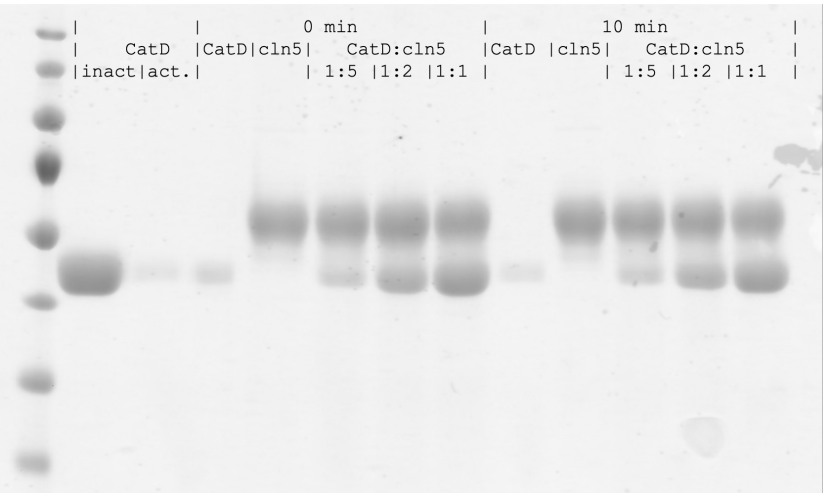
Next, the reaction solution containing active CTSD and cln5 was analyzed. Samples of the reaction solution were removed at given time intervals and SDS-PAGE used to separate the components (see Figure 4.11). The cln5 band remains stable and without any visible decay. SDS-gels were analyzed with ImageJ to obtain a first impression of cln5 concentration in presence of active CTSD (Table 4.10). These preliminary studies can be quantified by further investigation using other experimental techniques.

Table 4.10.: SDS-PAGE-based analysis of the CTSD processing of cln5. The ratio of CTSD to cln5 was varied and the amount of cln5 remaining after a certain time period was analyzed. The results are displayed as cln5 amount in contrast to the control.

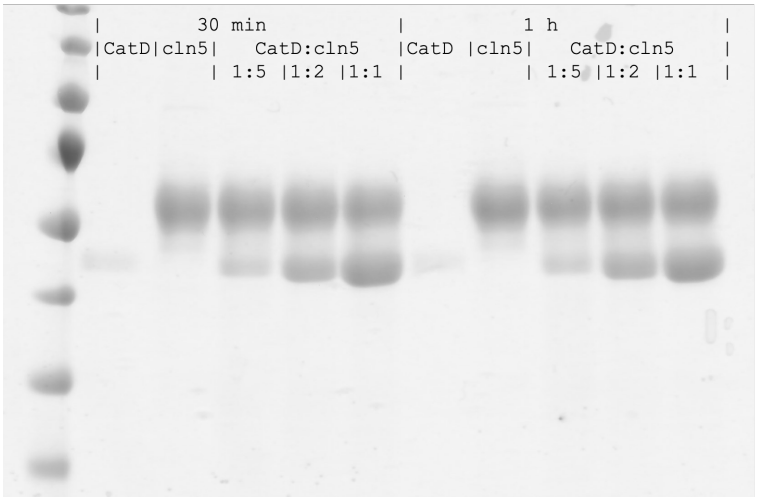
time	control	ratio CTSD:cln5			
		1:10	1:5	1:2	1:1
0 min	1	27.30			8.61
	1		0.98	0.96	0.93
	1		0.90	1.03	0.87
10 min	1	0.70			0.23
	1		0.98	0.93	0.73
	1		0.88	0.85	0.73
30 min	1	1.29			0.57
	1		0.85	0.91	0.98
	1		0.87	0.91	0.79
1 h	1	0.92			0.48
	1		1.04	1.01	0.99
	1		0.82	2.56	0.84

No decrease of cln5 can be found in the presence of active CTSD at lysosomal pH. A

processing of cln5 by CTSD within the time scope of the experiment was not measurable. It is possible that further studies with a longer exposure of cln5 to active CTSD can find evidence of cln5 processing by CTSD. Also a different experimental setup might yield a more precise analysis of an interaction between cln5 and CTSD.



(a) Before interaction (left) and after an incubation time of 10 minutes (right).



(b) Incubation time 30 minutes (left) and 60 minutes (right).

Figure 4.11.: SDS-PAGE of cln5 interaction with active CTSD to evaluate possible cln5 processing.

4.5.2. Circular dichroism

Circular dichroism (CD) was used to investigate the secondary structure of cln5 in solution. First, the difference between the two peaks eluted by Ni-affinity chromatography

is evaluated (see Section C.4). The resulting CD spectra show no variation and a further discrimination of the eluted protein is not considered necessary. It is assumed that this is a purely self-made distinction due to the step-wise gradient used during Ni-affinity chromatography, since neither SDS-PAGE nor CD analysis show a significant variation.

Next, different concentrations of the protein are measured to determine the likely secondary structure elements (see Figure 4.12).

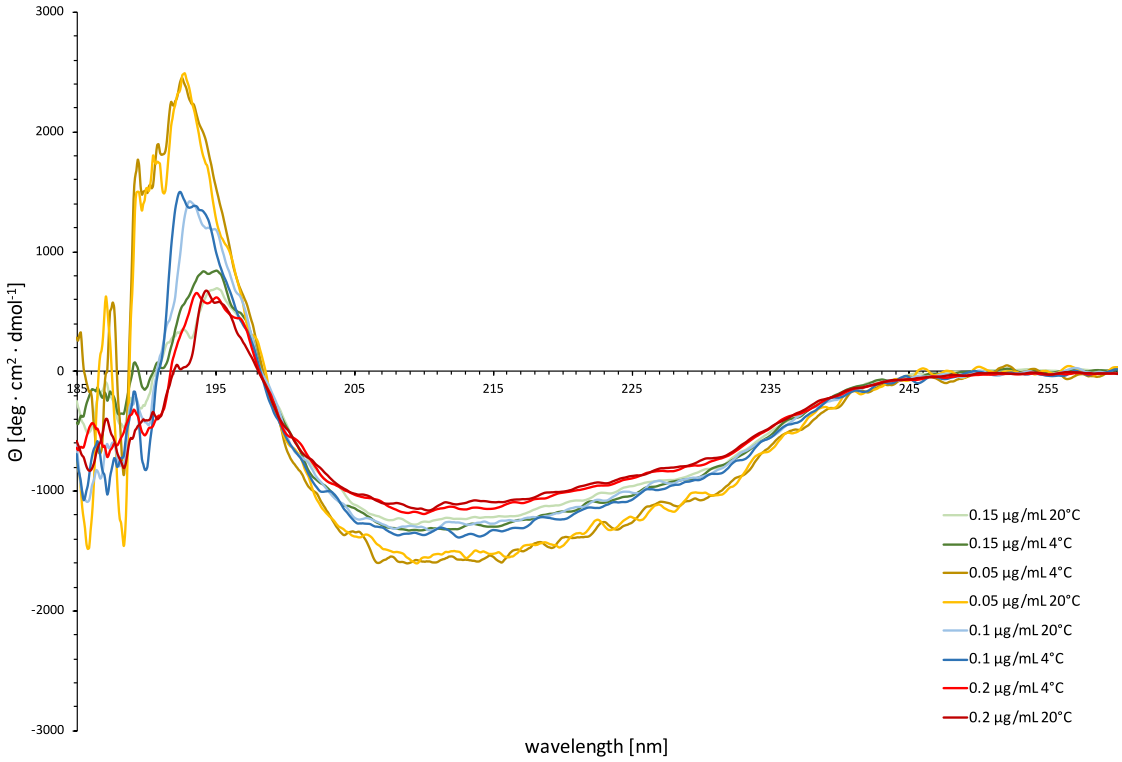


Figure 4.12.: Circular dichroism spectra calculated from circular dichroism of cln5 protein at different concentrations. Concentrations in µg/mL.

The CD analysis shows no dependence of the secondary structure on the concentration or the temperature at which the data was collected or the concentration. The temperatures of 4 °C and 20 °C are typical temperatures for the crystallization of proteins. No change in the secondary structure in the temperature range is favorable for crystallization.

The spectra clearly indicate that antiparallel β-sheets could be present in the structure of cln5 when compared to spectra as discussed in Greenfield (2006). But the intensity of the peaks is not as high as expected for a tertiary structure containing only this structure element. It can be ruled out that the protein is disordered or denatured.

In conclusion, the CD analysis confirms that a secondary and tertiary structure is

present in solution. Additionally, there is no measurable structural difference between the two peak fractions obtained from purification.

4.5.3. Structure description

One molecule of cln5 is observed in the asymmetric unit. The crystal structure reveals cln5 to be a globular protein with two domains consisting of α-helices and anti-parallel β-sheets (see Figure 4.13 and, for other viewpoints, Figure 4.14). A schematic overview of the structure is displayed in Figure C.7, Section C.7. In Figure 4.14 the structure is colored by regions depicting grouped secondary structure elements that can be considered as folds.

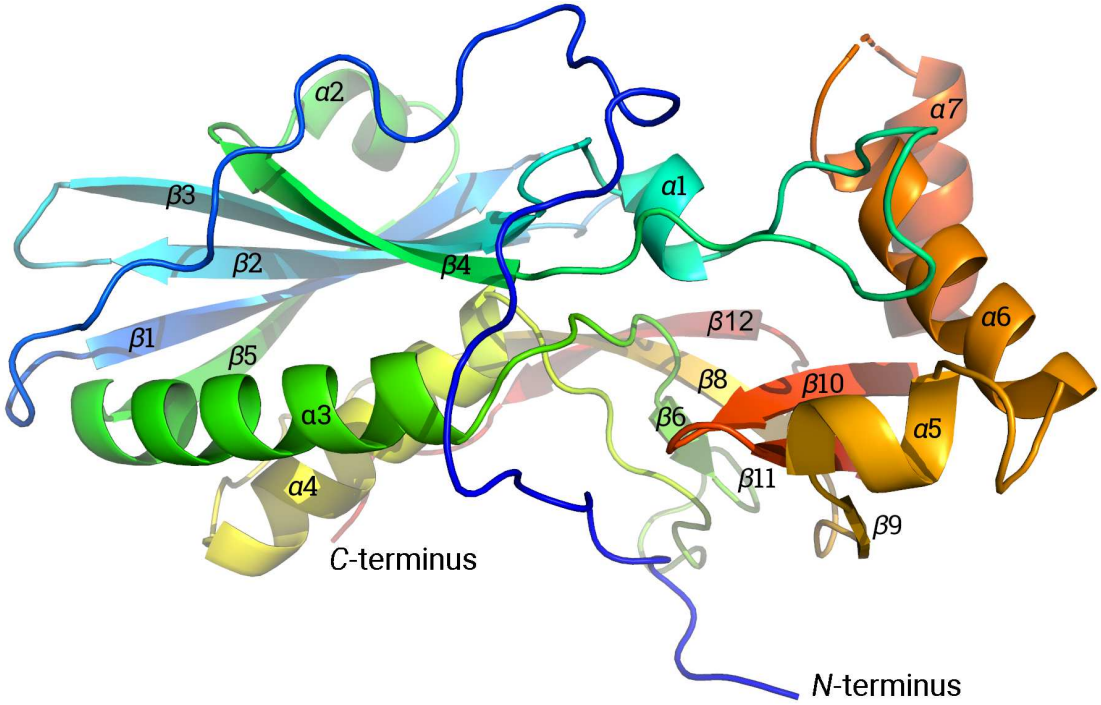


Figure 4.13.: Structure of cln5, colored by sequence, from blue at *N*-terminus to red at *C*-terminus. Secondary structure elements are labeled consecutively.

Starting from the *N*-terminus, the first 40 amino acids form an extended strand covering nearly one complete flank of the molecule. Two disulfide bridges, from Cys119 to Cys208 and from Cys126 to Cys214 keep the first region in place between two sugar modifications at Asn192 and Asn252 (see Figure 4.19). The strand proceeds into an antiparallel β-sheet consisting of five β-strands (β1–5). The sheet is intermitted by a long β-hairpin reaching over to the other side of the structure with a short α-helical

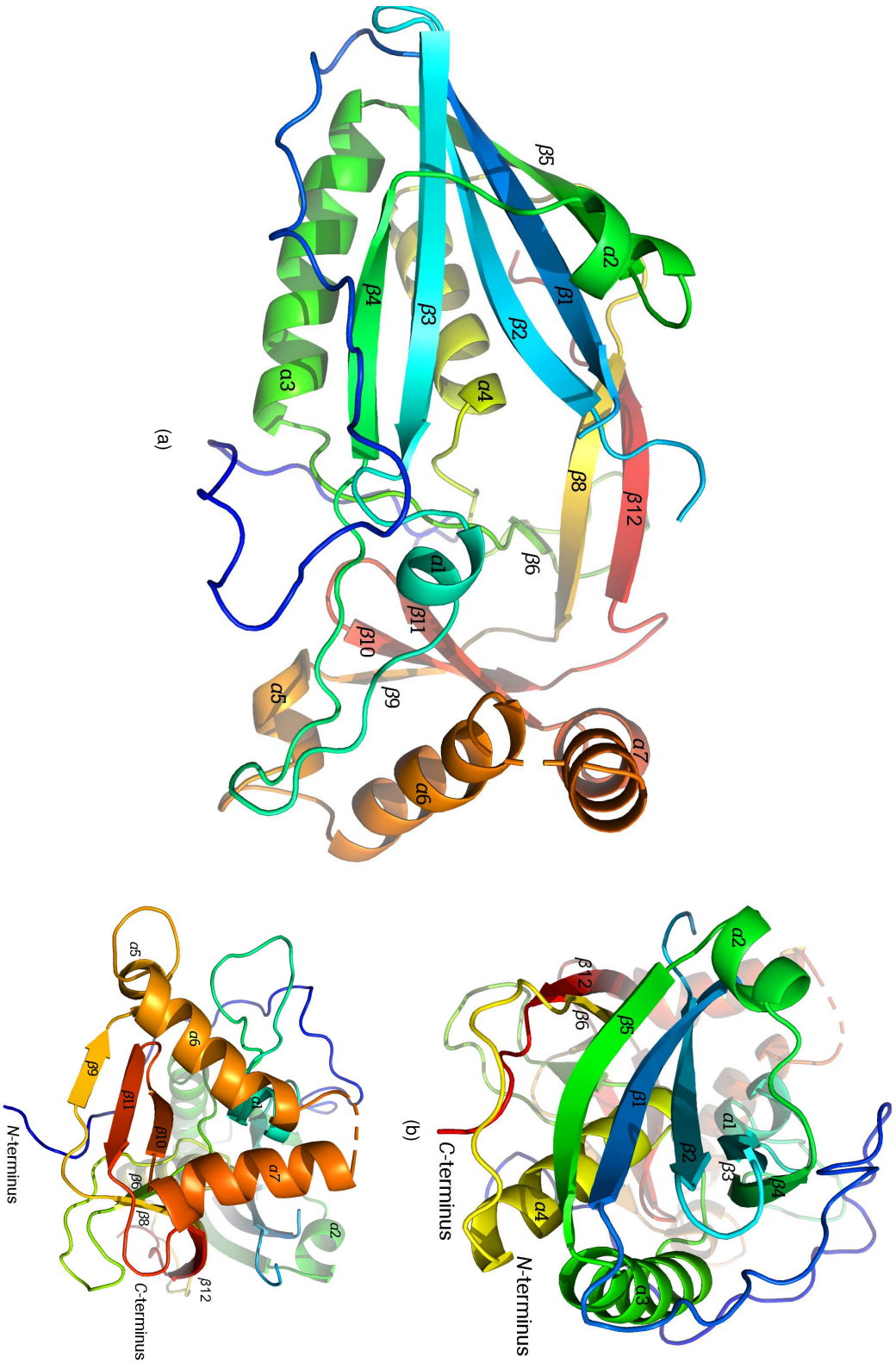


Figure 4.14.: Structure of cln5 in different orientations in reference to the front view in Figure 4.13. (a) structure as seen from above. (b) structure as seen from the left site. (c) structure as seen from the right side.

element (α1) between the third and fourth strand, and by a short α-helix (α2) between the fourth and fifth strand. While the helix α2 is lying across the sheet on one side, the region is completed by two prominent α-helices (α3 and α4) on the opposite side forming an $\alpha\alpha$ motif (see 4.15, blue fold).

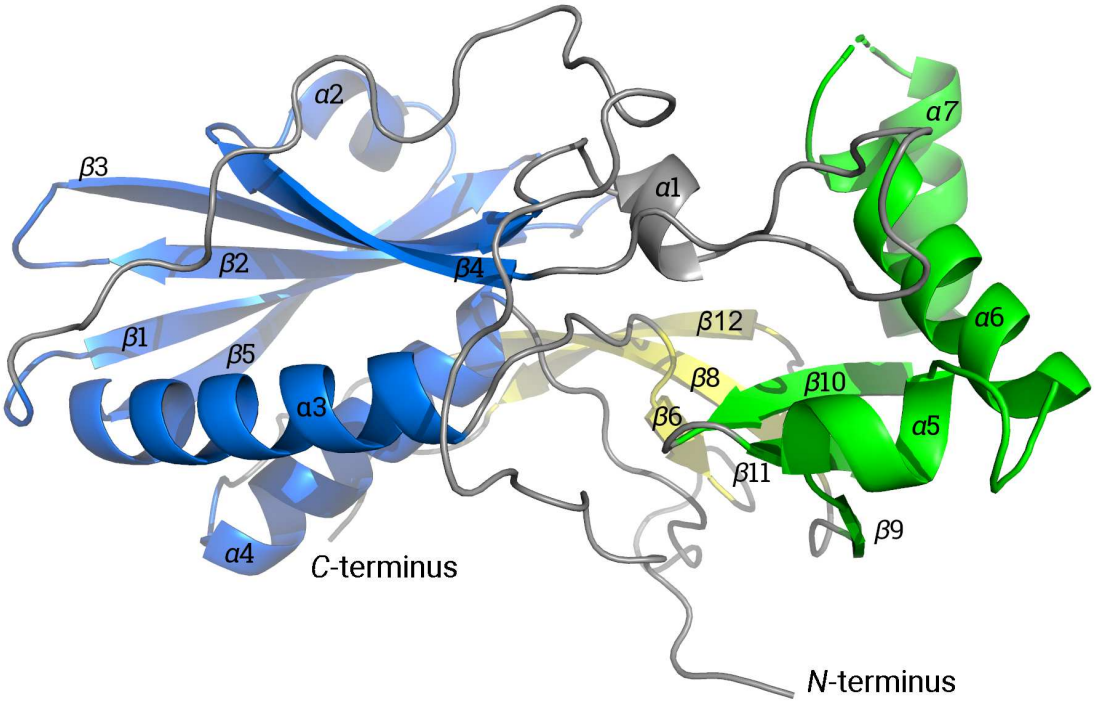


Figure 4.15.: Structure of cln5, colored by fold. The first fold, colored in blue, consists of the antiparrallel β-sheet (strands β1–5) and two α helices (α1–2). The second fold, colored in yellow, is a short antiparrallel β-sheet (strands β6, 8, and 12.). The last structure fold, colored in green, consists of a small antiparrallel β-sheet (strands β9, 10, and 11.) and three α helices (α5–7).

In the middle, a short antiparallel β-sheet consisting of three strands (β6, β8 and β12) is located (see 4.15, yellow fold). The first strand of the β-sheet is sequencially located between α3 and α4.

Between the strands β8 and β12 an adjoining antiparallel β-sheet (β9 – 11) is located (see 4.15, green fold). Also, two more α-helices (α6 and α7), forming another $\alpha\alpha$ motif, and a short α-helical element (α5) are part of the region. The strand β12 is the last structural element before the *C*-terminus, which is located next to helix α4.

The crystal structure shows electron density for the amino acids 99 to 399 with two gaps of ten and six amino acids between Val150–His161 and Pro346–Phe353, respectively (see Figure 4.16). Furthermore, significant electron density is present at four glycosylation sites, Asn179, Asn192, Asn252, and Asn304 (see Section 4.5.3).

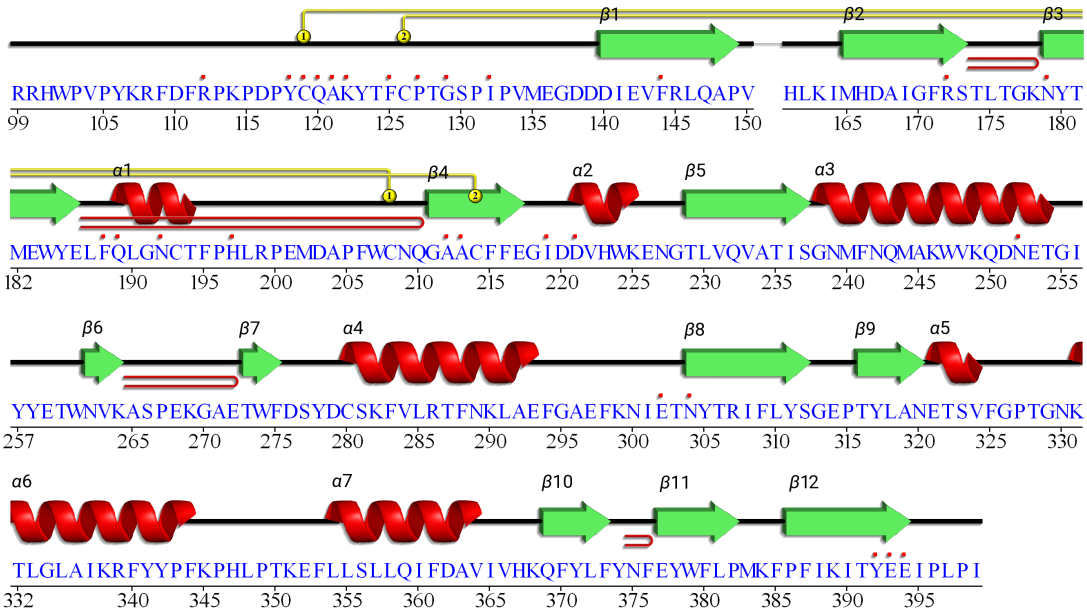


Figure 4.16.: Sequence of cln5 annotated with secondary structure elements. Secondary structure elements are visualized as helix in red and strands in green, numbered consecutively as α and β, respectively. β hairpins are illustrated with red lines. Furthermore, disulfide bridges are marked and numbered in yellow and residues with contacts to ligands are tagged with a red dot.

Comparison with protein structure predictions from literature

The structure of cln5 is a globular, soluble protein. Larkin *et al.* (2013) predicted a putative transmembrane helix reaching from amino acid 76 to 91 and an amphipathic helix located at amino acids 353–392. While the first transmembrane helix is in front of the reported signal peptide cleavage site at Ile96, the mature protein retains the amphipathic anchor region. Indeed, an α-helix is present from amino acids 353 to 364 (α7) in the region of the predicted second transmembrane helix (see Figure 4.16). This secondary structure element is only eleven residues long instead of the anticipated 39 residues and directly followed by two β-strands (β10 and β11), that form an antiparallel β-sheet. The structure prediction by Savukoski *et al.* (1998) appears more accurate, since here a hydrophobic region was predicted from amino acids 353 to 373. Both topology studies are based on bioinformatic analysis of the cln5 sequence but were conducted 15 years apart (Larkin *et al.*, 2013, Savukoski *et al.*, 1998). It is not clear why the deployed prediction tools reported different results.

Since only the residues from amino acids 99 to 399 are resolved in the crystal structure of cln5, the signal peptide cleavage position at Val93 or Ile96 is more probable than

at an earlier residue. But a conclusive statement based solely on the crystal structure presented here cannot be made here, of course.

Glycosylation sites

The *N*-glycosylation sites at Asn179, Asn192, Asn252 and Asn304 show excellent electron density for carbohydrate modification. Here high-mannose modification was modeled with up to three residues (see Figure 4.17). An overview of all sugar modifications and their immediate surroundings is presented in Figure 4.18 and in Figure C.9.

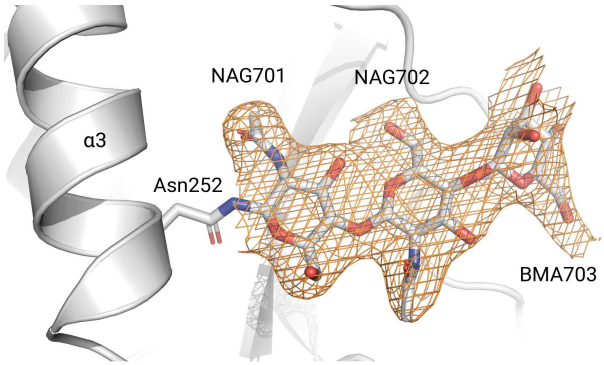
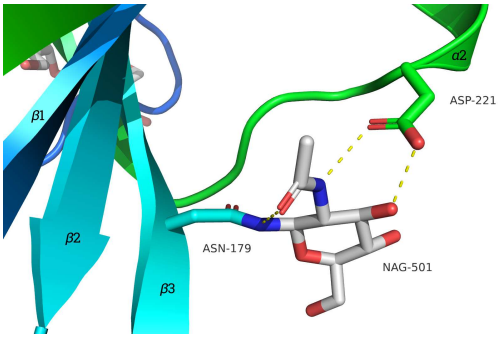


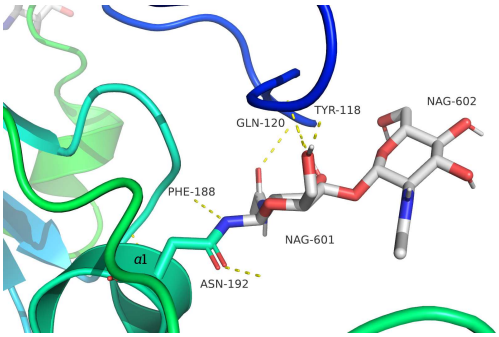
Figure 4.17.: The sugar modification at Asn252 could be modeled with three sugar molecules. The electron density from the final refinement is contoured at 1σ around the residues NAG701–702 and BMA703.

Asn179 is located at the start of strand β3 and the first sugar residue, NAG501, can form hydrogen bonds with the nearby side-chain of Asp221 at helix α2 (Figure 4.18a). Two sugar residues can be found at Asn192, NAG601 and NAG602 (Figure 4.18b). Hydrogen bonds connect NAG601 and the protein backbone at Gln120 as well as Tyr118. These interactions can hold the extanded strand in the first part of the protein in place, in addition to the disulfide bridges (see Figure 4.19).

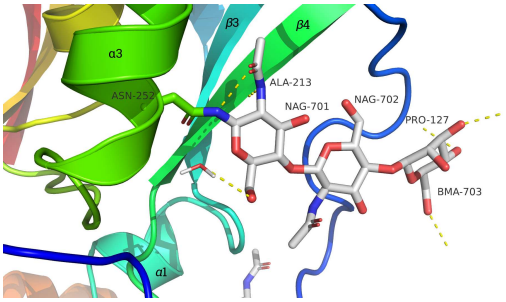
At Asn252 three sugar residues could be modeled, NAG701, NAG702, and BMA703 (Figure 4.18c). Hydrogen bonds to the protein backbone at Pro127 and Ala213 may contribute to the stability of the structure. The interaction of Pro127 with BMA703 is located to the other side of the *N*-terminal, extended strand in comparison to NAG601. This indicates a contribution to a stabilized overall motif at this position, visualized in Figure 4.19. The sugar residues around the peptide chain at amino acids 117 to 121 and the extended strand is further stabilized by disulfide bridges to the sides at Cys119 and Cys126.



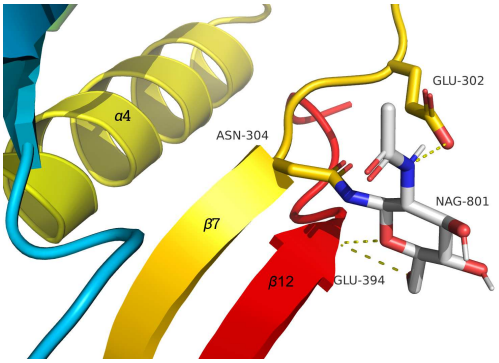
(a) Sugar modification at Asn179 (NAG 501).



(b) Sugar modification at Asn192 (NAG 601-602).



(c) Sugar modification at Asn252 (NAG 701-BMA703).



(d) Sugar modification at Asn304 (NAG 801).

Figure 4.18.: The sugar modification at the amino acids Asn179 (a), Asn192 (b), Asn252 (c), and Asn320 (d) are displayed with the nearest contact and the surrounding residues.

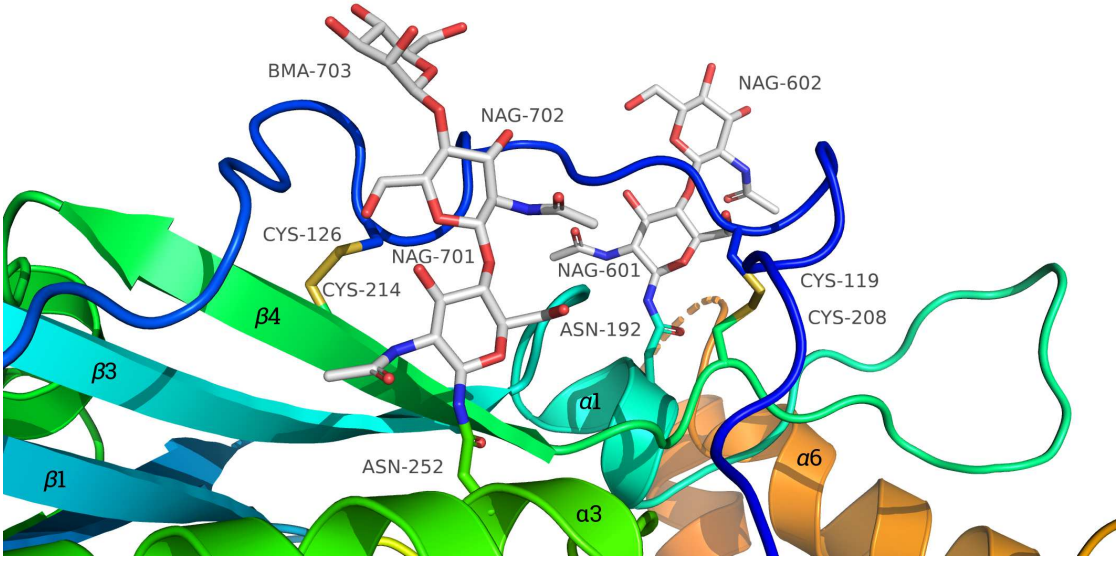


Figure 4.19.: Two disulfide bridges Cys119 to Cys208 and Cys126 to Cys214 are visualized and the nearby sugar modifications at Asn192 and Asn252 flanking the *N*-terminal extended strand are displayed.

One sugar molecule, NAG801, was modeled at Asn304 and displays hydrogen bonds to Glu302 and the amino acid backbone at strand β11 (Figure 4.18d). An overview of the nearest residues for all sugar modification is also given in Chapter C.7.1.

4.5.4. Structure homology

Structure prediction

At the very beginning of this study several secondary structure prediction web-based servers were employed. An idea of the secondary structure elements that can be expected might help modify crystallization conditions and, in the best case, a binding partner could be identified. The web based services of the structure prediction servers I-TASSER, BLAST, Swiss Model, HHPred and RaptorX were employed with the protein sequence of cln5. None of the programs resulted in a prediction that gave further insights into the structure of cln5, possible homologues or interaction partners (see Chapter C.8.1).

The final structure solution achieved in this thesis can be compared with past structure predictions. Huber and Mathavarajah (2018) published a binding prediction that anticipated a putative NGT binding side involving the residues Gln147, Pro148, Val101, Lys154, Gln243, Tyr311, Gly313, Glu314, Tyr317, Phe373[10]. This putative binding side was based on a predicted secondary structure from the RaptorX web server (see Figure 4.2). Figure 4.20 displays the XRD structure of cln5 with all predicted binding residues from the publication by Huber and Mathavarajah (2018) marked in green. All residues predicted to bind NGT are located widely apart in the structure of cln5. The difference between the predicted cln5 structure and the actual crystal structure reported in this thesis is quite prominent.

[10]The original publication numbered the residues starting from 1 for the first residue in the shortened sequence corresponding to Leu49 in the full sequence. Therefore the residue numbers were corrected to be consistent with the numbering used in this work.

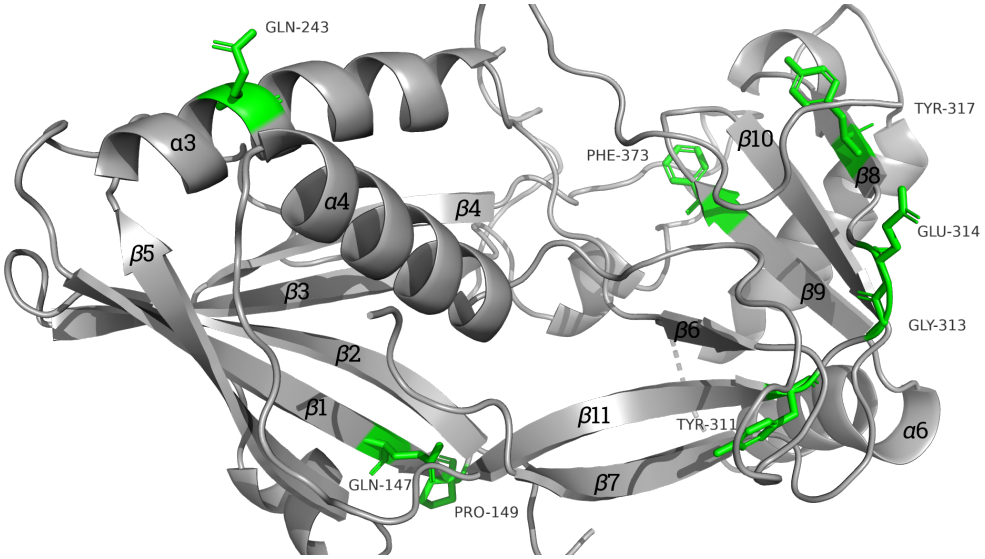


Figure 4.20.: Structure of cln5 with predicted 'binding residues' as discussed by Huber and Mathavarajah (2018) based on a secondary structure prediction from RaptorX server.

An attempt to reproduce the reported structure prediction was not successful (see Chapter C.8.1). Furthermore, no putative binding site could be predicted from the models obtained with the RaptorX web service. The secondary structure prediction results from I-TASSER, Swiss Model and HHPred are reported in section C.8.1. BLAST did not report any sequence matches outside of the cln5 orthologues.

In conclusion, the secondary structure prediction calculated by the RaptorX web server was not helpful in predicting the secondary structure in advance of the crystallization studies. Furthermore, no additional information was gained and results reported from other groups could not be reproduced. The additionally employed structure prediction methods resulted in poor or incomplete models of low confidence scores and were not regarded any further. Since most of the algorithms employed by this programs are based on a search for sequence similarity or homologies, none could predict a sensible structure of cln5. Even sophisticated programs like RaptorX cannot replace XRD structure solution for proteins with no known homologues.

4.5.5. Structure similarity

Initial structure similarity matches were found using the Dali web service. Dali sorts results by Dali Z-score, a measure for similarity. The structure similarity between cln5 and the two top matches found with Dali is weak but significant (see Table 4.11).

Table 4.11.: Structure similarity matches using Dali web service (Holm *et al.*, 2008, Holm and Laakso, 2016).

PDB-ID	chain	Z-score	RMSD[a] [Å]	lali[b]	N_{res}	$\%_{ID}$	protein description
3ebq	A	6.8	3.2	109	144	11	PPPDE1
2wp7	A	6.7	3.3	111	158	12	PPPDE2

[a] RMSD = root mean square deviation.

[b] lali = length of the alignment.

PPPDE1 (permuted papain fold peptidase of DsRNA viruses and eukaryotes 1) (Xie *et al.*, 2017) and PPPDE2 (Suh *et al.*, 2012), also known as DeSI2 (deSUMOylating isopeptidase 2) and DeSI1, respectively, show structural similarity with cln5 in the same region. While the Z-score is clearly above 2, indicating a significant similarity, the comparably low sequence identity of 11% and 12% does not support a strong match.

PDBeFold allows structure alignment via the identification of residues occupying equivalent geometrical positions. It is based on secondary structure matching (SSM) and can perform pairwise or multiple three-dimensional alignment (Krissinel and Henrick, 2004). The significance of a match is measured by Q-score, P-score, Z-score, r.m.s. deviation (RMSD) and percentage of sequence identity ($\%_{SSE}$) (see Chapter C.9).

Table 4.12.: Alignment statistics CLN5 and PPPDE1 from PDBeFold (Krissinel and Henrick, 2004).

cln5		Alignment				PPPDE1	
N_{res}	$\%_{res}$	Q	P	RMSD [Å]	N_{align}	N_{res}	$\%_{res}$
285	35	0.159	1.54	2.119	99	144	69
N_{SSE}	$\%_{SSE}$	$\%_{seq}$	Z	N_{SSE}	N_{gaps}	N_{SSE}	$\%_{SSE}$
18	39	12.1	6.63	7	9	12	58

When using the result from Dali as search query in PDBeFold, the result gave a poor Q-score and a sequence alignment of only 12% (Table 4.12). A sequence alignment of 99 residues is found, which corresponds to 69% of the PPPDE1 protein. Nonetheless, the match could describe a significant part of the cln5 protein structure – 35% of all modelled residues – and may help to assign a probable function. The match was therefore investigated further.

Next, the structure overlay of cln5 and PPPDE1 was visualized in a graphics program for visual inspection. The rotation-translation overlay was produced with PDBeFold, which also provides a table with the results of the 3D Cα alignment. In this alignment table the distance is rated as well as the amino acid similarity (see Figure 4.21). Several regions display a close distance match and better superposition. Some amino acids are identical and occupy the same position. These regions of interest were analyzed further.

Permuted papain fold peptidase of dsRNA viruses and eukaryotes 1

PPPDE1 was reported to have a deubiquitinating function (Iyer *et al.*, 2004, Xie *et al.*, 2017) and deSUMOylation activity (Shin *et al.*, 2012). It was suggested that PPPDE1 is part of the NlpC/P60 superfamily of papain-like enzymes due to its similar fold (see Figure 4.22) (Xu *et al.*, 2011). Embedded in this fold lies an arrangement of catalytic residues, including a Cys/His-containing dyad with an additional tyrosine.

The NlpC/P60 superfamily belongs to the cysteine peptidases, proteolytic enzymes that hydrolyze a peptide bond. The thiol group of a cysteine acts as a nucleophile and cysteine peptidases are often active at acidic pH. The superfamily is characterized by a papain-like fold consisting of a bundle of α-helices on one side and a β-barrel motif on the other side. The cysteine is here part of a catalytic triad consisting of a cysteine, a histidine, and an asparagine (aspartic acid). The NlpC/P60 superfamily consists of four main families, as suggested by sequence analysis (Xu *et al.*, 2011). P60-like, AcmB/LytN-like, Yaef/Yiif-like, and LRAT-like enzyme families are ubiquitous papain-like cysteine peptidases or other functionally related enzymes (Anantharaman and Aravind, 2003, Aramini *et al.*, 2008, Pai *et al.*, 2006, Xu *et al.*, 2010, 2009).

The families within the superfamily are sub-grouped. The canonical papain-like NlpC/P60 enzymes (CPNEs) contain the families of P60-like and AcmB/LytN-like enzymes. These are hydrolases with specificity for amide linkages in cell-wall components and have a catalytic core similar to that of papain (Xu *et al.*, 2011).

The permuted papain-like NlpC/P60 enzymes (PPNEs) consist of the families AcmB/-LytN-like and Yaef/Yiif-like enzymes. These were predicted to have a circularly permuted catalytic domain consisting of a conserved catalytic triad (Cys, His and a polar third residue). Bioinformatic studies have suggested that PPNEs are related to the PPPDE superfamily (Iyer *et al.*, 2004).

An overview of the structural folds associated with the NlpC/P60 super family is given in Figure 4.23, (b) and (c), respectively. The depicted protein structure folds are drawn similarly to the predicted topology diagrams by Iyer *et al.* (2004).

The structure of PPPDE1 was solved by the Structural Genomics Consortium (PDB-ID

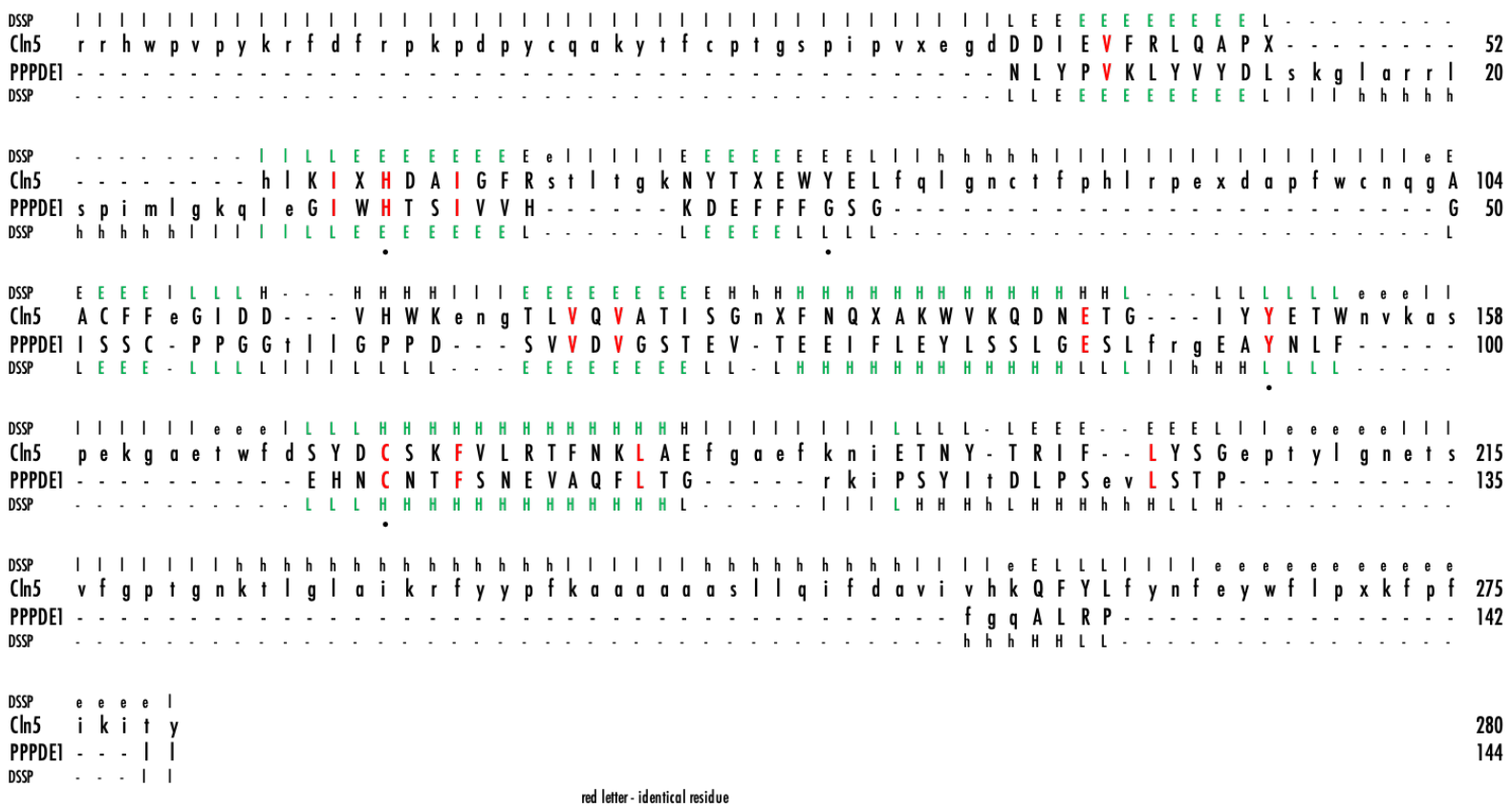


Figure 4.21.: Sequence alignment of cln5 and PPPDE1 of the NlpC/P60 super family with secondary structure elements above and below sequence. Capital letters if the residues overlay (both secondary structure and sequence), green letters when secondary structure matches and red letters show identical residues. H = helices, E = beta strand, L = loop/no structural feature. The point annotates the reactive residues.

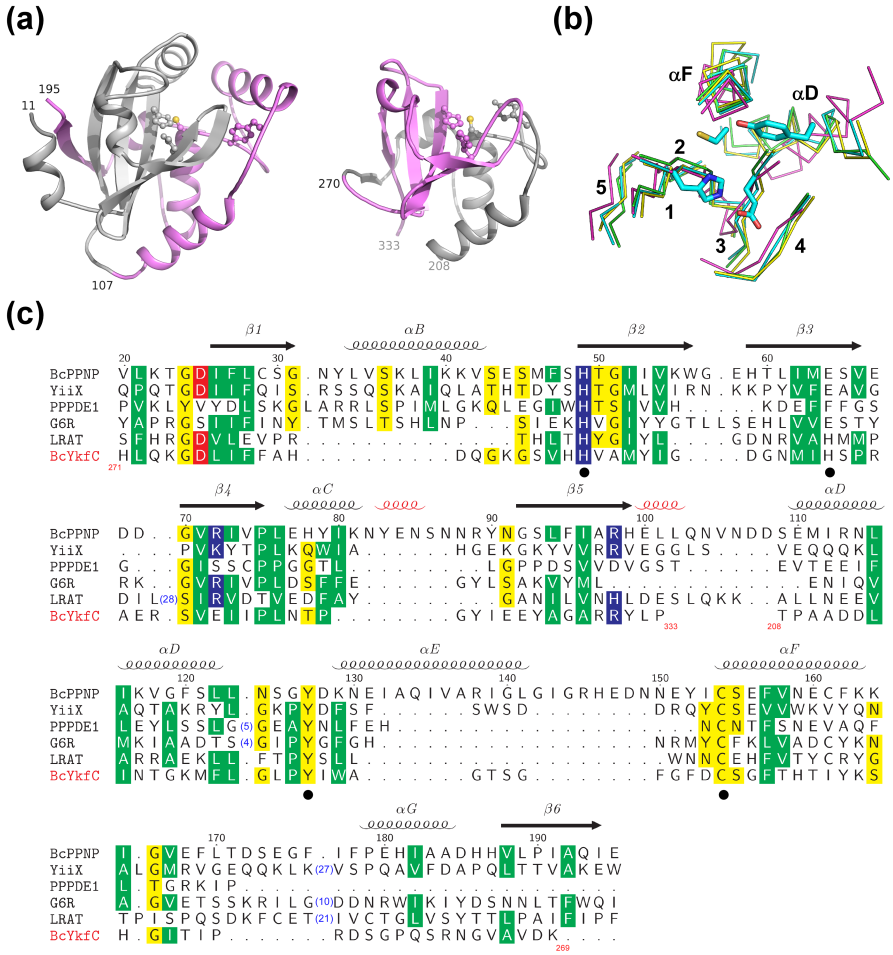
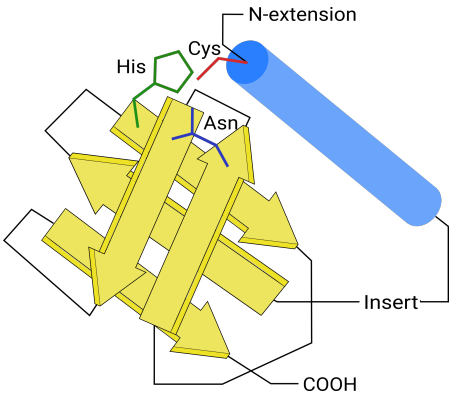
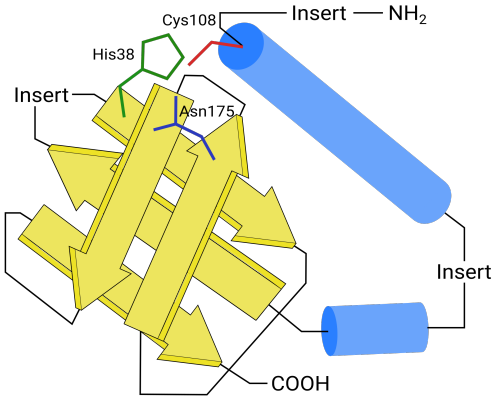


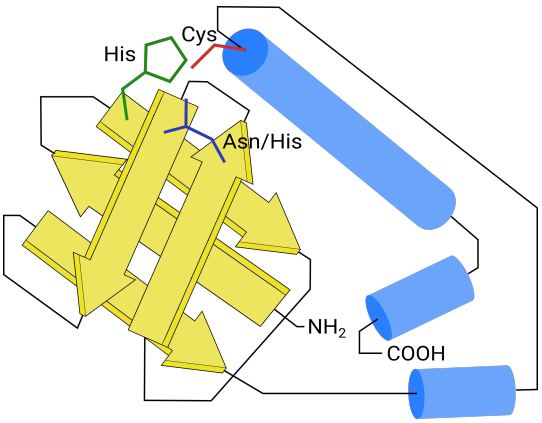
Figure 4.22.: Structural analysis of the papain-like proteins of the NlpC/P60 superfamily taken from Xu *et al.* (2011). a) depicts a structure comparisson between BcPPNE (left) and CPNE BcYkfC (right) with subdomains colored and catalytic residues as ball-and-sticks. b) The conservation of the core structure within the NlpC/P60 superfamily is visualized by superposition of conserved residues. c) Sequence alignment of PPNEs (BcPPNE, YiiX, PPPDE1, G6R and LRAT) and CPNE BcYkfC.



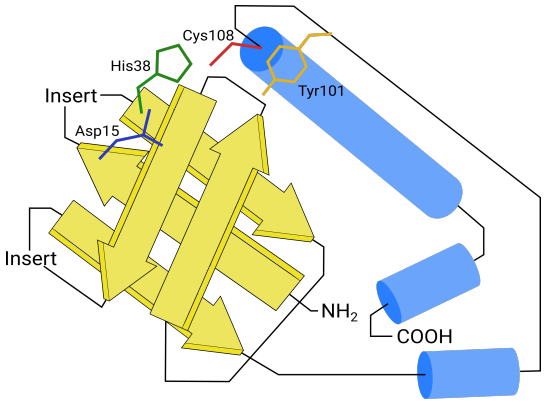
(a) Papain-like fold.



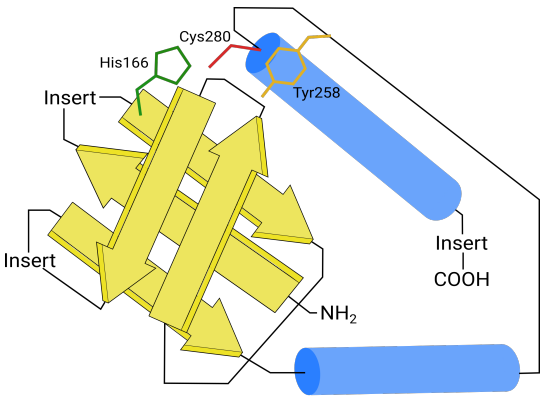
(b) Papain-like fold as present in the structure papain (1PPN).



(c) Predicted permuted papain-like fold in NlpC/P60.



(d) Permuted papain-like fold as present in the structure PPPDE1 (3EBQ).



(e) The topology diagram of the structure fold in cln5 is depicted. The fold is similar to the permuted papain-like fold in PPPDE1. The catalytic residues identified as catalytic triad in PPPDE1 can be found in cln5 in the same spatial arrangement.

Figure 4.23.: Topology diagrams of the NlpC/P60 super family protein folds (a–d) (Iyer *et al.*, 2004). The predicted permuted papain-like fold is derived from Anantharaman and Aravind (2003). In comparison the preserved structure fold found in cln5 is depicted in (e).

3ebq, unpublished results). The PPPDE family proteins have a conserved cysteine and histidine dyad for protease or acyltransferase activity in a circularly permuted papain-like fold. An overlay of the fold region and a sequence analysis was presented by Xu *et al.* (2011), see Figure 4.22. The dyad with an additional tyrosine forms a reactive triad consisting of the amino acids Cys108, His38, and Tyr101 in PPPDE1 (see Figure 4.23d).

Cln5 displays the circularly permuted papain-like fold

Taking a closer look at the reactive triad and the associated fold of the PPPDE proteins reveals that cln5 and PPPDE1 overlap and show conserved residues. Cln5 displays the triad of Cys, His and Tyr as present in PPPDE1. Furthermore, the complete structural fold of the circularly permuted papain-like fold around the reactive triad is present in the structure of cln5. The complete overlay was generated using PDBeFold, see Figure 4.24. In general, the structure fold involving α3 and α4 as well as β1–5 is arranged in a similar manner as in PPPDE1.

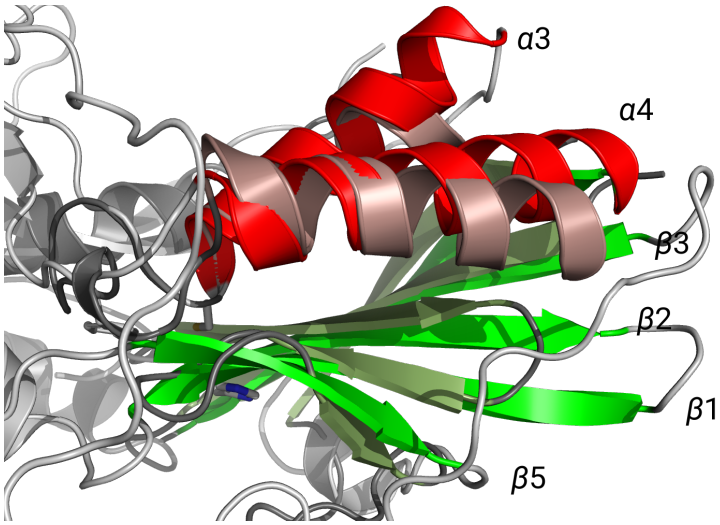
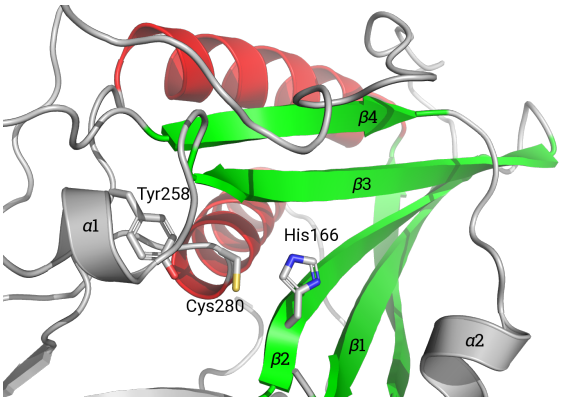
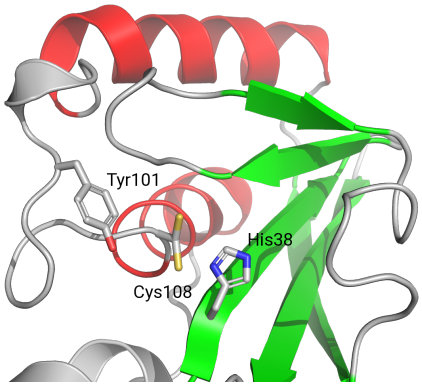


Figure 4.24.: Overlay of permuted papain-like fold with cln5 in dark gray and PPPDE1 in light grey, the similarity is colored as calculated by PDBeFold (red/green or light red/light green, respectively). Secondary structure numbering as in Figure 4.13, orientation ↻ 180°.

The reactive triad present in PPPDE1 is located to the side of the conserved structure fold, consisting of Cys108, His38 and Tyr101. As typical in the whole NlpC/P60 superfamily, the reactive center is located in between an α-helix bundle and β-sheet. These amino acids are visualized in Figure 4.25b. The cln5 structure displayed in Figure 4.25a in the same orientation shows a similar positioning for the residues Cys280, His166 and Tyr258. The triad is depicted in Figure 4.26.



(a) The triad in cln5 consisting of the amino acids Cys280, His166, and Tyr258.



(b) The triad in PPPDE1 consisting of amino acids Cys108, His38, and Tyr101.

Figure 4.25.: Structural overlay of the conserved secondary structure of cln5 and PPPDE1: Similarity marked in green (β-sheet) and red (α-helix). Orientation of structure in (a)/(b) as in Figure 4.13 and in 4.25a/4.25b, horizontal ↻ by 90°. The PPPDE1 structure displays residue Cys108 as disordered, both positions of the sulfur atom are displayed in (b).

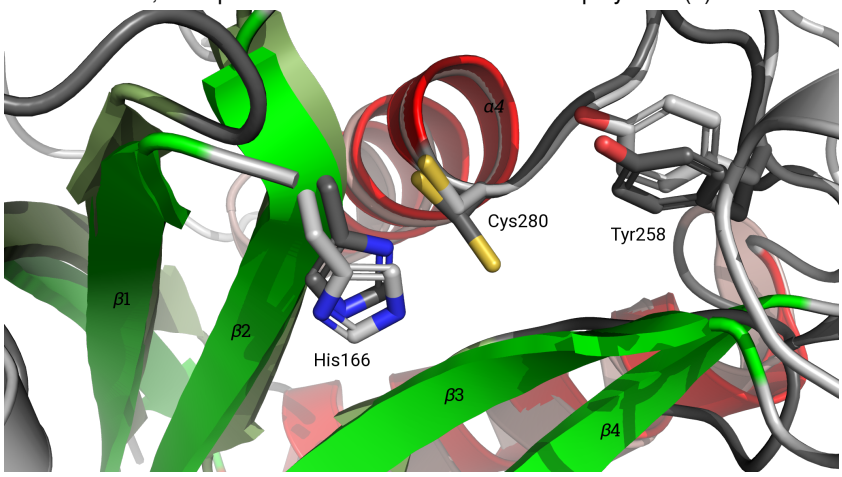


Figure 4.26.: The triad present in PPPDE1 is depicted as overlay with cln5. In PPPDE1 the triad consists of the residues Cys108, His38, and Tyr101. In cln5 the corresponding residues are Cys280, His166, and Tyr258. The residue His38 in PPPDE1 was reported disordered in two conformations, both are depicted here. The overlay of permuted papain-like fold is colored as in 4.25, the secondary structure numbering is given as in Figure 4.13, and the orientation is depicted as in the topology diagrams in Figure 4.23.

In conclusion, cln5 shows a strong structure similarity to PPPDE1 in the putative catalytic domain. While the function of cln5 remains unknown, the similarity to PPPDE1 with a conserved reactive triad gives a strong indication that cln5 might have an enzymatic function. The triad in PPPDE1 is the reactive center for the protein's deSUMOylation activity (Gillies and Hochstrasser, 2012, Shin *et al.*, 2012). Cln5 displays a circularly permuted papain-like fold in one of its domains, the triad is present in a conserved form at the edge of this fold. This is a strong indicator that the residues Cys280, His166, Tyr258 may have a catalytic function in cln5 as well.

Recently, PPPDE1 was reported to be a deubiquitinating peptidase (DUB) (Xie *et al.*, 2017) in addition to the deSUMOylating function associated with the cysteine peptidase (Iyer *et al.*, 2004). Neither function has been attributed to cln5 before. Nonetheless, as demonstrated here, the structure of cln5 could contain valuable clues to function. Further enzyme activity studies with a focus on a possible cysteine peptidase function should be conducted in addition to testing for a deSUMOylation activity.

4.5.6. *CLN5* mutation analysis

Since several dozen mutations of *CLN5* are known to date, the focus will be on point mutations leading to disease onset as depicted in Figure 4.27. A loss of function can be expected if no mature protein is expressed, as it is the case in missense mutations. With point mutations, it may be interesting to study why such small changes also lead to a loss of function.

Mutation cln5.003 transforms Asp279 to an asparagine residue, a change from acidic to polar amino acid. This residue is situated in the midst of the preserved catalytic triad (see Figure 4.28a) but is not part of it. Even small changes in a catalytic center could lead to total loss of function.

Another mutation affects the triad directly, cln5.011 describes a change of Tyr258 to asparagine. This tyrosine residue is part of the conserved triad as it is found in the PPPDE family. Since the mutation leads to disease onset and the amino acid is associated with catalytic function in other proteins, it it may point to the role of this residue in the function of cln5.

The mutation cln5.029 results in a change of Asn192 to a serine amino acid and thereby in the removal of a glycosylation site. In the XRD structure glycosylation at this residue was well ordered as indicated by the observed electron density. More interestingly, the sugar modification frames a part of the *N*-terminus of the protein with an extended strand (see Figure 4.19). Nonetheless, that part is held in place by two disulfide bridges and flanked by two sugar residues at either side, suggesting a well

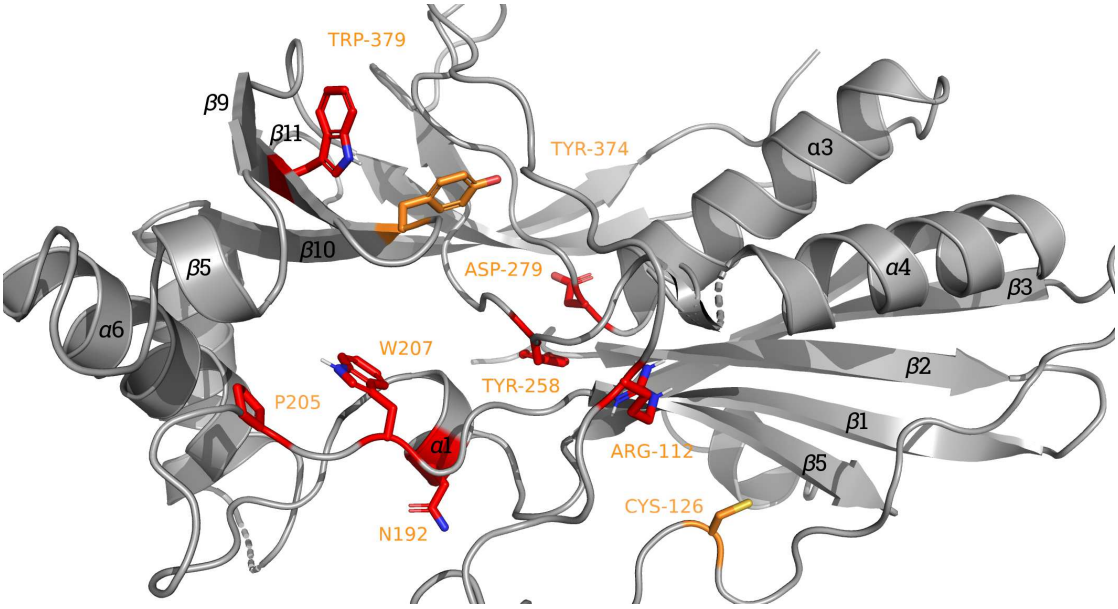


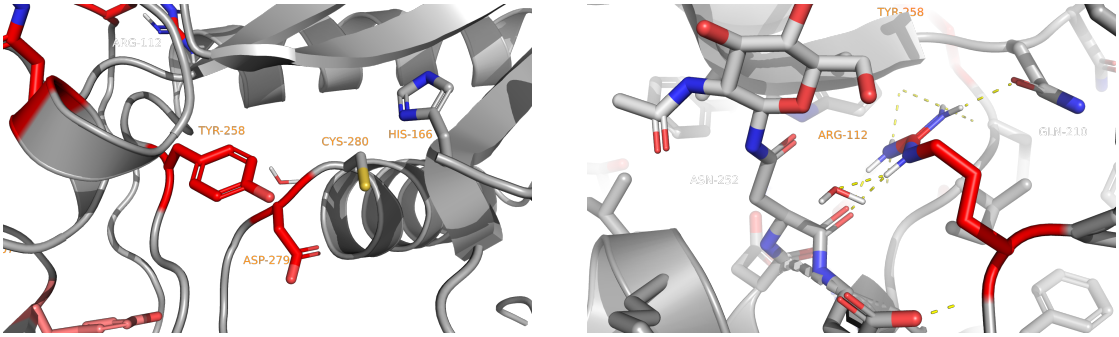
Figure 4.27.: Structure of cln5 is displayed and the currently known missense mutations are highlighted. All point mutations leading to symptoms of late infantile NCL are marked in red and orange for adult onset causing mutations.

ordered region (see Figure C.9b). The elimination of this glycosylation site might not only destabilize the pH-resilience but also the secondary structure itself.

The mutation cln5.021 transforms Cys126 into a tyrosine residue resulting in the loss of a disulfide bridge. The disulfide bridge binds Cys126 to Cys214, part of the β-sheet in a structure motif. This disulfide bridge is in the middle of the aforementioned part of the *N*-terminus framed by sugar modifications at Asn192 and Asn252.

Likewise located in this region is Arg112, which is changed to histidine by mutation cln5.006. A change from the basic arginine to histidine represents a very significant change in space requirement and distance for possible contacts. This residue is located at the edge of both the conserved triad region as well as the sugar moiety at Asn252 (see Figure 4.28b). Mutations in this part of the protein might not only disturb the folding and stability of the secondary structure but also result in loss of function, should the conserved core, as in the NlpC/P60 superfamily of papain-like enzymes, be a catalytic center.

Interestingly, the premature stop codon of Fin major mutation results in only a slightly shorted protein with a termination at Tyr392. The termination occurs shortly before the glycosylation site at Asn401. This part of the protein is not part of the conserved structure fold, therefore it is not possible to evaluate if the mutation affects the structure.



(a) Mutation site cln5.003.

(b) Mutation site cln5.006

Figure 4.28.: Mutation of Asp279 (4.28a) and Arg112 (4.28b) leads to disease onset in patients. Asp279 is directly neighboring the preserved reactive triad consisting of Cys108, His166 and Tyr258.

Localization studies showed that the mutation of Asn401 (N401Q), the glycosylation site itself, most likely resulted in a correctly folded protein that was mis-localized to the Golgi. This *N*-glycosylation site might be important for cln5 trafficking to the lysosome (Moharir *et al.*, 2013).

Further studies are mandatory to ascertain the function and interactions of cln5 and to understand the grave effect of mutations in patients.

4.6. Conclusion and outlook

In this work the protein cln5 was successfully crystallized in two different variants. XRD experiments were conducted on multiple crystals and structure solution was possible from crystals with the intrinsic anomalous scatterer selenium. The crystal structure of cln5 was measured to an average resolution for protein crystals. The structure gave vital insights into structural relationships not discovered until now.

The cln5 crystal structure shows three folds, the largest one of which can be found in a similar configuration in the protein PPPDE1. This relationship could not be discovered by bioinformatic analysis. PPPDE1 was reported to have a deSUMOylating function and a deubiquitinating function. The main structural fold of PPPDE1 is conserved within the family of permuted papain-like NlpC/P60 enzymes. The motif of the permuted papain-like fold is present in cln5 and reveals the structural relationship of cln5 to the PPPDE family.

The conserved reactive center of PPPDE1, consisting of a cysteine, a histidine, and a tyrosine residue, is present in cln5. This suggest that a putative catalytic site is present in cln5 as well. The reactive center is typical for cysteine peptidases with a papain-like fold. Most cysteine peptidases are active at acidic pH, as it can be found in the animal

lysosome were cln5 is located.

So far only one of the folds in the structure of cln5 could be associated with known folds. In the future it might be possible to find a structural relationship for the other part of the protein as well.

With a lack of mechanism-based therapies, the study of the protein structure can give insights into the function of cln5 and offer prospects for further therapies. A more fundamental understanding of the molecular mechanism and interactions is desirable.

As only part of the secondary structure of cln5 could be identified as a known fold, it is reasonable that cln5 may display other functions as well. An additional function remains elusive and further studies have to be conducted. Due to the structural similarity to the permuted papain-like NlpC/P60 enzymes, activity studies with their substrates should be considered. Studies to reveal a possible cysteine peptidase activity of cln5 might be sensible. It is essential to further establish biochemically the proposed function of cln5 based on the structure.

Appendices

A. Appendix Poly(rA)

A.1. Single crystal data

All X-ray diffraction (XRD) measurements of Poly(rA) crystals are listed in the tables in this chapter. Overall 37 scans were collected from seven crystals at three different beamlines. The data is presented in the order of measurement number (#) and not by crystal name. At first, the three crystals measured at the DESY synchrotron PetraIII PX11 undulator beamline are listed. Next, all crystals and their scans measured at the SLS synchrotron beamlines are given. Here crystals were measured at the undulator beamlines PXII X10SA and PXI X06SA. The measurements are listed in this order. Furthermore all data in this chapter is presented ordered by measurement number, from 1 to 37.

In Table A.1 addition information of the data collection are presented. An overview of different data quality indicators for each scan is given in Table A.2. The calculation of the data quality indicators was performed by PHENIX_xtriage. During the course of the investigation, data quality indicators were also collected from the programs XDS, XSCALE, and XPREP.

Table A.1.: Overview of the data collection from Poly(rA) crystals.

#	crystal	scan	beamline	rotation [°]	slicing [°]	exposure [sec]	transmission [%]	No of frames
1	A1	1	P11	90	0.2	0.1	1	450
2		2	P11	180	0.2	0.1	1	900
3		3	P11	180	0.2	0.1	1	900
4		4	P11	180	0.2	0.1	90	900
5		5	P11	180	0.2	0.1	1	900
6	A3	1	P11	180	0.2	0.1	2	900
7		2	P11	180	0.2	0.1	20	900
8		3	P11	180	0.2	0.1	50	900
9	A4	1	P11	180	0.2	0.1	5	900
10		2	P11	180	0.2	0.1	5	900
11		3	P11	180	0.2	0.1	5	900
12		4	P11	180	0.2	0.1	5	900
13		5	P11	180	0.2	0.1	50	900
14		6	P11	180	0.2	0.1	20	900
15	B3	1	X10SA	360	0.1	0.1	20	3 600
16		2	X10SA	180	0.1	0.1	10	1 800
17		3	X10SA	180	0.1	0.1	70	1 800
18	B4	1	X10SA	180	0.1	0.1	25	1 800
19		2	X10SA	180	0.1	0.1	50	1 800
20		3	X10SA	180	0.1	0.1	80	1 800
21	C3	1	X10SA	180	0.1	0.1	15	1 800
22		2	X10SA	360	0.1	0.1	15	3 600
23		3	X10SA	180	0.1	0.1	40	1 800
24		4	X10SA	180	0.1	0.1	50	1 800
25		5	X10SA	180	0.1	0.1	75	1 800
26		6	X10SA	180	0.1	0.1	90	1 800
27	C2	1	X10SA	180	0.1	0.1	10	1 800
28		2	X10SA	360	0.1	0.1	50	3 600
29		3	X10SA	180	0.1	0.1	30	1 800
30		4	X06SA	180	0.1	0.1	20	1 800
31		5	X06SA	180	0.1	0.1	50	1 800
32		6	X06SA	180	0.1	0.1	90	1 800
33	C1	1	X06SA	180	0.1	0.1	20	1 800
34		2	X06SA	180	0.1	0.04	50	1 800
35		3	X06SA	180	0.1	0.1	50	1 800
36		4	X06SA	180	0.1	0.05	90	1 800
37		5	X06SA	180	0.1	0.1	90	1 800

Table A.2.: Data Quality for each scan from the Poly(rA) measurements. Data quality indicators calculated by PHENIX_xtriage.

#	completeness [%]	redundancy	mean I/σ(I)	R_{meas}	R_{pim}	Wilson B [Å^2]
1	99.69	11.3	4.2	0.227	0.067	5.81
2	82.90	9.6	11.8	0.088	0.027	6.56
3	90.64	9.9	6.1	0.132	0.040	7.00
4	92.17	10.1	7.0	0.126	0.038	7.41
5	83.32	9.5	8.4	0.114	0.035	7.57
6	76.74	10.4	20.0	0.057	0.017	5.89
7	94.11	10.6	10.9	0.085	0.025	6.06
8	85.84	10.3	13.0	0.079	0.024	6.02
9	99.69	11.3	4.2	0.227	0.067	5.81
10	82.90	9.6	11.8	0.088	0.027	6.56
11	90.64	9.9	6.1	0.132	0.040	7.00
12	92.17	10.1	7.0	0.126	0.038	7.41
13	81.88	7.7	9.5	0.076	0.025	8.59
14	83.38	10.0	6.7	0.092	0.029	11.78
15	97.12	21.8	21.8	0.059	0.012	6.09
16	96.49	10.9	16.3	0.090	0.027	5.73
17	97.89	11.4	15.0	0.085	0.025	6.05
18	98.88	10.9	12.1	0.073	0.022	5.86
19	96.65	10.6	12.1	0.099	0.029	6.77
20	98.41	11.3	16.2	0.062	0.018	5.35
21	99.41	12.3	12.9	0.12	0.034	5.71
22	94.31	22.4	36.5	0.051	0.011	6.40
23	99.14	12.4	19.1	0.072	0.02	5.63
24	97.74	12.4	14.1	0.086	0.024	5.97
25	99.51	12.3	27.2	0.049	0.014	6.07
26	99.24	12.3	22.7	0.056	0.016	6.74
27	95.76	11.1	24.1	0.051	0.015	6.13
28	99.26	23.9	33.1	0.045	0.009	5.91
29	99.6	12.2	29.4	0.02	0.012	6.05
30	94.75	9.2	33.5	0.031	0.01	10.07
31	99.87	11.8	21.7	0.046	0.014	9.07
32	99.88	11.8	29.9	0.037	0.011	8.43
33	94.66	9.2	24.2	0.048	0.015	7.30
34	99.92	11.5	21.2	0.056	0.016	6.25
35	99.90	11.4	23.900	0.046	0.014	6.73
36	99.95	11.7	23.1	0.051	0.015	6.55
37	99.84	11.4	25.7	0.040	0.012	4.57

A.1.1. Radiation damage

The possible change in the size of the unit cell due to radiation damage is evaluated for cell axis a in Figure A.1.

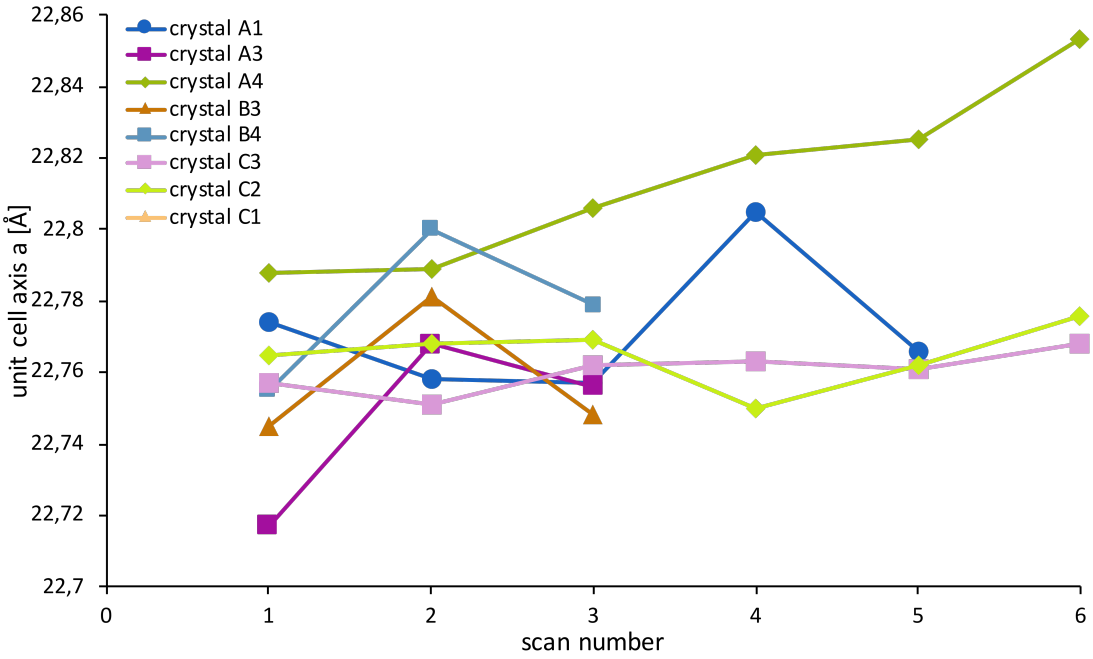


Figure A.1.: Change of unit cell axis a in consecutive measurements.

A.1.2. Overload correction

The influence of the overload correction on the parameter R_{anom} calculated by XPREP is plotted against the data set number (#) in Figure A.2.

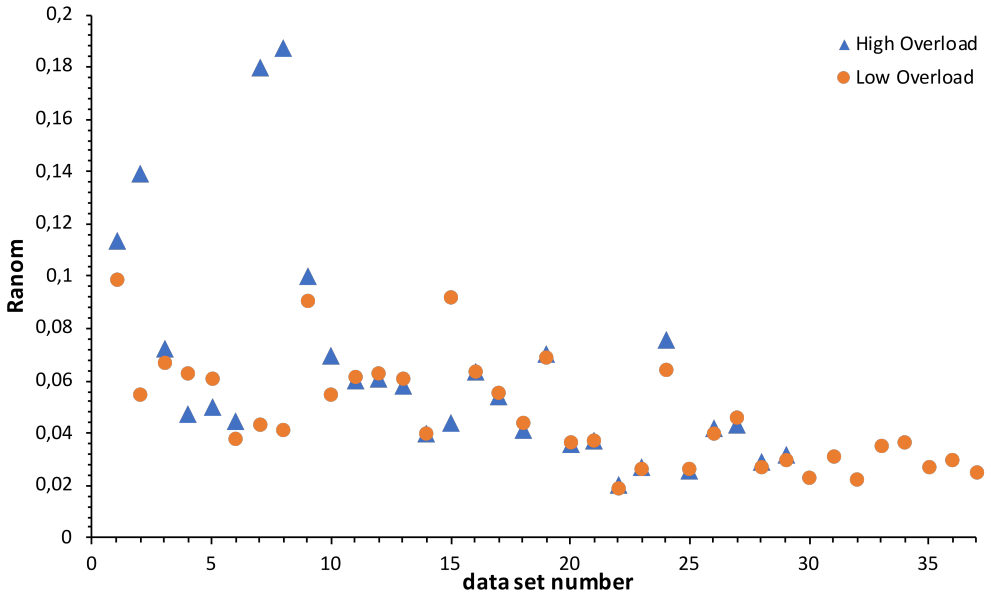


Figure A.2.: Detector parameter overload correction and it's influence on R_{anom}.

A.1.3. Absorption correction

The influence of the absorption correction employed during data reduction in XDS is evaluated. The values with absorption correction and without for the quality indicators ISa, CC_{anom}, and d"/σ(d") are listed in Table A.3. All quality indicators were calculated with XDS. The values for CC_{anom} and d"/σ(d") are given for a specific resolution. The values for ISa with and without absorption correction are plotted against the data set number (#) in Figure A.3.

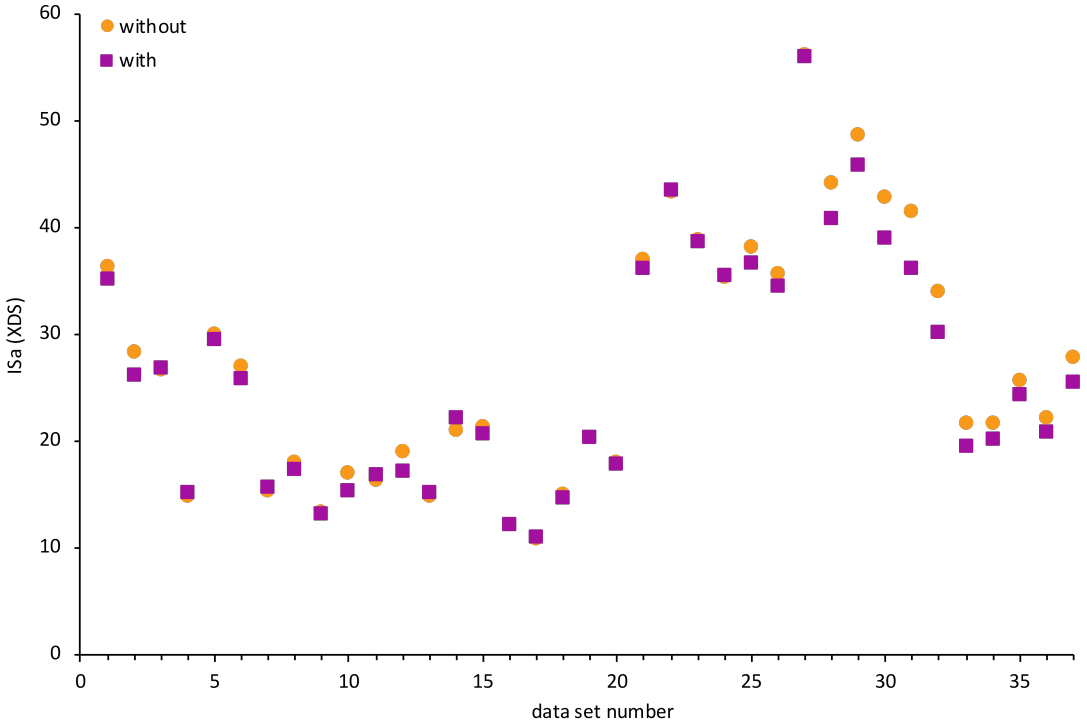


Figure A.3.: Influence of the strict absorption correction on CC_{anom} during data reduction with XDS.

Table A.3.: Influence of absorption correction during data reduction on data quality. The values for CC_{anom} and d"/σ(d") are given for a specific resolution for each data set (#).

#	ISa		resolution	CC_{anom}		d"/σ(d")	
	without	with	[Å]	without	with	without	with
1	36.27	35.18	2.57	13	19	0.977	1.099
2	28.4	26.22	2.09	11	16	1.007	1.099
3	26.61	26.86	2.32	10	18	1.08	1.167
4	14.89	15.22	2.32	21	30	0.908	0.948
5	29.97	29.5	2.55	3	10	0.93	1.034
6	27.05	25.88	2.51	9	13	0.87	0.914
7	15.35	15.73	2.25	15	16	0.811	0.806
8	17.95	17.27	2.1	12	16	0.863	0.888
9	13.29	13.12	2.14	7	13	0.937	1.014
10	16.98	15.32	2.52	2	0	0.901	0.9
11	16.34	16.84	2.09	11	12	0.998	1.025
12	19.03	17.1	2.27	12	12	0.962	1.001
13	14.88	15.17	2.28	27	22	0.943	0.959
14	21.05	22.16	2.84	27	29	1.059	0.971
15	21.3	20.73	2.18	18	17	1.007	1.029
16	12.17	12.16	3.41	16	11	1.076	1.026
17	10.91	11.06	2.46	19	22	1.123	1.142
18	14.93	14.58	2.1	24	25	1.041	1.076
19	20.3	20.33	3.41	0	3	0.93	0.954
20	17.96	17.78	2.19	13	15	0.949	0.984
21	37.02	36.21	2.69	8	7	0.944	0.977
22	43.32	43.52	3.44	4	15	0.883	0.901
23	38.88	38.65	2.4	13	17	1.044	1.049
24	35.26	35.49	2.51	5	15	0.941	1.036
25	38.17	36.69	2.4	24	28	0.96	1.035
26	35.62	34.49	2.4	20	26	1.009	1.053
27	56.22	56	3.1	18	19	0.929	0.946
28	44.15	40.88	2.2	31	33	1.108	1.142
29	48.64	45.76	2.09	19	26	1.015	1.048
30	42.8	38.94	3.73	28	14	0.915	0.915
31	41.5	36.18	2.37	27	33	1.034	1.203
32	33.98	30.08	2.37	24	28	0.93	0.999
33	21.59	19.43	3.76	10	1	0.786	0.879
34	21.74	20.23	2.33	14	14	0.944	0.959
35	25.7	24.34	2.33	27	30	0.992	1.041
36	22.09	20.85	2.37	11	19	0.927	0.997
37	27.86	25.47	2.33	40	38	0.998	1.07

A.1.4. Correlation of quality indicators

The correlation between the quality indicators Isa (XDS) and I /σ(I) with the anomalous signal specific indicator R_{anom} is evaluated. Furthermore their significance for the resulting averaged anomalous density at the positions of the phosphor atoms is discussed. In this section an overview of the indicators for all data sets is presented in Table A.4. A possible correlation is visualized in the Figures A.4, A.5, A.6, and A.7 by plotting the quality indicators against one another.

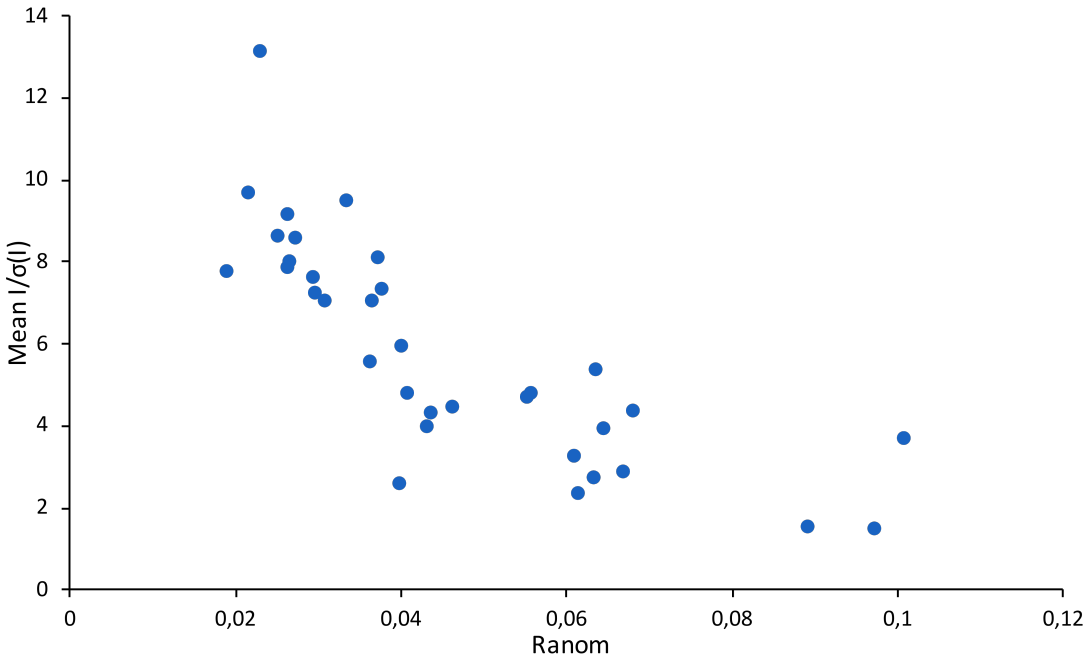


Figure A.4.: The mean I/σ(I) is plotted against the R_{anom}. The mean I/σ(I) was calculated for each data set by XPREP and the R_{anom} was calculated by Phenix_xtriage.

Table A.4.: Data quality for each scan from the Poly(rA) data sets (#). Data quality indicators are compared for correlation.

#	Isa (XDS)	mean I/σ(I) (XPREP)	R_{anom} (PHENIX)	averaged anomalous density (ANODE)
1	35.18	1.47	0.0975	1.37
2	26.22	4.65	0.0555	2.88
3	26.86	2.86	0.067	2.45
4	15.22	2.68	0.0634	2.39
5	30.38	3.25	0.061	2.35
6	25.97	7.3	0.0379	3.74
7	15.73	3.93	0.0433	3.79
8	17.27	4.76	0.0409	3.86
9	13.12	1.52	0.0894	1.55
10	15.32	4.65	0.0555	2.17
11	16.48	2.32	0.0616	2.07
12	17.1	2.68	0.0634	1.95
13	15.27	3.25	0.061	1.72
14	22.16	2.56	0.0401	2.1
15	20.73	3.67	0.101	0
16	12.16	5.34	0.0637	1.95
17	11.06	4.74	0.0558	1.65
18	14.64	4.26	0.0439	2.68
19	20.33	4.34	0.0683	1.51
20	17.78	5.55	0.0365	3.46
21	36.21	3.9	0.0646	1.59
22	43.51	7.81	0.0266	3.73
23	39.46	5.89	0.0403	2.74
24	35.49	4.42	0.0464	2.89
25	38.37	8.54	0.0274	4.06
26	35.33	7.19	0.0299	3.54
27	56.26	8.05	0.0374	2.88
28	41.27	7.75	0.0192	6.32
29	47.52	9.11	0.0266	4.29
30	38.94	13.12	0.0232	4.56
31	36.18	7	0.0309	4.8
32	30.08	9.65	0.0218	6.08
33	19.43	9.48	0.0335	3.14
34	20.23	7.01	0.0368	4.33
35	24.34	7.99	0.0268	5.51
36	20.85	7.58	0.0295	4.85
37	25.47	8.6	0.0253	5.97

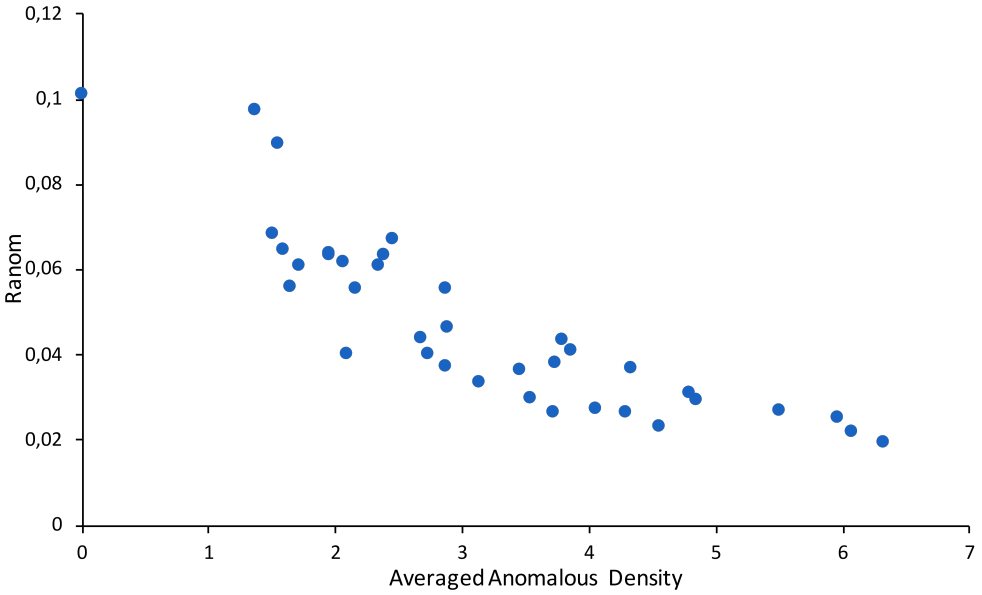


Figure A.5.: R_{anom} is plotted against the averaged anomalous density. The R_{anom} was calculated for each data set by PHENIX_xtriage and the averaged anomalous density of the phosphorus atom positions was calculated by ANODE.

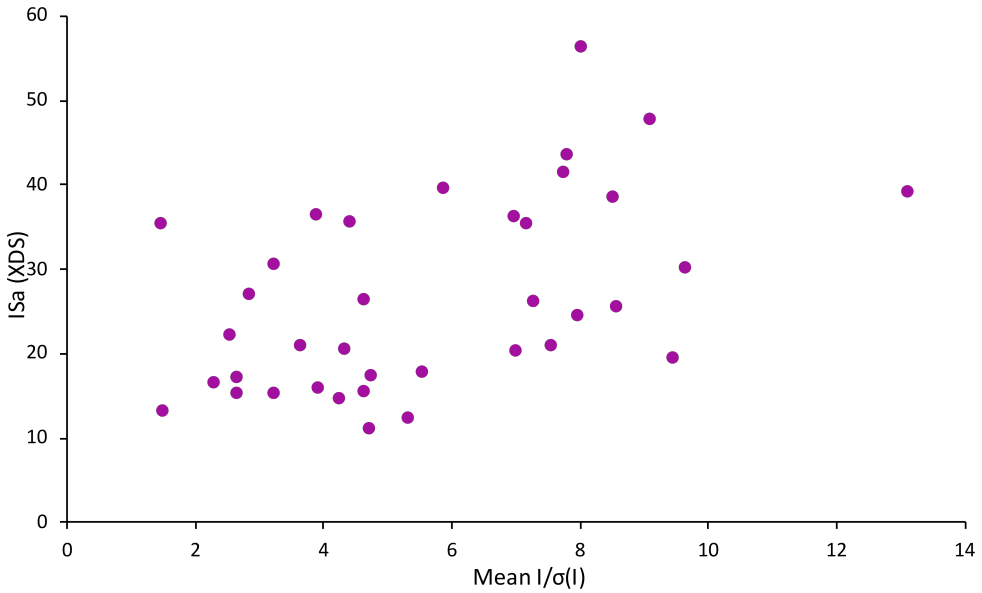


Figure A.6.: The mean I/σ(I) is plotted against the limit of the asymptotic signal-to-noise ratio (ISa). The mean I/σ(I) was calculated for each data set by XPREP and the ISa was calculated by XDS.

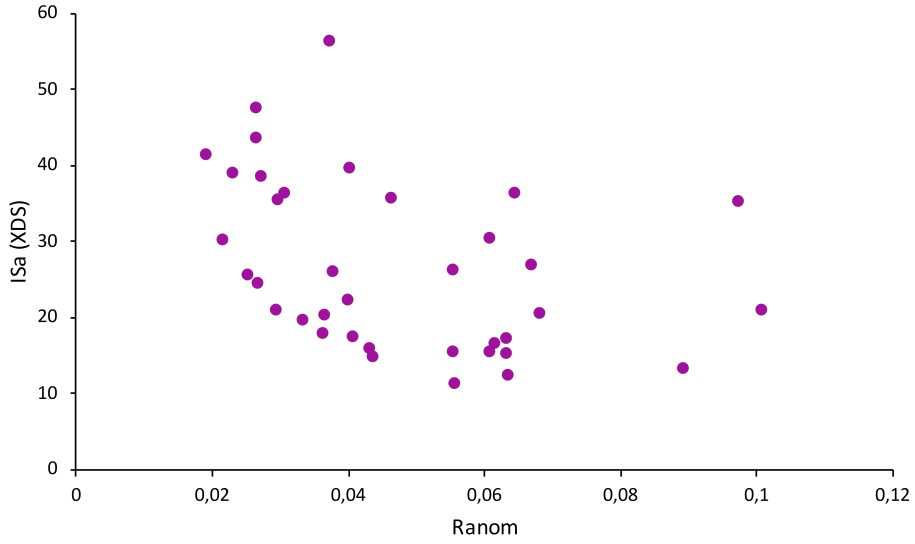


Figure A.7.: The R_{anom} is plotted against the limit of the asymptotic signal-to-noise ratio (ISa). The R_{anom} was calculated for each data set by Phenix_xtriage and the ISa was calculated by XDS.

A.1.5. Averaging statistics

The programs XRPEP, XSCALE, and PHENIX_scale_and_merge were used to prepare averaged data set combinations of varying content. Data sets for averaging were selected by various criteria. The resulting averaged combined files were analyzed for the strength of the anomalous signal. For this purpose the averaged anomalous density of the selenium atoms was calculated. The results are listed in Table A.5.

Table A.5.: For selected merged data sets the average anomalous density calculated by ANODE is listed. The data sets were merged with the programs XPREP, XSCALE or PHENIX_scale_and_merge.

data set[a]	XPREP	XSCALE	PHENIX
crystal-A1	3.24	3.12	3.49
crystal-A3	5.29	5.43	5.65
crystal-A4	3.32	3.15	4.01
crystal-B3	5.37	5.24	5.04
crystal-B4	4.34	4.27	4.37
crystal-C3	5.62	5.64	5.92
crystal-C2	5.44	7.69	8.68
crystal-C1	7.64	7.75	7.98
PetraIII-P11	3.18	3.38	5.78
SLS2015-PX10	6.31	6.32	6.37
SLS2016-PX10	7.57	7.38	8.36
SLS2016-PX06SA	6.02	7.04	9.49
All SLS	9.03	9.09	10.23
Best5MeanIs	5.55	6.15	8.53
Best10MeanIs	7.93	8.60	9.68
Best15MeanIs	8.63	9.41	10.46
Best20MeanIs	9.21	9.48	10.62
Best25MeanIs	8.91	8.88	10.70
All37	8.09	8.88	10.61

[a] The single data sets are merged by crystal (crystal-A1–C1), by beamline (PetraIII or SLS), or by the highest mean I/σ(I) value (Best5MeanIs–Best25MeanIs). The averaged data sets contain from 3–37 single data sets.

B. Appendix PDB2INS

B.1. PDB file format

The protein database (PDB) consists of record types, which can be divided into ten sections.

- title section
- primary structure section
- heterogen section
- secondary structure section
- connectivity annotation section
- miscellaneous features section
- crystallographic and coordinate transformation section
- coordinate section
- connectivity section
- bookkeeping section

Record entries can be mandatory or optional and have to appear in a specific order. Mandatory records, by record type, are listed and have to appear in a defined order. All mandatory record types are listed in Table B.1 with a short description.

Table B.1.: Keywords used in the *pdb* file format as defined and described in the PDB format guide.

record type	description[a]
HEADER	first line in file, contains PDB ID code, classification and date of deposition.
TITLE	description of the experiment.
COMPND	description of macromolecular contents.
SOURCE	biological source of the macromolecule.
KEYWDS	list of keywords describing the macromolecule.
EXPDTA	experimental technique used for structure determination.
AUTHOR	list of contributors.
REVDAT	revision data and related information.
REMARK	general remarks, subdivided by number.
SEQRES	primary sequence of backbone residues.
CRYST1	unit cell parameters, space group, and Z.
ORIGXn	(n = 1, 2, or 3) transformation from the orthogonal coordinates to the submitted coordinates.
SCALEn	(n = 1, 2, or 3) transformation from orthogonal to fractional coordinates.
MASTER	control record for book keeping.
END	last record in file.

[a] *http://www.wwpdb.org/documentation/file-format-content/format32/v3.2.html*, Worldwide Protein Data Bank Foundation, Piscataway, USA.

B.2. PDB test results

A total of 23974 data sets from the PDB were used to identify issues with the data conversion program PDB2INS or SHELXL. Of this test collection, 964 data sets resulted in an issue and did not finish a refinement with SHELXL when used in an automated fashion. Of all issues tables characterizing the issue are presented, based on rate of occurrence (Table B.2 and B.3). In the description a short form of the keywords appearing as error message are displayed in bold font. For most problems a work around is available while using PDB2INS in the interactive mode or by manual editing. Some issues require special attention.

Table B.2.: Overview of issues reported while using PDB2INS, sorted by frequency. The error keyword is given in bold font.

$\%_E$	$\%_T$	description
10.37	0.42	**no structure factor found** The inquiry file *-sf.cif* did not contain a complete set of reflection keywords. PDB2INS only converts a reflection file if a keyword set is complete. Accepted keyword sets are listed in Table 3.2.
4.67	0.19	**model error** It is legal to deposit more than one model in one *pdb* file. PDB2INS cannot handle those files and will terminate without writing a new file. Removing all but one model from the input file will enable PDB2INS to convert the file.
1.36	0.05	**Syntax error** A numerical parameter contains a non-digit character. This can occur in structure factor files when a reflection or standard deviation contains e.g. a question mark.
1.24	0.05	**Illegal atom name** For all atoms in natural amino acids a specific naming scheme is standard (Markley *et al.*, 1998). When one or more atoms are not named correctly, the program will write a warning and will mention the problematic residue.
1.14	0.05	**residue renumbering** PDB2INS is capable of renumbering residues when an insertion code is used. This is necessary since no specific method of handling insertion codes is available. When this renumbering is affected by complications, the residue that could not be renumbered is included in the error message.
0.93	0.04	**other**

$\%_E$ percent in error.

$\%_T$ percent in all test sets.

Both programs, PDB2INS and SHELXL, give extensive error messages and display the problematic section of the file. Following these hints and instructions nearly all issues can be solved with little time and effort.

Table B.3.: Overview of issues reported while using SHELXL, sorted by frequency. The error keyword is given in bold font.

$\%_E$	$\%_T$	description
31.74	1.28	**bad resi** SHELXL terminated without refinement when one of the residues has a residue name containing only digits. This is allowed in *pdb* files. PDB2INS will prompt the user to change this residue names when run in interactive mode. However, in an automated mode, PDB2INS only writes a warning but continues without renaming.
22.41	0.90	**reflection format** One reflection does not adhere to the *hkl* format as specified in section 3.3.2.
11.20	0.45	**wrong element** The *pdb* file contains an element name in the SFAC instruction line that does not correspond to the first 98 elements of the periodic table. Most commonly, one of the elements is specified as 'X', an unknown element, which does not correspond to any structure factors in SHELXL.
10.68	0.43	**wrong restraint** One or more atoms do not have the appropriate restraints specifying them. The user can check if the mentioned residues are named correctly.
5.30	0.20	**other**

$\%_E$ percent in error.

$\%_T$ percent in all test sets.

C. Appendix CLN5

C.1. Background

C.1.1. Pathogenesis of neuronal ceroid lipofuscinoses

Since the identification of the first genes causing neuronal ceroid lipofuscinoses (NCLs) a great number of mutations have been identified. With large heterogeneity the encoded proteins were identified as soluble lysosomal proteins and putative transmembrane proteins, nonetheless NCLs demonstrate a consistent morphological phenotype.

In spite of their genetic heterogeneity the NCLs share two archetypal features:

1. the accumulation of auto-fluorescent, electron dense material, containing sphingolipid activator proteins (ssaposins) A and D or subunit c of mitochondrial adenosine triphosphate synthase (ATPase), in the lysosomes of nerve cells and even many other cell types.
2. mental, motor, and visual problems caused by selective and progressive loss of neurons.

The neuronal and extra-neuronal pigment accumulations in NCLs are found in the lysosomes, the diseases are considered lysosomal storage disease (LSD). While many of the NCL proteins are present in the lysosomes, the NCLs display characteristics not typical for LSD (Mink, 2010) as cited in (Nita *et al.*, 2016). The periodic acid-Schiff (PAS)- and Sudan black B-positive accumulated material is not disease specific and takes different form when viewed under the electron microscope. Granular osmiophilic deposits (GROD), rectilinear complex (RLC), fingerprint profiles (FPP), as well as curvilinear profiles (CLP) or 'condensed forms' have been reported.

In almost all cases the disease is inherited in a recessive manner when mutations are present on both gene alleles. Exceptions occur in adult-onset NCL via *CLN4* gene mutations, were a dominant inheritance is described (Nosková *et al.*, 2011), and in a patient with complete isodisomy of chromosome 8 giving rise to homozygosity of deletion in *CLN8* (Vantaggiato *et al.*, 2009).

C.1.2. Pathogenesis of cln5

Earliest cases were almost exclusively found in Finland, giving rise to the name Finnish variant late infantile NCL (vLINCL$_{Fin}$) disease next to variant Janský-Bielschowsky disease. Beginning with slight motor mis-coordination and muscular hypotonia, the typical age of onset is 3 to 7 years for late infantile phenotype CLN5 in patients with CLN5 'Fin major' (Santavuori *et al.*, 1991, Xin *et al.*, 2010). To this day, there have been reports of an age of onset as early as 18 months and as late as 17 years, the later is classified as adult disease (Xin *et al.*, 2010). Most patients present at first with motor impairment or regression while, with some patients, visual loss or seizures were the first presenting symptom (Xin *et al.*, 2010). In 2017, a study reported a high number of patients with language decline as presenting symptom in line with decline of cognitive abilities (Simonati *et al.*, 2017).

Within 1-9 years of first clinical presentation the patients developed seizures, visual loss, motor difficulty, and cognitive regression. The disease shows rapid progression of symptoms in patients with CLN5 'Fin major' between the ages of 9 to 11 years. Also behavior disturbances were reported, such as hyperactivity, aggression, intolerance and motor stereotypes as early features in child patients (Simonati *et al.*, 2017). With disease progression anxiety, obsessive activities, hallucinations, and autistic features were observed.

Early symptoms are followed by impaired concentration, learning problems and mental retardation (Haltia, 2003). Onset of epileptic seizures in patients manifest by 7 to 8 years of age, followed by myoclonia. Some patients reported optical illusions that may be caused by seizures but are difficult to differentiate from symptoms related to declining vision. Visual deficiency can be apparent from the beginning of the disease and macular dystrophy can be found in an early stage. Progressive optic atrophy with functional blindness is found after the age of 7 to 9 years (Haltia, 2003).

While cognitive decline is an early sign, it also progresses rapidly (Santavuori *et al.*, 1991). Patients loose their ability to walk and become bedridden by the age of 9 to 13 years (Haltia, 2003, Simonati *et al.*, 2017). Later, spastic contractures develop and speech impairment leads to the inability to produce speech by the age of 11 years. Better retained is the understanding of speech, which is preserved with some patients till the age of 14 or 15 years (Mole *et al.*, 2012).

Feeding difficulties appear at the age of 9 to 13 years and can lead to weight loss due to swallowing difficulties and slow eating when the disease progresses (Mole *et al.*, 2012). The disease leads to premature death between 12 to 23 years and shows a delay of at least 6 to 8 years with adult onset between early and subsequent major symptoms

(Haltia, 2003).

The phenotype classification should be used with caution, since publications differ in the definition of 'late infantile' and 'juvenile' when referring to the age of onset. The mutation database[1] adheres to the following age classification: infantile from 6 to 18 month, late infantile from 2 to 4 years, juvenile from 5 to 10 years and adult afterwards. There are some publications classifying patients up to the age of 3, 4 or 5 as late-infantile, the early nomenclature as *variant* late infantile NCL is reminiscent of this. Other publications do not use the age of onset when describing the phenotype, a patient with cln5.014 mutation and age of onset of 3 years is still classified as the juvenile NCL phenotype (Kousi *et al.*, 2012, Xin *et al.*, 2010). Conflicting is the listing of a patient described with an age of onset of 5 years and listed as late infantile phenotype in the mutation database but referred to as juvenile NCL by Kousi *et al.* (2012).

C.1.3. Clinical features of CLN5

Magnetic resonance imaging (MRI) displays abnormalities at time of diagnosis and reveals severe cerebellar atrophy. Brain imaging studies revealed pronounced atrophy of the cerebellum and more generalized effects upon the cortical mantle (Haltia, 2003, Savukoski *et al.*, 1998). T_2-weighted images show a low thalamic signal intensity when compared to caudate nuclei signal. Also, an increased signal intensity is observed in periventricular white matter and posterior limbs of internal capsules.

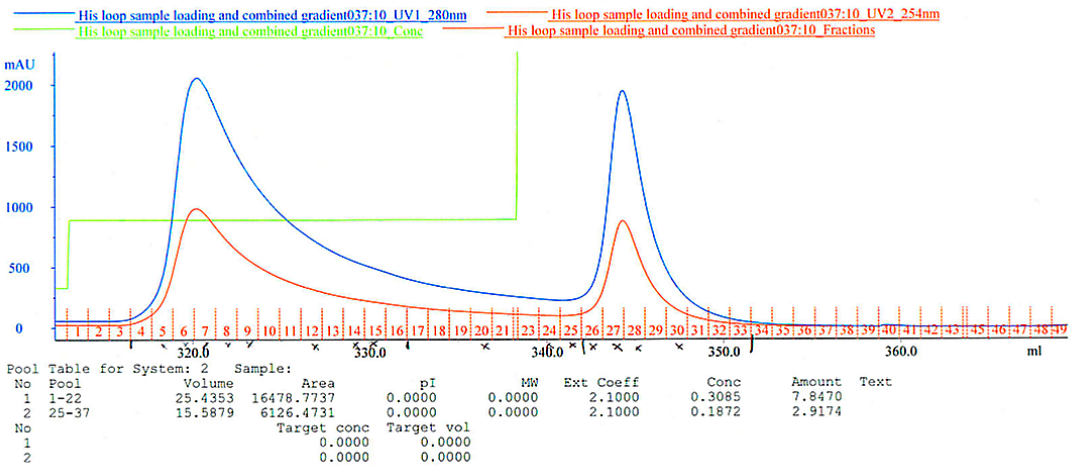
Electroretinogram (ERG) presents abnormalities at an early stage of disease and is abolished by 8 to 9.5 years (Lebrun *et al.*, 2009). Giant visual evoked potentials (VEPs) can be seen 8 – 9.5 years and the macula becomes sharply outlined.

Electroencephalogram (EEG) displays posterior spikes to low frequency photic stimulation in patients between 7 to 13 years of age (Holmberg *et al.*, 2000). Giant somatosensory evoked potentials (SEPs) were reported by the age of 7.5 to 9.5 years and often reported in patients with myoclonus (Berkovic *et al.*, 1991).

C.2. Protein purification

The chromatogram of the purification of cln5 with Ni affinity chromatography is given in Figure C.1. The step-wise gradient is depicted as green line and the UV absorption spectra of the protein at 280 nm and 254 nm are given as blue and red line, respectively. During eluation, two peaks were obtained and pooled separately.

[1] http://www.ucl.ac.uk/ncl/mutation.shtml, hosted by the MRC Laboratory for Molecular Cell Biology,



Pool Table for System: 2 Sample:

No	Pool	Volume	Area	pI	MW	Ext Coeff	Conc	Amount	Text
1	1-22	25.4353	16478.7737	0.0000	0.0000	2.1000	0.3085	7.8470	
2	25-37	15.5879	6126.4731	0.0000	0.0000	2.1000	0.1872	2.9174	

No	Target conc	Target vol
1	0.0000	0.0000
2	0.0000	0.0000

Figure C.1.: Chromatogram of cln5 purification via Ni affinity chromatography. The UV absorption at wavelength 280 nm and 254 nm is depicted.

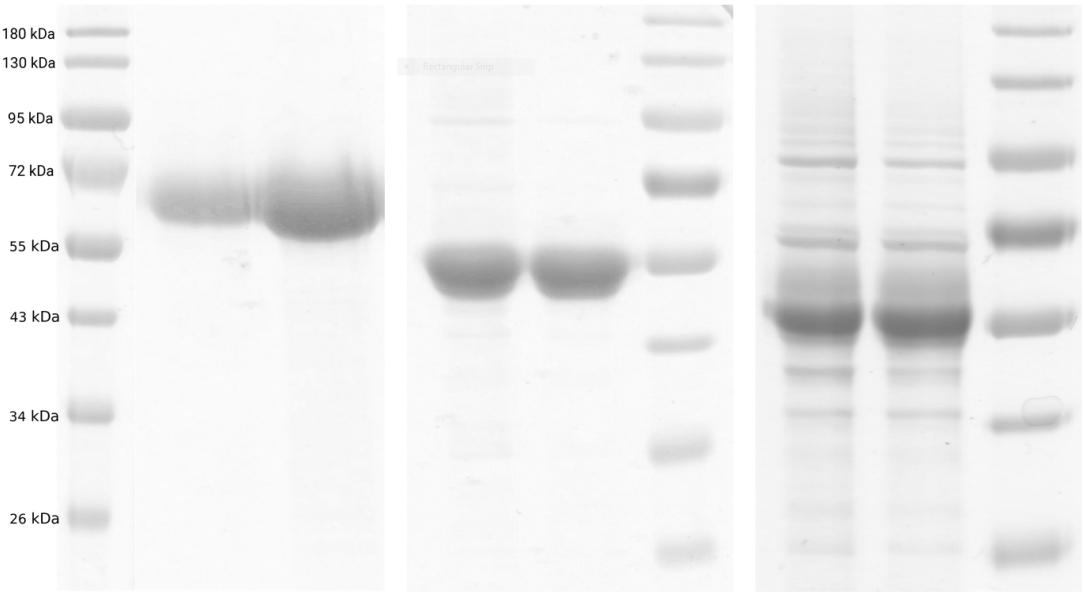


Figure C.2.: SDS-PAGE gels depicting cln5 native protein (left), cln5-kifunensine (middle), and cln5-kifunense-SeMet (right).

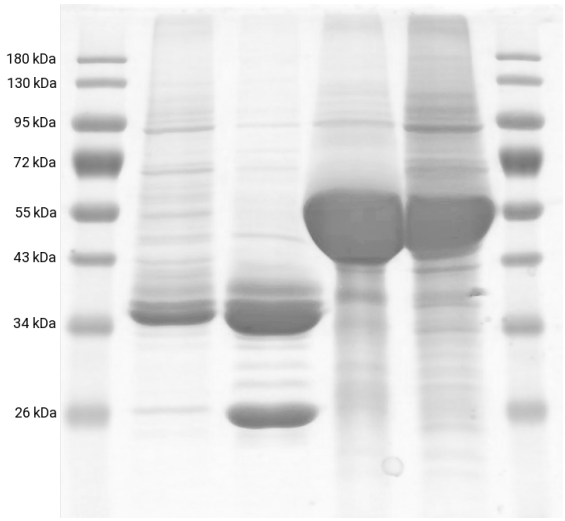


Figure C.3.: SDS-PAGE gel depicting deglycosylation of cln5 with EndoH. cln5 kifunensine protein (right) and treated with EndoH (left).

C.3. Interaction Studies

Table C.1.: Measurement of CTSD activity at different pH levels.

activity [nmol g^{-1} h^{-1}]	inactive	pH 3.0	pH 4.5
10 ng	419.1	3078.5	2700.4
ratio		1	0.88
10 ng	23097	195550	154410
ratio		1	0.79
10 ng	23097	195550	166301
ratio		1	0.85

University College London, Gower Street, London. WC1E 6BT. United Kingdom.

C.4. Circular dichroism

The circular dichroism spectra obtained from the different peaks (see Section C.2) is depicted in Figure C.4.

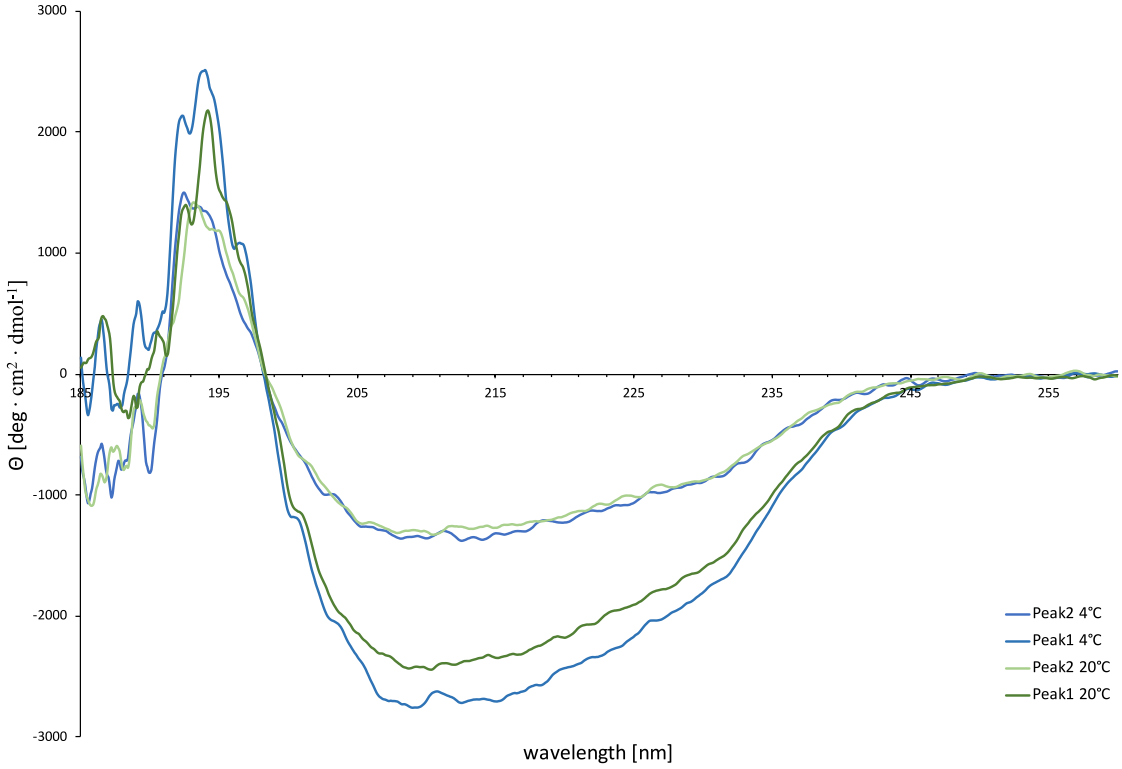


Figure C.4.: Circular dichroism spectra calculated from circular dichroism of cln5 protein at 0.1 µg/mL concentration. The two peaks obtained from Ni affinity chromatography are compared.

C.5. Data collection

the results of the on-beamline fluorescence scan is depicted in Figure C.5.

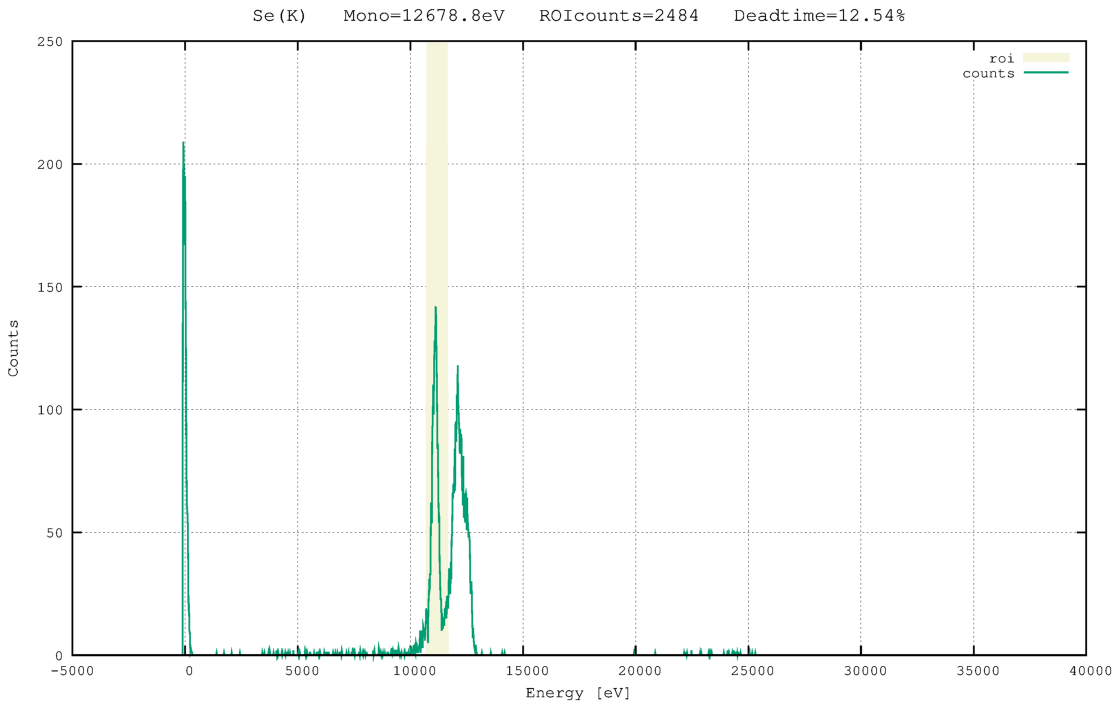


Figure C.5.: On-beamline fluorescence scan to verify the presence of significant selenium signal. The region of interest is marked.

Various resolution limits used to define the cutoff during data integration are presented for the scans of cln5k-Se crystals in Table C.4.

Table C.4.: Approximate resolution limits according to various commonly used criteria.

resolution [Å]	c8-1	c8-2	c8-3	c8-4	c11-1	c11-2	c11-3	c11-4	merged
R_{meas}≤80%	2.70	2.78	2.98	2.82	3.19	3.12	3.26	3.35	3.25
mean I/σ(I)≥2.0	2.79	2.90	3.00	2.68	3.37	3.18	3.25	3.13	2.64
$CC_{1/2}$ ≥30%	<2.50	2.58	2.75	2.38	3.04	2.87	<2.92	2.84	<2.50

Table C.2.: Data collection statistics for all data sets measured from crystal cln5k-Se-8.

data statistics	cln5k-Se-8 scan1	cln5k-Se-8 scan2	cln5k-Se-8 scan3	cln5k-Se-8 scan4
X-ray source	PSI SLS PII-X10SA			
detector	Pilatus 6M (25 Hz)			
wavelength [Å]	0.97898		0.978985	
space group	$P3_221$			
unit cell [Å]	a = 58.42 c = 179.05	a = 58.48 c = 179.19	a = 58.47 c = 179.15	a = 58.51 c = 179.22
number of observations	52146	49255	89216	236419
number of unique reflections	10252	10340	9335	12254
redundancy	5.0	4.8	9.6	19.3
resolution range[a] [Å]	48.67–2.65 (2.74–2.65)	48.70–2.70 (2.79–2.70)	48.73–2.80 (2.89–2.80)	48.76–2.55 (2.64–2.55)
completeness[a] (all) [%]	96.56 (97.62)	99.75 (99.70)	99.99 (100)	99.99 (100)
completeness (anomalous) [%]	95.72	97.41	100	100
mean intensity	6.9	8.0	6.5	4.3
mean I/σ(I)[a]	7.6 (1.2)	7.7 (1.1)	9.3 (1.3)	13.1 (1.3)
$CC_{1/2}$[a]	99.3 (63.9)	99.6 (46.7)	99.7 (55.6)	99.9 (68.9)
ISa (XDS)	26.04	30.93	23.48	24.09
R_{merge}[a]	0.143 (1.104)	0.136 (1.081)	0.180 (1.433)	0.170 (1.894)
R_{meas}[a]	0.159 (1.231)	0.152 (1.277)	19.1 (1.515)	0.174 (1.943)
R_{pim}[a]	0.069 (0.532)	0.068 (0.546)	6.1 (0.485)	0.039 (0.430)

[a] highest resolution shell in parenthesis.

Table C.3.: Data collection statistics for all data sets measured from crystal cln5k-Se-11.

data statistics	cln5k-Se-11 scan1	cln5k-Se-11 scan2	cln5k-Se-11 scan3	cln5k-Se-11 scan4
X-ray source	PSI SLS PII-X10SA			
detector	Pilatus 6M (25 Hz)			
wavelength [Å]	0.97898		0.978985	
space group	$P3_221$			
unit cell [Å]	a = 58.42	a = 58.14	a = 58.42	a = 58.41
	c = 179.05	c = 177.62	c = 178.40	c = 178.27
number of observations	31057	34579	78145	148584
number of unique reflections	6667	7472	8365	7962
redundancy	4.7	4.6	9.3	18.7
resolution range[a] [Å]	48.43–3.10	48.44–3.00	48.67–2.90	48.66–2.95
	(3.19–3.10)	(3.09–3.00)	(2.99–2.90)	(3.04–2.95)
completeness[a] (all) [%]	97.97 (98.97)	99.83 (100)	99.99 (100)	99.99 (100)
completeness (anomalous) [%]	98.05	98.88	100	100
mean intensity	16.4	15.5	16.2	6.1
mean I/σ(I)[a]	7.3 (1.2)	7.9 (1.3)	9.2 (1.2)	10.1 (1.3)
$CC_{1/2}$[a]	99.7 (61.0)	99.7 (65.9)	99.8 (61.3)	99.8 (63.3)
ISa (XDS)	29.68	33.19	21.04	21.34
R_{merge}[a]	0.176 (1.278)	0.146 (1.036)	0.185 (1.694)	0.254 (2.504)
R_{meas}[a]	0.198 (1.427)	0.164 (1.161)	0.196 (1.787)	0.261 (2.569)
R_{pim}[a]	0.088 (0.622)	0.075 (0.515)	0.064 (0.564)	0.060 (0.572)

[a] highest resolution shell in parenthesis.

C.6. Data merging and refinement

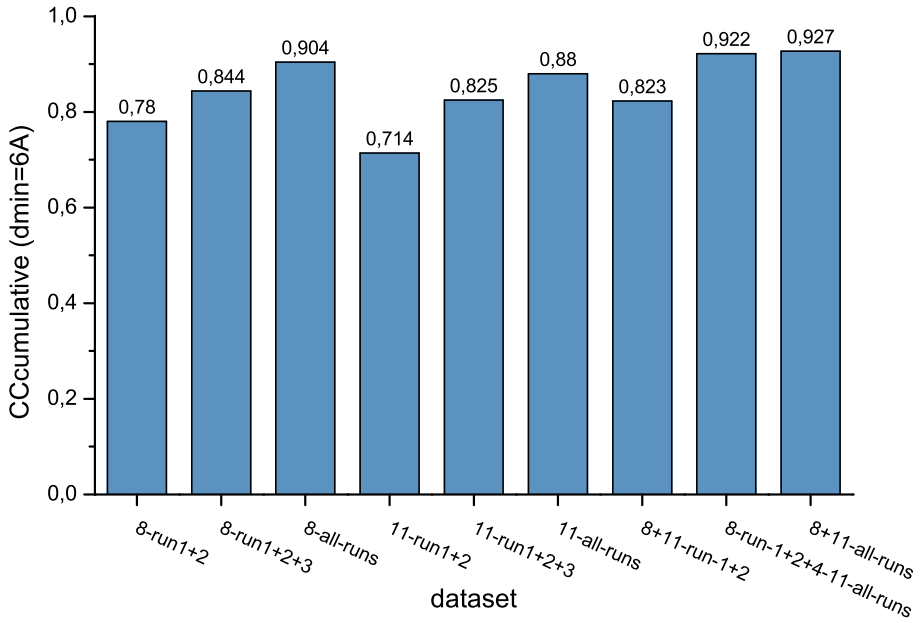


Figure C.6.: Value of $CC_{cumulative}$ at 6.00 Å for various combinations of cln5Se data collected from two crystals. The value at highest resolution should give a good indicator of the quality of the merged file. All values above 0.5 are considered a good match.

Table C.5.: Refinement of the occupancy of the selenium/sulfur sides by step-wise addition of refined residues in SHELXL. The occupancy for the selenium containing residues is listed per refinement cycle. No number was given for residues with an occupancy of one that was not refined.

occupancy [%]	MET residue number						
cycle	202	383	135	165	240	182	244
1	73.37						
2	79.43	96.12					
3	78.73	95.93	106.5				
4	79.41	100.5	-	98.22			
5	97.38	-	-	91.83	68.01	89.01	
6	107.4	-	-	89.00	65.00	95.09	89.36

C.7. Structure

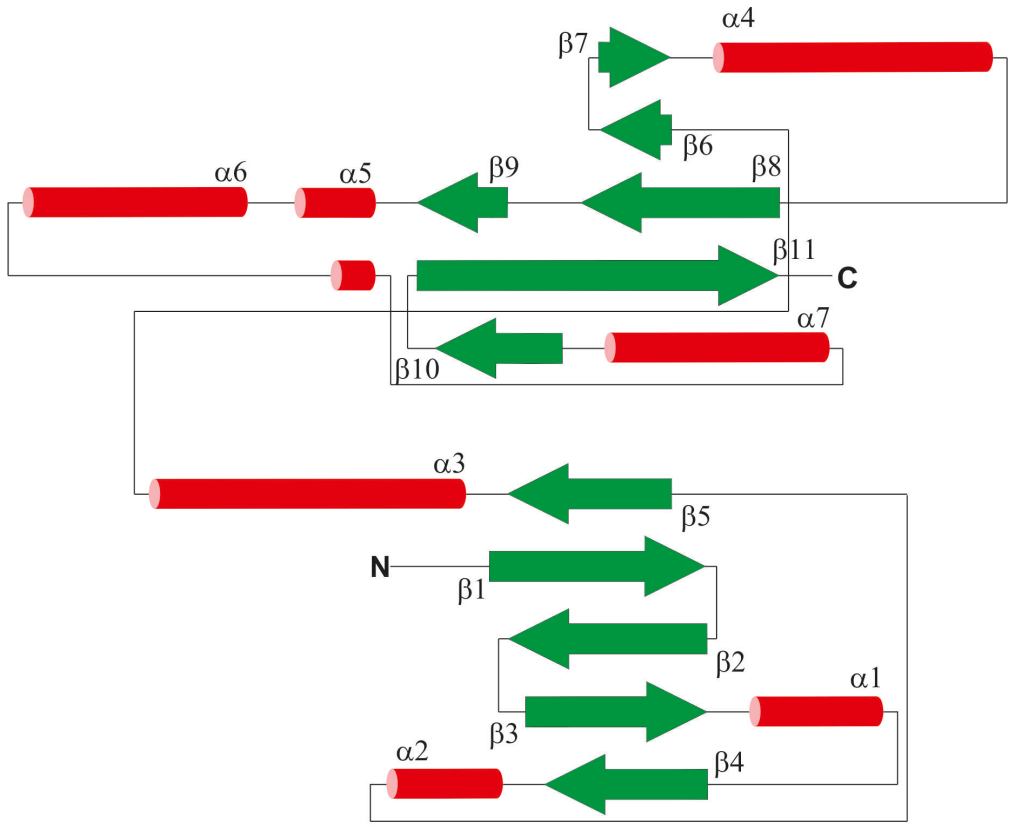


Figure C.7.: Topology diagram of the cln5 structure (from PDBsum (Laskowski, 2001)).

The average B-factor of the structure is 57 Å^2. Some regions of the structure have a significantly higher B-factor. A B-factor of 79 Å^2 corresponds to a thermal displacement of around 1 Å but a high value can also indicate disorder. The B-factor of the main chain is displayed in Figure C.8.

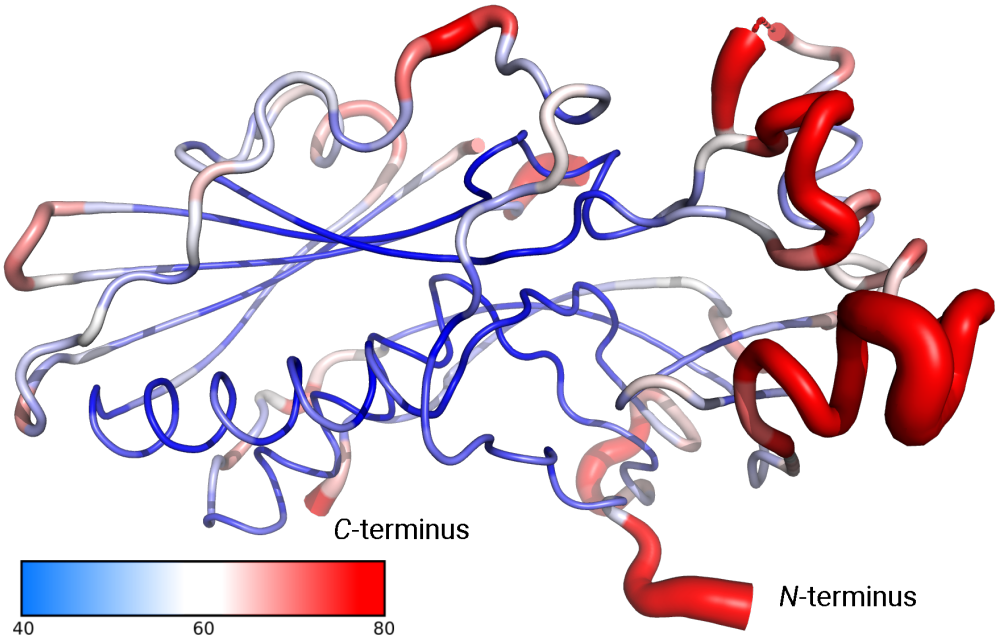
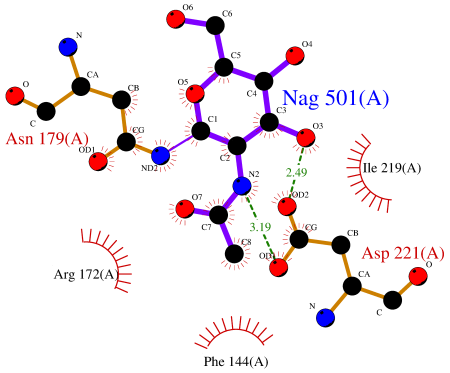


Figure C.8.: The main chain of cln5 is colored by B-factor from blue (40 Å^2) to red (80 Å^2).

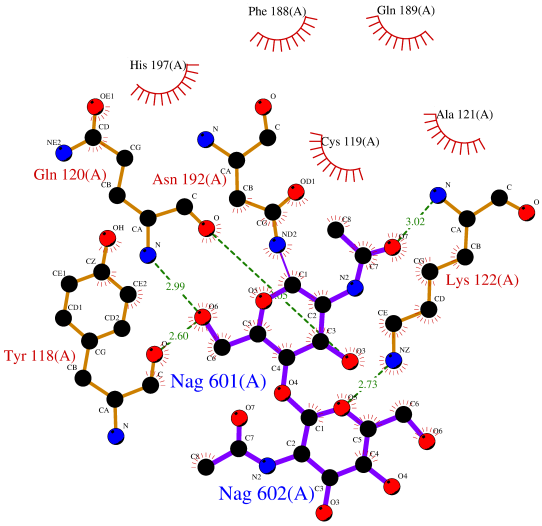
C.7.1. Sugar modifications

The following diagrams (Figure C.9) of the sugar modification have been created with the web service PDBsum[2] (Laskowski, 2001). The nearest residues for the sugar modifications at Asn179 (C.9a), Asn192 (C.9b), Asn252 (C.9c), and Asn179 (C.9d) are depicted.

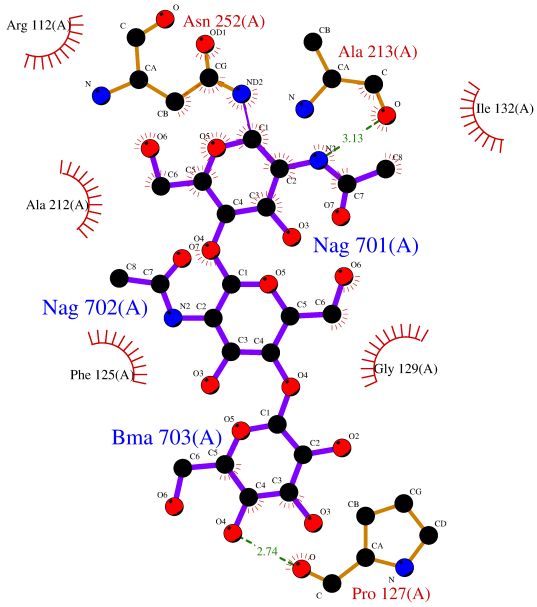
[2] *http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html.*



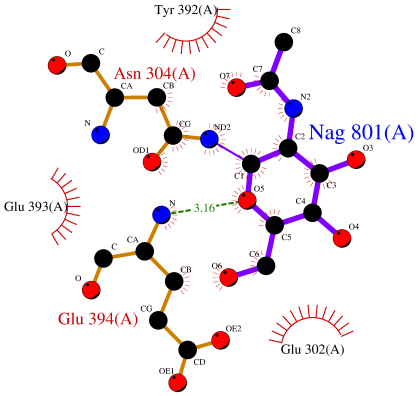
(a) Sugar modification NAG501 at Asn179.



(b) Sugar modification NAG601 at Asn192.

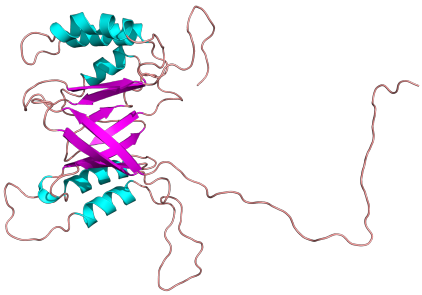


(c) Sugar modification NAG701 at Asn252.

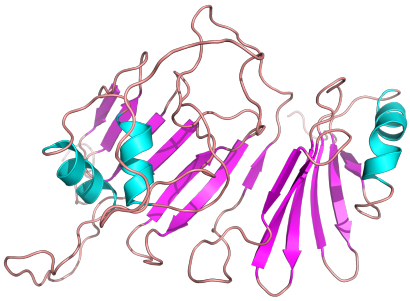


(d) Sugar modification NAG801 at Asn304.

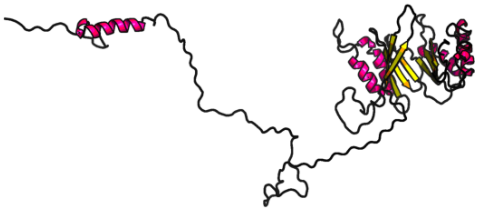
Figure C.9.: An overview of the sugar modifications in the structure of cln5 and their surrounding residues, generated with PDBsum LigPlot (Laskowski, 2001).



(a) Model 1 predicted from the full cln5 sequence (amino acids 1–407).



(b) Model 2 predicted from the full cln5 sequence (amino acids 1–407).



(c) Model 3 predicted from the shortened cln5 sequence (amino acids 49–407).

(d) Model 4 predicted from the shortened cln5 sequence (amino acids 93–407).

Figure C.10.: RaptorX predictions of cln5 protein structure calculated from different sequence ranges. The secondary structure elements are highlighted.

C.8. Structure prediction

C.8.1. Structure prediction methods

Huber and Mathavarajah (2018) reported an analysis of the predicted secondary structure of cln5 with amino acids 49 to 407, the 'pre-processed form' of cln5[3]. Based on this work the RaptorX server was used here on the full sequence (407 amino acids), the shortened sequence used by Huber and Mathavarajah (2018) (amino acids 49–407) and the even shorter mature protein reported by Jules *et al.* (2017) (amino acids 93–407). The secondary structure predictions obtained from the RaptorX web server in this work were visually evaluated and compared to the published prediction (see Figures 4.2 and C.10). None of the various predictions showed a visual match to the published prediction of the secondary structure.

[3]UniProt accession number O75503.

The calculated models displayed an unordered secondary structure for the larger part of the protein sequence. For the remaining part mostly a mixture of short α helices and small antiparallel β sheets was predicted. This stands in contrast to the results reported before (Huber and Mathavarajah, 2018). A search for binding pockets was not successful with the calculated models.

The secondary structure prediction tool based on the sequence, HHPred was employed with the full sequence of cln5[4]. HHPred reported the first match PPPDE2 (2wq7_A) with an identity of 15% and a similarity of 0.113 giving a sum probability of 53.4 (see Table C.6). The second hit Lmo2511 protein (3k2t_A) was calculated with an identity of 19% and a similarity of 0.425 leading to a sum probability of 21.3. These probability scores are far below the recommended 95% for a nearly certain homology. The E-value, as a measure of statistical significance, evaluates how many chance hits in the database better than the present one could be obtained. Only a match with a high probability and an E-value below one should be considered, therefore the results were categorized as not significant.

Table C.6.: Statistics for the homology based secondary structure prediction reported by HHPred (Soding *et al.*, 2005).

PDB-ID	Prob	E-value	P-value	Score	SS	Aligned Cols	Query HMM	Template HMM
2wq7_A	53.4	84	0.0015	28.2	6.7	115	50-208	3-119 (168)
3k2t_A	21.3	0.022	0.0039	21.3	3.1	26	263-293	11-36 (57)

I-Tasser generates secondary and tertiary structure models based on the calculated alignments[5]. The models generated on the template alignment matches to cln5 are displayed in Figure C.11. The quality of the prediction is measured by a confidence score (C-score) ranging from -5 to 2, with higher values signifying a model with a high confidence. The template modeling (TM) score is a scale of measure for the similarity between two structures and a TM-score of $\geq$ 0.5 indicates a model of correct topology.

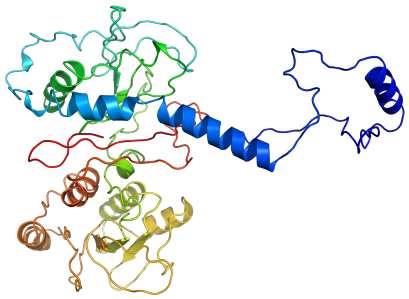
All obtained models displayed a low confidence score and a low score on parameters describing the accuracy of the modeling. These results were not considered further.

Results obtained from SWISS-MODEL are depicted in Figure C.12 (Waterhouse *et al.*, 2018).

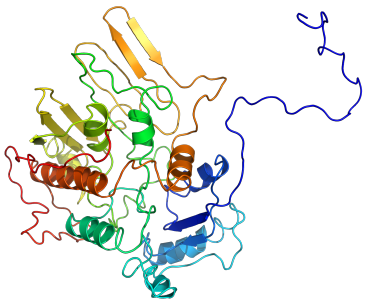
[4]The search was conducted on the 28th November 2017.

[5]The search was conducted on 24th January 2016.

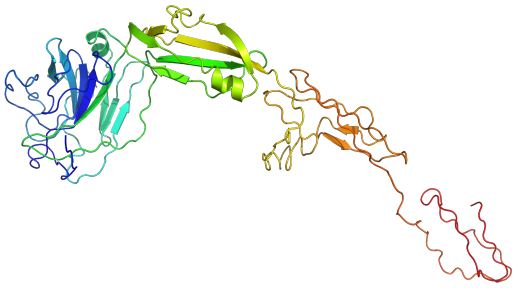
(a) Model 1, C-score -4.56.



(b) Model 2, C-score -3.70.



(c) Model 3, C-score -4.20.



(d) Model 4, C-score -4.02.

Figure C.11.: Models based on the secondary structure prediction by I-TASSER. The TM score for the first model was reported as 0.24.

C.9. Structure similarity

C.9.1. Three dimensional structural overlay of cln5 with NlpC/P60 superfamily proteins

Additional structural overlays based on the structure similarity are depicted in this section. Figure C.13 shows the structural overlay of cln5 with selected proteins of the NlpC/P60 super family.

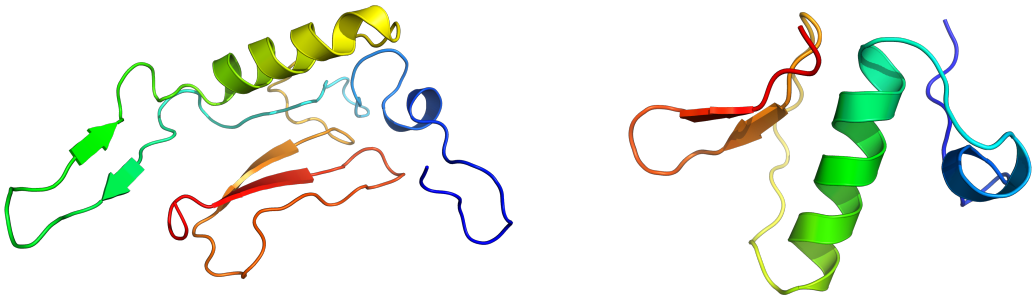
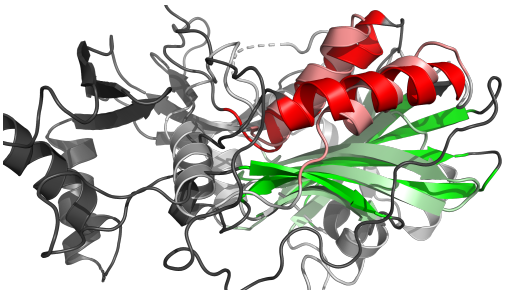
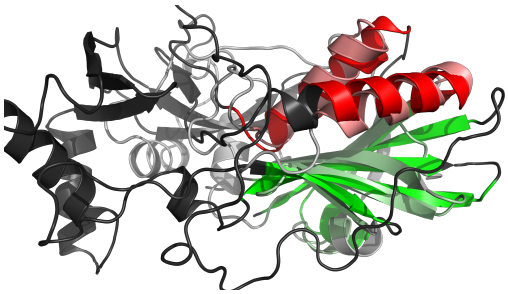


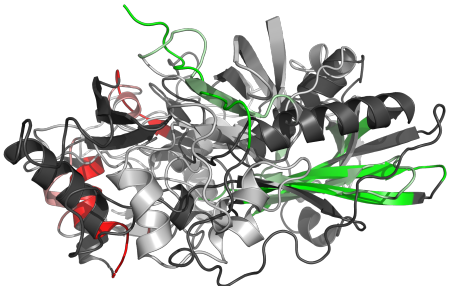
Figure C.12.: Secondary structure prediction by SWISS-MODEL. Only for part of the sequence was a target found and a secondary structure predicted.



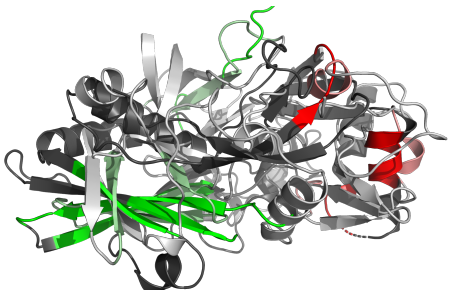
(a) Overlay of cln5 and BcPPNE (ekw0).



(b) Overlay of cln5 and Yiix (2if6).



(c) Overlay of cln5 and BcYkfC (3h41).



(d) Overlay of cln5 and BcYkfC (3h41).

Figure C.13.: Conserved secondary structure across NlpC/P60 superfamily: similarity as calculated by PDBeFold is marked. cln5 in dark gray/red/green, PPNE protein in light grey/light red/light green. PDB codes are given in parantheses.

C.10. Graphics Software

Table C.7.: Graphics software used for processing and picture generation.

software	usage	distributor /reference
ChemDraw	2D molecule figures	PerkinElmer, Waltham, USA.
Coot	molecule visualization	(Emsley *et al.*, 2010).
GIMP	graphic editing	The GIMP Team, *https://www.gimp.org/*.
HKL2MAP	quality indicator plots (SHELXC/D/E)	(Pape and Schneider, 2004).
ImageJ	SDS-gel and western blot analysis	MRC Laboratory of Molecular Biology, Cambridge, England and University of York, York, England.
Inkscape	topology diagrams	The Inkscape Project, *https://inkscape.org*.
Origin	plots from data	OriginLab, Northampton, USA.
PyMol	molecule rendering	DeLano Scientific LLC, Schrödinger , Cambridge, USA.
Xprep	quality indicator plots	Bruker AXS, Madison, USA.

Bibliography

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. and Zwart, P. H. (2010). *Acta Crystallographica*, **D66**(2), 213–221.

Agirre, J., Davies, G. J., Wilson, K. S. and Cowtan, K. D. (2017). *Current Opinion in Structural Biology*, **44**, 39–47.

Agirre, J., Iglesias-Fernández, J., Rovira, C., Davies, G. J., Wilson, K. S. and Cowtan, K. D. (2015). *Nature structural & molecular biology*, **22**(11), 833–4.

Akey, D. L., Brown, W. C., Konwerski, J. R., Ogata, C. M. and Smith, J. L. (2014). *Acta Crystallographica*, **D70**(10), 2719–2729.

Akey, D. L., Terwilliger, T. C. and Smith, J. L. (2016). *Acta Crystallographica*, **D72**(3), 296–302.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). *Journal of Molecular Biology*, **215**(3), 403–410.

Anantharaman, V. and Aravind, L. (2003). *Genome biology*, **4**(2), R11.

Anderson, G. W., Goebel, H. H. and Simonati, A. (2013). *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1832**(11), 1807–1826.

Arakaki, N., Nagao, T., Niki, R., Toyofuku, A., Tanaka, H., Kuramoto, Y., Emoto, Y., Shibata, H., Magota, K. and Higuti, T. (2003). *Molecular Cancer Research*, **1**(13), 931–939.

Aramini, J. M., Rossi, P., Huang, Y. J., Zhao, L., Jiang, M., Maglaqui, M., Xiao, R., Locke, J., Nair, R., Rost, B., Acton, T. B., Inouye, M. and Montelione, G. T. (2008). *Biochemistry*, **47**(37), 9715–9717.

Armstrong, D., Koppang, N. and Rider, J. (eds.) (1982). *Ceroid-lipofuscinosis (Batten's disease)*. Amsterdam: Elsevier Biomedical Press.

Arnold, E., Himmel, D. M. and Rossmann, M. G. (eds.) (2012). *International Tables for Crystallography*, vol. F. Chester, England: International Union of Crystallography.

Aroyo, M. I. (ed.) (2016). *International Tables for Crystallography*, vol. A. Chester, England: International Union of Crystallography.

Assmann, G., Brehm, W. and Diederichs, K. (2016). *Journal of Applied Crystallography*, **49**, 1021–1028.

Barton, W. A., Tzvetkova-Robev, D., Erdjument-Bromage, H., Tempst, P. and Nikolov, D. B. (2006). *Protein Science*, **15**(8), 2008–2013.

Batten, F. E. (1903). *Trans Ophthalmol Soc UK*, **23**, 386–390.

Batten, F. E. (1909). *Proceedings of the Royal Society of Medicine*, **2**, 35.

Benes, P., Vetvicka, V. and Fusek, M. (2008). *Critical Reviews in Oncology/Hematology*, **68**(1), 12–28.

Berg, J. M., Tymoczko, J. L. and Stryer, L. (2012). *Biochemistry*. New York: W.H. Freeman Palgrave MacMillan, 7th ed.

Berkovic, S. F., So, N. K. and Andermann, F. (1991). *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, **8**(3), 261–74.

Berman, H., Henrick, K. and Nakamura, H. (2003). *Nature Structural Biology*, **10**, 980.

Bessa, C., Teixeira, C. A. F., Mangas, M., Dias, A., Sá Miranda, M. C., Guimarães, A., Ferreira, J. C., Canas, N., Cabral, P. and Ribeiro, M. G. (2006). *Molecular Genetics and Metabolism*, **89**(3), 245–253.

Bielschowsky, M. (1913). *Zeitschrift für Nervenheilkunde*, **50**, 7–29.

Bijelic, A. and Rompel, A. (2017). *Accounts of Chemical Research*, **50**(6), 1441–1448.

Bonifacino, J. S. and Hurley, J. H. (2008). *Current Opinion in Cell Biology*, **20**(4), 427–436.

Borchardt-Ott, W. (2012). *Crystallography*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bragg, W. L. (1962). In *Fifty years of X-ray diffraction*, edited by P. P. Ewald, chap. 8, pp. 120–135. Uetrecht: International Union of Crystallography.

Brünger, A. T. (1992). *X-PLOR v.3.1. A System for X-ray Crystallography and NMR*. New Haven: Yale University Press.

Brunger, A. T. (2007). *Nature Protocols*, **2**, 2728.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. and Warren, G. L. (1998). *Acta Crystallographica*, **D54**(5), 905–921.

Cannelli, N., Nardocci, N., Cassandrini, D., Morbin, M., Aiello, C., Bugiani, M., Criscuolo, L., Zara, F., Striano, P., Granata, T., Bertini, E., Simonati, A. and Santorelli, F. (2007). *Neuropediatrics*, **38**(1), 46–49.

Cárcel-Trullols, J., Kovács, A. D. and Pearce, D. A. (2015). *Biochimica et Biophysica Acta - Molecular Basis of Disease*, **1852**(10), 2242–2255.

Casanas, A., Warshamanage, R., Finke, A. D., Panepucci, E., Olieric, V., Nöll, A., Tampé, R., Brandstetter, S., Förster, A., Mueller, M., Schulze-Briese, C., Bunk, O. and Wang, M. (2016). *Acta Crystallographica*, **D72**(9), 1036–1048.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. and Richardson, D. C. (2010). *Acta Crystallographica*, **D66**(1), 12–21.

Cismondi, I. A., Cannelli, N., Aiello, C., Santorelli, F. M., Kohan, R., Oller Ramírez, A. M. and Halac, I. N. (2008). *Human Genetics*, **123**(5), 554.

Corpet, F. (1988). *Nucleic Acids Research*, **16**(22), 10881–10890.

Cowtan, K. (2006). *Acta Crystallographica*, **D62**(9), 1002–1011.

Danyukova, T., Ariunbat, K., Thelen, M., Brocke-Ahmadinejad, N., Mole, S. E. and Storch, S. (2018). *Human Molecular Genetics*, **27**(10), 1711–1722.

D'Arcy, A., Bergfors, T., Cowan-Jacob, S. W. and Marsh, M. (2014). *Acta crystallographica*, **F70**(9), 1117–26.

Das, A., Jolly, R. and Kohlschütter, A. (1999). *Molecular Genetics and Metabolism*, **66**(4), 349–355.

Dauter, Z. and Adamiak, D. A. (2001). *Acta Crystallographica*, **D57**(7), 990–995.

Dauter, Z. and Jaskolski, M. (2010). *Journal of Applied Crystallography*, **43**(5), 1150–1171.

Diederichs, K. (2009). *Acta Crystallographica*, **D65**(6), 535–542.

Diederichs, K. (2010). *Acta Crystallographica*, **D66**(6), 733–740.

Diederichs, K. (2016). *Nucleic Acid Crystallography: Methods and Protocols*, **1320**.

Diederichs, K. (2017). *Acta Crystallographica*, **D73**(4), 286–293.

Diederichs, K. and Karplus, P. A. (1997). *Nature Structural Biology*, **4**(4), 269–275.

Diederichs, K. and Karplus, P. A. (2013). *Acta Crystallographica*, **D69**(7), 1215–1222.

Diederichs, K., McSweeney, S. and Ravelli, R. B. G. (2003). *Acta Crystallographica*, **D59**(5), 903–909.

Duke, E. M. H. and Johnson, L. N. (2010). *Proceedings of the Royal Society A*, **466**(2124), 3421–3452.

Dutta, S., Burkhardt, K., Young, J., Swaminathan, G. J., Matsuura, T., Henrick, K., Nakamura, H. and Berman, H. M. (2009). *Molecular Biotechnology*, **42**(1), 1–13.

Dyken, P. R. (1989). *Journal of child neurology*, **4**(3), 165–174.

Eckert, M. (2012). *Annalen der Physik*, **524**(5), A83–A85.

Eiberg, H., Gardiner, R. M. and Mohr, J. (2008). *Clinical Genetics*, **36**(4), 217–218.

Einspahr, H. M. and Weiss, M. S. (2012). In *International Tables for Crystallography Volume F*, edited by E. Arnold, D. M. Himmel and M. Rossmann, chap. 2.2, pp. 64–74. Chester, England: International Union of Crystallography, 2nd ed.

Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010). *Acta Crystallographica*, **D66**(4), 486–501.

Engh, R. A. and Huber, R. (1991). *Acta Crystallographica*, **A47**(4), 392–400.

Engilberge, S., Riobé, F., Di Pietro, S., Lassalle, L., Coquelle, N., Arnaud, C.-A., Pitrat, D., Mulatier, J.-C., Madern, D., Breyton, C., Maury, O. and Girard, E. (2017). *Chemical Science*, **8**(9), 5909–5917.

Eskelinen, E. L. and Saftig, P. (2009). *Biochimica et Biophysica Acta - Molecular Cell Research*, **1793**(4), 664–673.

Evans, P. R. (2006). *Acta Crystallographica*, **D62**(1), 72–82.

Evans, P. R. (2011). *Acta Crystallographica*, **D67**(4), 282–292.

Evans, P. R. and Murshudov, G. N. (2013). *Acta Crystallographica*, **D69**(7), 1204–1214.

Fischer, W. and Koch, E. (2006). In *International Tables for Crystallography Volume A*, edited by T. Hahn, chap. Chapter 11, pp. 810–816. Dordrecht: Springer Netherlands, 5th ed.

Garman, E. and Murray, J. W. (2003). *Acta Crystallographica - Section D Biological Crystallography*, **59**(11), 1903–1913.

Getty, A. L. and Pearce, D. A. (2011). *Cellular and Molecular Life Sciences*, **68**(3), 453–474.

Giacovazzo, C., Monaco, H. L., Artioli, G., Viterbo, D., Milanesio, M., Gilli, G., Gilli, P., Zanotti, G., Ferraris, G. and Catti, M. (2011). *Fundamentals of Crystallography*. Oxford University Press.

Gillies, J. and Hochstrasser, M. (2012). *EMBO reports*, **13**(4), 284–285.

Goebel, H. (1997). *Neuropediatrics*, **28**(01), 67–68.

Goebel, H. H., Mole, S. E. and Lake, B. D. (eds.) (1999). *The Neuronal Ceroid Lipofuscinoses (Batten Disease)*. Amsterdam: IOS Press, 1st ed.

Greenfield, N. J. (2006). *Nature Protocols*, **1**(6), 2876–2890.

Grosse-Kunstleve, R. W. and Adams, P. D. (2002). *Journal of Applied Crystallography*, **35**(4), 477–480.

Gruene, T., Hahn, H. W., Luebben, A. V., Meilleur, F. and Sheldrick, G. M. (2013). *Journal of Applied Crystallography*, **47**(1), 462–466.

Grune, T. (2008). *Journal of Applied Crystallography*, **41**(1), 217–218.

Gulick, A. M., Horswill, A. R., Thoden, J. B., Escalante-Semerena, J. C. and Rayment, I. (2002). *Acta Crystallographica*, **D58**(2), 306–309.

Haddad, S. E., Khoury, M., Daoud, M., Kantar, R., Harati, H., Mousallem, T., Alzate, O., Meyer, B., Boustany, R.-M. M., El Haddad, S., Khoury, M., Daoud, M., Kantar, R., Harati, H., Mousallem, T., Alzate, O., Meyer, B. and Boustany, R.-M. M. (2012). *Electrophoresis*, **33**(24), 3798–3809.

Haidar, B., Kiss, R. S., Sarov-Blat, L., Brunet, R., Harder, C., McPherson, R. and Marcel, Y. L. (2006). *Journal of Biological Chemistry*, **281**(52), 39971–39981.

Haltia, M. (2003). *Journal of Neuropathology & Experimental Neurology*, **62**(1), 1–13.

Haltia, M., Rapola, J. and Santavuori, P. (1973). *Acta Neuropathologica*, **26**(2), 157–170.

Harp, J., Pallan, P. and Egli, M. (2016). *Crystals*, **6**(10), 125.

Hasegawa, H. and Holm, L. (2009). *Current Opinion in Structural Biology*, **19**(3), 341–348.

Hendrickson, W. A. and Teeter, M. M. (1981). *Nature*, **290**(5802), 107–113.

Hirokawa, T., Boon-Chieng, S. and Mitaku, S. (1998). *Bioinformatics*, **14**(4), 378–379.

Hofman, S. L. and Peltonen, L. (2002). In *The Metabolic and Bolecular Basis of Inherited Disease*, edited by C. R. Scriver, A. L. Beaudet, W. Sly, D. Valle, B. Childs, K. W. Kinzler and B. Vogelstein, pp. 3877–3894. New York: McGraw-Hill, 8th ed.

Holbrook, S. R. and Kim, S.-H. (2004). *Biopolymers*, **44**(1), 3–21.

Holm, L., Kaariainen, S., Rosenstrom, P. and Schenkel, A. (2008). *Bioinformatics*, **24**(23), 2780–2781.

Holm, L. and Laakso, L. M. (2016). *Nucleic Acids Research*, **44**(W1), W351–W355.

Holmberg, V., Jalanko, A., Isosomppi, J., Fabritius, A.-L., Peltonen, L. and Kopra, O. (2004). *Neurobiology of Disease*, **16**(1), 29–40.

Holmberg, V., Lauronen, L., Autti, T., Santavuori, P., Savukoski, M., Uvebrant, P., Hofman, I., Peltonen, L. and Jarvela, I. (2000). *Neurology*, **55**(4), 579–581.

Huang, C.-Y., Olieric, V., Ma, P., Panepucci, E., Diederichs, K., Wang, M. and Caffrey, M. (2015). *Acta Crystallographica*, **D71**(6), 1238–1256.

Huber, R. J. and Mathavarajah, S. (2018). *Cellular Signalling*, **42**, 236–248.

Hübschle, C. B., Sheldrick, G. M. and Dittrich, B. (2011). *Journal of Applied Crystallography*, **44**(6), 1281–1284.

Hülsen, G., Broennimann, C., Eikenberry, E. F. and Wagner, A. (2006). *Journal of Applied Crystallography*, **39**(4), 550–557.

Isosomppi, J., Vesa, J., Jalanko, A. and Peltonen, L. (2002). *Human molecular genetics*, **11**(8), 885–891.

Iyer, L. M., Koonin, E. V. and Aravind, L. (2004). *Cell Cycle*, **3**(11), 1440–1450.

Jabs, S., Quitsch, A., Kkel, R., Koch, B., Tyynel, J., Brade, H., Glatzel, M., Walkley, S., Saftig, P., Vanier, M. T. and Braulke, T. (2008). *Journal of Neurochemistry*, **106**(3), 1415–1425.

Janský, J. (1908). *Sborn Lék*, **13**, 85–139.

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

Jules, F., Sauvageau, E., Dumaresq-Doiron, K., Mazzaferri, J., Haug-Kröper, M., Fluhrer, R., Costantino, S. and Lefrancois, S. (2017). *Experimental Cell Research*, **357**(1), 40–50.

Kabeya, Y. (2000). *The EMBO Journal*, **19**(21), 5720–5728.

Kabsch, W. (2010). *Acta Crystallographica*, **D66**(2), 125–132.

Käkelä, R., Somerharju, P. and Tyynelä, J. (2003). *Journal of Neurochemistry*, **84**(5), 1051–1065.

Karplus, K., Barrett, C. and Hughey, R. (1998). *Bioinformatics*, **14**(10), 846–856.

Karplus, P. A. and Diederichs, K. (2012). *Science*, **336**(6084), 1030–1033.

Karplus, P. A. and Diederichs, K. (2015). *Current Opinion in Structural Biology*, **34**, 60–68.

Kim, B.-W., Choo, H.-J., Lee, J.-W., Kim, J.-H. and Ko, Y.-G. (2004). *Experimental & Molecular Medicine*, **36**(5), 476–485.

Klepeis, J. L., Wei, Y., Hecht, M. H. and Floudas, C. A. (2004). *Proteins: Structure, Function, and Bioinformatics*, **58**(3), 560–570.

Kollmann, K., Mutenda, K. E., Balleininger, M., Eckermann, E., von Figura, K., Schmidt, B. and Lübke, T. (2005). *Proteomics*, **5**(15), 3966–3978.

Köpfer, D. A., Song, C., Gruene, T., Sheldrick, G. M., Zachariae, U. and de Groot, B. L. (2014). *Science*, **346**(6207), 352–355.

Kousi, M., Lehesjoki, A.-E. and Mole, S. E. (2012). *Human Mutation*, **33**(1), 42–63.

Kousi, M., Siintola, E., Dvorakova, L., Vlaskova, H., Turnbull, J., Topcu, M., Yuksel, D., Gokben, S., Minassian, B. A., Elleder, M., Mole, S. E. and Lehesjoki, A.-E. (2009). *Brain*, **132**(3), 810–819.

Kraft, P., Bergamaschi, A., Broennimann, C., Dinapoli, R., Eikenberry, E. F., Henrich, B., Johnson, I., Mozzanica, A., Schlepütz, C. M., Willmott, P. R. and Schmitt, B. (2009). *Journal of Synchrotron Radiation*, **16**(3), 368–375.

Krissinel, E. and Henrick, K. (2004). *Acta Crystallographica*, **D60**(12), 2256–2268.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001). *Journal of Molecular Biology*, **305**(3), 567–580.

Kufs, H. (1925). *Zeitschrift für die gesamte Neurologie und Psychiatrie*, **95**(1), 169–188.

Kuronen, M., Hermansson, M., Manninen, O., Zech, I., Talvitie, M., Laitinen, T., Gröhn, O., Somerharju, P., Eckhardt, M., Cooper, J. D., Lehesjoki, A.-E., Lahtinen, U. and Kopra, O. (2012). *Neuropathology and Applied Neurobiology*, **38**(5), 471–486.

Laemmli, U. K. (1970). *Nature*, **227**(5259), 680–685.

Langer, G., Cohen, S. X., Lamzin, V. S. and Perrakis, A. (2008). *Nature Protocols*, **3**(7), 1171–1179.

Larkin, H., Ribeiro, M. G. and Lavoie, C. (2013). *Human Mutation*, **34**(12), 1688–1697.

Laskowski, R. A. (2001). *Nucleic acids research*, **29**(1), 221–222.

Lebrun, A.-H., Storch, S., Rüschendorf, F., Schmiedt, M.-L., Kyttälä, A., Mole, S. E., Kitzmüller, C., Saar, K., Mewasingh, L. D., Boda, V., Kohlschütter, A., Ullrich, K., Braulke, T. and Schulz, A. (2009). *Human Mutation*, **30**(5), E651—-E661.

Leonarski, F., Redford, S., Mozzanica, A., Lopez-Cuenca, C., Panepucci, E., Nass, K., Ozerov, D., Vera, L., Olieric, V., Buntschu, D., Schneider, R., Tinti, G., Froejdh, E., Diederichs, K., Bunk, O., Schmitt, B. and Wang, M. (2018). *Nature Methods*, **15**(10), 799–804.

Lerner, T. J., Boustany, R.-M. N., Anderson, J. W., D'Arigo, K. L., Schlumpf, K., Buckler, A. J., Gusella, J. F. and Haines, J. L. (1995). *Cell*, **82**(6), 949–957.

Liu, Q., Dahmane, T., Zhang, Z., Assur, Z., Brasch, J., Shapiro, L., Mancia, F. and Hendrickson, W. A. (2012). *Science*, **336**(6084), 1033–1037.

Liu, Q., Zhang, Z. and Hendrickson, W. A. (2011). *Acta Crystallographica*, **D67**(1), 45–59.

Liu, W., Tian, F., Wang, X., Yu, H. and Bi, Y. (2013). *Chem. Comm.* **49**(29), 2983–5.

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

Luebben, J. and Gruene, T. (2015). *Proceedings of the National Academy of Sciences*, **112**(29), 8999–9003.

Lyly, A., Marjavaara, S. K., Kyttälä, A., Uusi-Rauva, K., Luiro, K., Kopra, O., Martinez, L. O., Tanhuanpää, K., Kalkkinen, N., Suomalainen, A., Jauhiainen, M. and Jalanko, A. (2008). *Human Molecular Genetics*, **17**(10), 1406–1417.

Lyly, A., von Schantz, C., Heine, C., Schmiedt, M.-L., Sipilä, T., Jalanko, A. and Kyttälä, A. (2009). *BMC Cell Biology*, **10**, 83.

Maley, F., Trimble, R. B., Tarentino, A. L. and Plummer, T. H. (1989). *Analytical Biochemistry*, **180**(2), 195–204.

Mamo, A., Jules, F., Dumaresq-Doiron, K., Costantino, S. and Lefrancois, S. (2012). *Molecular and Cellular Biology*, **32**(10), 1855–1866.

Mark, B. L., Mahuran, D. J., Cherney, M. M., Zhao, D., Knapp, S. and James, M. N. (2003). *Journal of Molecular Biology*, **327**(5), 1093–1109.

Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E. and Wuethrich, K. (1998). *Pure And Applied Chemistry*, **70**(1), 117–142.

Markmann, S., Thelen, M., Cornils, K., Schweizer, M., Brocke-Ahmadinejad, N., Willnow, T., Heeren, J., Gieselmann, V., Braulke, T. and Kollmann, K. (2015). *Traffic*, **16**(7), 743–759.

Martinez, L. O., Jacquet, S., Esteve, J.-P., Rolland, C., Cabezón, E., Champagne, E., Pineau, T., Georgeaud, V., Walker, J. E., Tercé, F., Collet, X., Perret, B. and Barbaras, R. (2003). *Nature*, **421**(6918), 75–79.

Massa, W. (2009). *Kristallstrukturbestimmung*. Wiesbaden: Vieweg+Teubner, 6th ed.

McCoy, A. J. (2007). *Acta Crystallographica*, **D63**(1), 32–41.

Moews, P. and Kretsinger, R. (1975). *Journal of Molecular Biology*, **91**(2), 201–225.

Moharir, A., Peck, S. H., Budden, T. and Lee, S. Y. (2013). *PLoS ONE*, **8**(9), e74299.

Mole, S. (2004). *European Journal of Paediatric Neurology*, **8**(2), 101–103.

Mole, S., Williams, R. and Goebel, H. (2012). *The Neuronal Ceroid Lipofuscinoses (Batten Disease)*, vol. 1. Oxford: Oxford University Press, 2nd ed.

Mueller, M., Wang, M. and Schulze-Briese, C. (2012). *Acta Crystallographica*, **D68**(1), 42–56.

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

Muhammad, A., Flores, I., Zhang, H., Yu, R., Staniszewski, A., Planel, E., Herman, M., Ho, L., Kreber, R., Honig, L. S., Ganetzky, B., Duff, K., Arancio, O. and Small, S. A. (2008). *Proceedings of the National Academy of Sciences*, **105**(20), 7327–7332.

Munford, R. S., Sheppard, P. O. and O'Hara, P. J. (1995). *Journal of Lipid Research*, **36**(8), 1653–1663.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. and Vagin, A. A. (2011). *Acta Crystallographica*, **D67**(4), 355–367.

Nicholls, R. A., Long, F. and Murshudov, G. N. (2012). *Acta Crystallographica*, **D68**(4), 404–417.

Nijssen, P. C., Brusse, E., Leyten, A. C., Martin, J., Teepen, J. L. and Roos, R. A. (2002). *Movement Disorders*, **17**(3), 482–487.

Nita, D. A., Mole, S. E. and Minassian, B. A. (2016). *Epileptic Disorders*, **18**(S2), 73–88.

North, A. C. T. (1965). *Acta Crystallographica*, **18**(2), 212–216.

Nosková, L., Stránecký, V., Hartmannová, H., Přistoupilová, A., Barešová, V., Ivánek, R., Hůlková, H., Jahnová, H., van der Zee, J., Staropoli, J. F., Sims, K. B., Tyynelä, J., Van Broeckhoven, C., Nijssen, P. C., Mole, S. E., Elleder, M. and Kmoch, S. (2011). *The American Journal of Human Genetics*, **89**(2), 241–252.

Oram, J. F., Wolfbauer, G., Vaughan, A. M., Tang, C. and Albers, J. J. (2003). *Journal of Biological Chemistry*, **278**(52), 52379–52385.

Pai, C. H., Chiang, B. Y., Ko, T. P., Chou, C. C., Chong, C. M., Yen, F. J., Chen, S., Coward, J. K., Wang, A. H. and Lin, C. H. (2006). *EMBO Journal*, **25**(24), 5970–5982.

Palmer, D. N., Bayliss, S. L., Clifton, P. A. and Grant, V. J. (1993). *Journal of Inherited Metabolic Disease*, **16**(2), 292–295.

Palmer, D. N., Husbands, D. R., Winter, P. J., Blunt, J. W. and Jolly, R. D. (1986). *Journal of Biological Chemistry*, **261**(4), 1766–1772.

Palmer, D. N., Oswald, M. J., Westlake, V. J. and Kay, G. W. (2002). *Archives of Gerontology and Geriatrics*, **34**(3), 343–357.

Panjikar, S. and Tucker, P. A. (2002). *Journal of Applied Crystallography*, **35**(2), 261–266.

Pape, T. and Schneider, T. R. (2004). *Journal of Applied Crystallography*, **37**(5), 843–844.

Pattabiraman, N. (1986). *Biopolymers*, **25**, 1603–1606.

Pineda-Trujillo, N., Cornejo, W., Carrizosa, J., Wheeler, R. B., Múnera, S., Valencia, A., Agudelo-Arango, J., Cogollo, A., Anderson, G., Bedoya, G., Mole, S. E. and Ruíz-Linares, A. (2005). *Neurology*, **64**(4), 740–742.

Pohlmann, R., Boeker, M. W. C. and von Figura, K. (1995). *Journal of Biological Chemistry*, **270**(45), 27311–27318.

Porebski, P. J., Cymborowski, M., Pasenkiewicz-Gierula, M. and Minor, W. (2016). *Acta Crystallographica*, **D72**(2), 266–280.

Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P. R., Jenkins, H. T., Keegan, R., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S. J., Nicholls, R. A., Noble, M., Pannu, N. S., Roth, C., Sheldrick, G., Skubak, P., Turkenburg, J., Uski, V., von Delft, F., Waterman, D., Wilson, K., Winn, M. and Wojdyr, M. (2018). *Acta Crystallographica*, **D74**(2), 68–84.

Qureshi, Y. H., Patel, V. M., Berman, D. E., Kothiya, M. J., Neufeld, J. L., Vardarajan, B., Tang, M., Reyes-Dumeyer, D., Lantigua, R., Medrano, M., Jiménez-Velázquez, I. J., Small, S. A. and Reitz, C. (2018). *Molecular and Cellular Biology*, **38**(20), e00011–18.

Reyes, F. E., Garst, A. D. and Batey, R. T. (2009). In *Methods in Enzymology*, vol. 469, chap. 6, pp. 119–139. Elsevier Inc., 1st ed.

Rice, L. M., Earnest, T. N. and Brunger, A. T. (2000). *Acta Crystallographica*, **D56**(11), 1413–1420.

Rich, A., Davies, D. R., Crick, F. H. and Watson, J. D. (1961). *Journal of Molecular Biology*, **3**(1), 71–86.

Rossmann, M. G. (1961). *Acta Crystallographica*, **14**(4), 383–388.

Roy, S. C., Pratt, R. H. and Kissel, L. (1993). *Radiation Physics and Chemistry*, **41**(4-5), 725–738.

Rupp, B. (2011). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. New York: Garland Science.

Safaee, N., Noronha, A. M., Rodionov, D., Kozlov, G., Wilds, C. J., Sheldrick, G. M. and Gehring, K. (2013). *Angewandte Chemie - International Edition*, **52**(39), 10370–10373.

Šali, A. and Blundell, T. L. (1993). *Journal of Molecular Biology*, **234**(3), 779–815.

Sammito, M., Meindl, K., de Ilarduya, I. M., Millán, C., Artola-Recolons, C., Hermoso, J. A. and Usón, I. (2014). *The FEBS journal*, **281**(18), 4029–4045.

Santavuori, P., Rapola, J., Nuutila, A., Raininko, R., Lappi, M., Launes, J., Herva, R. and Sainio, K. (1991). *Neuropediatrics*, **22**(02), 92–96.

Santavuori, P., Rapola, J., Sainio, K. and Raitta, C. (1982). *Neuropediatrics*, **13**(03), 135–141.

Sarma, G. N. and Karplus, P. A. (2006). *Acta Crystallographica*, **D62**(7), 707–716.

Savukoski, M., Kestilä, M., Williams, R., Järvelä, I., Sharp, J., Harris, J., Santavuori, P., Gardiner, M. and Peltonen, L. (1994). *American Journal of Human Genetics*, **55**(4), 695–701.

Savukoski, M., Klockars, T., Holmberg, V., Santavuori, P., Lander, E. S. and Peltonen, L. (1998). *Nature Genetics*, **19**(3), 286–288.

Schmiedt, M. L., Blom, T., Blom, T., Kopra, O., Wong, A., von Schantz-Fant, C., Ikonen, E., Kuronen, M., Jauhiainen, M., Cooper, J. D. and Jalanko, A. (2012). *Neurobiology of Disease*, **46**(1), 19–29.

Schmiedt, M.-L. L., Bessa, C., Heine, C., Ribeiro, M. G., Jalanko, A. and Kyttälä, A. (2010). *Human Mutation*, **31**(3), 356–365.

Schneider, T. R. and Sheldrick, G. M. (2002). *Acta Crystallographica*, **D58**(10), 1772–1779.

Shabalin, I. G., Porebski, P. J. and Minor, W. (2018). *Crystallography Reviews*, **24**(4), 236–262.

Shakked, Z. (1983). *Acta Crystallographica*, **A 39**(3), 278–279.

Sheldrick, G. M. (1996). *Crystallographic Computing*, pp. 4–6.

Sheldrick, G. M. (2010). *Acta Crystallographica*, **D66**(4), 479–485.

Sheldrick, G. M. (2015). *Acta crystallographica*, **C71**(1), 3–8.

Sheldrick, G. M. and Schneider, T. R. (1997). *Methods in Enzymology*, **277**, 319–43.

Shin, E. J., Shin, H. M., Nam, E., Kim, W. S., Kim, J. H., Oh, B. H. and Yun, Y. (2012). *EMBO Reports*, **13**(4), 339–346.

Simonati, A., Williams, R. E., Nardocci, N., Laine, M., Battini, R., Schulz, A., Garavaglia, B., Moro, F., Pezzini, F. and Santorelli, F. M. (2017). *Developmental Medicine and Child Neurology*, **59**(8), 815–821.

Simpkin, A. J., Simkovic, F., Thomas, J. M. H., Savko, M., Lebedev, A., Uski, V., Ballard, C., Wojdyr, M., Wu, R., Sanishvili, R., Xu, Y., Lisa, M.-N., Buschiazzo, A., Shepard, W., Rigden, D. J. and Keegan, R. M. (2018). *Acta Crystallographica*, **D74**(7), 595–605.

Sleat, D. E. (1997). *Science*, **277**(5333), 1802–1805.

Sleat, D. E., Ding, L., Wang, S., Zhao, C., Wang, Y., Xin, W., Zheng, H., Moore, D. F., Sims, K. B. and Sims, K. B. (2009). *Molecular & Cellular Proteomics*, **8**(7), 1708–1718.

Sleat, D. E., Wang, Y., Sohar, I., Lackland, H., Li, Y., Li, H., Zheng, H. and Lobel, P. (2006). *Molecular & Cellular Proteomics*, **5**(10), 1942–1956.

Soding, J., Biegert, A. and Lupas, A. N. (2005). *Nucleic Acids Research*, **33**(suppl_2), W244–W248.

Spielmeyer, W. (1905). *Neurol Cbl*, **24**, 620–621.

Steinfeld, R., Steinke, H. B., Isbrandt, D., Kohlschütter, A. and Gärtner, J. (2004). *Human Molecular Genetics*, **13**(20), 2483–2491.

Stengel, O. C. (1826). *Eyr Medicinsk Tidskrift*, **1**, 347–352.

Suh, H.-Y. Y., Kim, J.-H. H., Woo, J.-S. S., Ku, B., Shin, E. J., Yun, Y. and Oh, B.-H. H. (2012). *Proteins: Structure, Function, and Bioinformatics*, **80**(8), 2099–2104.

Sygusch, J. and Allaire, M. (1988). *Acta Crystallographica*, **A44**(4), 443–448.

Tanida, I., Ueno, T. and Kominami, E. (2004). *The International Journal of Biochemistry & Cell Biology*, **36**(12), 2503–2518.

Tarentino, A. L., Gomez, C. M. and Plummer, T. H. (1985). *Biochemistry*, **24**(17), 4665–4671.

Taylor, G. L. (2010). *Acta Crystallographica*, **D66**(4), 325–338.

Teng, T.-y. and Moffat, K. (2000). *Journal of Synchrotron Radiation*, **7**(5), 313–317.

Terry, R. D. and Korey, S. R. (1960). *Nature*, **188**(4755), 1000–1002.

Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. and Adams, P. D. (2016). *Acta Crystallographica*, **D72**(3), 346–358.

Thorn, A. and Sheldrick, G. M. (2011). *Journal of Applied Crystallography*, **44**(6), 1285–1287.

Tronrud, D. E. (1997). In *Methods in Enzymology - Volume 277*, edited by C. W. J. Carter and R. M. Sweet, chap. 16, pp. 306–319. Elsevier Inc.

Tronrud, D. E. (2007). *Macromolecular Crystallography Protocols, Volume 2*, **364**, 231–254.

Tronrud, D. E., Ten Eyck, L. F. and Matthews, B. W. (1987). *Acta Crystallographica*, **A43**(4), 489–501.

Tyynelä, J., Palmer, D. N., Baumann, M. and Haltia, M. (1993). *FEBS Letters*, **330**(1), 8–12.

Vagin, A. and Teplyakov, A. (2010). *Acta Crystallographica*, **D66**(1), 22–25.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. and Murshudov, G. N. (2004). *Acta Crystallographica*, **D60**(12), 2184–2195.

Vantaggiato, C., Redaelli, F., Falcone, S., Perrotta, C., Tonelli, A., Bondioni, S., Morbin, M., Riva, D., Saletti, V., Bonaglia, M. C., Giorda, R., Bresolin, N., Clementi, E. and Bassi, M. T. (2009). *Human Mutation*, **30**(7), 1104–1116.

Varilo, T., Savukoski, M., Norio, R., Santavuori, P., Peltonen, L. and Järvelä, I. (1996). *American journal of human genetics*, **58**(3), 506–512.

Vesa, J., Chin, M. H., Oelgeschläger, K., Isosomppi, J., DellAngelica, E. C., Jalanko, A. and Peltonen, L. (2002). *Molecular Biology of the Cell*, **13**(7), 2410–2420.

Vesa, J., Hellsten, E., Verkruyse, L. A., Camp, L. A., Rapola, J., Santavuori, P., Hofmann, S. L. and Peltonen, L. (1995). *Nature*, **376**(6541), 584–587.

Vesa, J. and Peltonen, L. (2002). *Current Molecular Medicine*, **2**(5), 439–44.

Vogt, H. (1906). *European Neurology*, **18**(2), 161–171.

Wang, J. (2010). *Acta Crystallographica*, **D66**(9), 988–1000.

Wang, Z., Zhao, F., Peng, J. and Xu, J. (2011). *Proteomics*, **11**(19), 3786–3792.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T. (2018). *Nucleic Acids Research*, **46**(W1), W296–W303.

Weiss, M. S. (2001). *Journal of Applied Crystallography*, **34**(2), 130–135.

Weiss, M. S. and Hilgenfeld, R. (1997). *Journal of Applied Crystallography*, **30**(2), 203–205.

Westhof, E. and Sundaralingam, M. (1980). *Proceedings of the National Academy of Sciences*, **77**(4), 1852–1856.

Williams, R., Topçu, M., Lake, B., Mitchell, W. and Mole, S. (1999). In *The neuronal ceroid lipofuscinoses (Batten disease)*, edited by H. Goebel, S. Mole and B. Lake, pp. 114–116. Amsterdam: IOS Press.

Williams, R. E., Goebel, H. H., Mole, S. E., Boustany, R.-M., Elleder, M., Kohlschütter, A., Mink, J. W., Niezen-de Boer, R. and Simonati, A. (2011). In *The Neuronal Ceroid Lipofuscinoses (Batten Disease)*, pp. 20–23. Oxford University Press.

Williams, R. E. and Mole, S. E. (2012). *Neurology*, **79**(2), 183–191.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. a., Powell, H. R., Read, R. J., Vagin, A. and Wilson, K. S. (2011). *Acta Crystallographica*, **D67**, 235–42.

Winter, E. and Ponting, C. P. (2002). *Trends in Biochemical Sciences*, **27**(8), 381–383.

Wisniewski, K. E., Kida, E., Golabek, A. A., Kaczmarski, W., Connell, F. and Zhong, N. (2001). In *Batten Disease: Diagnosis, Treatment, and Research*, edited by K. E. Wisniewski and N. Zhong, vol. 45, pp. 1–34. Academic Press.

Wlodawer, A. (2007). *Acta Crystallographica*, **D63**(3), 421–423.

Xie, X., Wang, X., Jiang, D., Wang, J., Fei, R., Cong, X., Wei, L., Wang, Y. and Chen, H. (2017). *Biochemical and Biophysical Research Communications*, **488**(2), 291–296.

Xin, W., Mullen, T. E., Kiely, R., Min, J., Feng, X., Cao, Y., O'Malley, L., Shen, Y., Chu-Shore, C., Mole, S. E., Goebel, H. H. and Sims, K. (2010). *Neurology*, **74**(7), 565–571.

Xu, Q., Abdubek, P., Astakhova, T., Axelrod, H. L., Bakolitsa, C., Cai, X., Carlton, D., Chen, C., Chiu, H. J., Chiu, M., Clayton, T., Das, D., Deller, M. C., Duan, L., Ellrott, K., Farr, C. L., Feuerhelm, J., Grant, J. C., Grzechnik, A., Han, G. W., Jaroszewski, L., Jin, K. K., Klock, H. E., Knuth, M. W., Kozbial, P., Krishna, S. S., Kumar, A., Lam, W. W., Marciano, D., Miller, M. D., Morse, A. T., Nigoghossian, E., Nopakun, A., Okach, L., Puckett, C., Reyes, R., Tien, H. J., Trame, C. B., Van Den Bedem, H., Weekes, D., Wooten, T., Yeh, A., Hodgson, K. O., Wooley, J., Elsliger, M. A., Deacon, A. M., Godzik, A., Lesley, S. A. and Wilson, I. A. (2010). *Acta Crystallographica*, **F66**(10), 1354–1364.

Xu, Q., Rawlings, N. D., Chiu, H.-J., Jaroszewski, L., Klock, H. E., Knuth, M. W., Miller, M. D., Elsliger, M.-A., Deacon, A. M., Godzik, A., Lesley, S. A. and Wilson, I. A. (2011). *PLoS ONE*, **6**(7), e22013.

Xu, Q., Sudek, S., McMullan, D., Miller, M. D., Geierstanger, B., Jones, D. H., Krishna, S. S., Spraggon, G., Bursalay, B., Abdubek, P., Acosta, C., Ambing, E., Astakhova, T., Axelrod, H. L., Carlton, D., Caruthers, J., Chiu, H. J., Clayton, T., Deller, M. C., Duan, L., Elias, Y., Elsliger, M. A., Feuerhelm, J., Grzechnik, S. K., Hale, J., Won Han, G., Haugen, J., Jaroszewski, L., Jin, K. K., Klock, H. E., Knuth, M. W., Kozbial, P., Kumar, A., Marciano, D., Morse, A. T., Nigoghossian, E., Okach, L., Oommachen, S., Paulsen, J., Reyes, R., Rife, C. L., Trout, C. V., van den Bedem, H., Weekes, D., White, A., Wolf, G., Zubieta, C., Hodgson, K. O., Wooley, J., Deacon, A. M., Godzik, A., Lesley, S. A. and Wilson, I. A. (2009). *Structure*, **17**(2), 303–313.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. (2015). *Nature Methods*, **12**(1), 7–8.

Yang, J. and Zhang, Y. (2015). *Nucleic Acids Research*, **43**(W1), W174–W181.

Yasuda, Y., Kageyama, T., Akamine, A., Shibata, M., Kominami, E., Uchiyama, Y. and Yamamoto, K. (1999). *Journal of Biochemistry*, **125**(6), 1137–1143.

Zeman, W. and Donahue, S. (1963). *Acta Neuropathologica*, **3**(2), 144–149.

Zeman, W. and Dyken, P. (1969). *Pediatrics*, **44**(4), 570–83.

Zwart, P. H. (2005). *Acta Crystallographica Section D: Biological Crystallography*, **61**(11), 1437–1448.

Acknowledgment

First I would like to thank my supervisor George M. Sheldrick for his support and the introduction into an interesting field of research I explored under his kind guidance.

I would like to thank Prof. Dr. Kai Tittmann and Prof. Dr. Dr. med Robert Steinfeld for their time and instruction as my thesis committee.

I would like to thank Prof. Dr. Inke Siewert, Jun.-Prof. Dr. Nathalie Kunkel, and Prof. Dr. Dietmar Stalke for being part of my examination commission.

I would like to thank Prof. Dr. Dr. med Robert Steinfeld for the interesting oppertunities arising from the cooperation on the topic of cln5. As part of this cooperation I was welcomed into the laboratory of the Department of Paediatrics and Paediatric Neurology and would like to thank Dr. Ralf Krätzner, Karin Schreiber, Annika Wolf and Marc Ziegenbein for the pleasure of working with them.

I would like to thank Stefan Becker for his guidance and patience when introducing me to protein crystallization. Furthermore, I am thankful for many helpful discussions and the oppertunity to accompany him to synchrotron facilities.

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

I would like to thank Stefan Becker, Christine Hansen and Jens Lübben for proof-reading 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 four years as enjoyable as they were. This includes Claudia Wandtke and Jens Lübben with whom I shared an office, Annika Münch, Helena Keil, Christian Köhler, and Paul Niklas Ruth, (who where so kind to share their office with me), and Christian Schürmann from the Stalke group, Massimo Sammito, Claudia Milan, Rafael Borges from the Usón group in Barcelona, Tim Grüne from the PSI in Switzerland, and all the other fellow researchers I had the pleasure of meeting.

I would like to thank everyone who was kind enough to test my program and provide feedback.

Finally I would like to thank my husband Jens for his continuing support.