Linking data and model quality in X-ray Crystallography

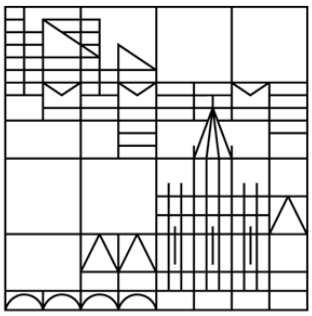
Master Thesis

submitted by

Jonathan Adler

at the

Universität Konstanz



Molecular Bioinformatics

Department of Biology

1st evaluator: Prof. Dr. Kay Diederichs

2nd evaluator: Prof. Dr. Olga Mayans

Konstanz, 2021

Table of Contents

Abstract

In this project, the usefulness of data simulation is demonstrated by exploring the impacts of four parameters (wavelength dispersion, beam divergence, cell axis variation, and orientation variation) on the quality of diffraction images and the structure models obtained from refinement against them. Several data quality indicators are plotted as functions of increasing parameter values, and examples of simulated diffraction images are shown, as well as electron density maps of refined models. Ideal reference models are simulated from insulin and lysozyme crystal data. Results show that the relative impact of each parameter differs between proteins, and possible explanations like the influence of unit cell size are discussed. Additionally, it is shown that model geometry is mostly unaffected by "realistic" values of the simulated parameters, while B-factors respond more strongly, in some cases to an unwanted extent. The meanings, reliability, and possible interpretations of internal data quality statistics are evaluated.

Zusammenfassung

Dieses Projekt demonstriert den Nutzen von Datensimulation durch Untersuchung der Effekte von vier Parametern (Wellenlängendispersion, Strahlendivergenz, Zellachsenvariation, und Variation der Orientierung von Mosaik-Kristallen) auf die Qualität von Diffraktionsbildern und den dagegen verfeinerten Strukturmodellen. Der Zusammenhang zwischen steigenden Parameterwerten und verschiedenen Datenqualitätsindikatoren wird veranschaulicht, sowie Beispiele für den Einfluss der Parameter auf Diffraktionsmuster und verfeinerte Elektronendichten. Daten von Insulin- und Lysozymkristallen werden zur Simulation der idealen Referenzmodelle verwendet. Ergebnisse zeigen, dass sich die relative Auswirkung der einzelnen Parameter abhängig von den Eigenschaften des verwendeten Proteins unterscheidet. Mögliche Erklärungen, wie der Einfluss der Einheitszellengröße, werden diskutiert. Außerdem wird gezeigt, dass die Geometrie von Strukturmodellen bei "realistischen" Parameterwerten großteils unbeeinflusst bleibt, während B-Faktoren stärker reagieren, teils in unerwünschtem Ausmaß. Die Bedeutungen, Verlässlichkeit, und möglichen Interpretationen interner Datenqualitätsindikatoren werden evaluiert.

1. Introduction

Macromolecular X-ray crystallography is the most commonly used method in structural biology, leading to the deposition of thousands of new structures to the Protein Data Bank every year. The insights gained from many of these structures are of great help to anyone studying the properties, functions, and interactions of biomolecules. However, the reliability of such insights is limited by the quality of the deposited models, which only represent an approximation of the physical reality. This is due to the complex and flexible nature of most biomolecules, as well as difficulties related to the involved mathematical methods. The greatest of these difficulties is known as the "phase problem", which is caused by the loss of phase information during measurement of X-ray intensities. Solving the phase problem requires highly accurate and precise diffraction data, which necessitates optimization of experimental setups. Possible sources of experimental error include the properties of the crystal (size, disorder, mosaicity, twinning, radiation damage), beam (instability, divergence, dispersion), detector (distance, sensitivity, gain, pixel size, point spread, shutter jitter), goniometer, and cooling system (vibrations, absorption, shading). All these parameters affect the pixel intensities of reflection profiles recorded in diffraction images, which are translated to structure factors by data reduction programs like **XDS** [Kabsch, 2010]. Determining the impacts of specific experimental parameters on reflection shapes, as well as how these impacts are interpreted by data reduction programs, and how the downstream consequences of these interpretations affect model refinement, should be of interest to crystallographers. Synthetic data, simulated with programs like **SIM_MX** [Diederichs, 2009], can be used to answer such questions. For this project, the effects of four simulated parameters on the operation of data reduction and analysis programs, and ultimately the refined models, are studied using several data quality indicators.

1.1. R-factor

The term "R-factor" refers to various indicators used in crystallographic data analysis. Most commonly, it describes a value measuring the agreement between structure factor amplitudes calculated from a structure model (F_{model}) and those from the original X-ray diffraction data (F_{obs}) [Rupp, 2009]. This measure is defined as:

$$R = \frac{\sum |F_{obs} - kF_{model}|}{\sum |F_{obs}|}$$

With k being a scaling factor. Instead of F_{obs}, structure factor amplitudes obtained from simulated diffraction images will be designated F_{sim} in the following. Theoretical values for R lie between 0 (perfect agreement) and 0.6 (no agreement) [IUCr Commission on Crystallographic Nomenclature, 2017]. To avoid over-fitting data to a model, a subset of the experimentally determined intensities (usually 5-10%) is removed from the rest and not used for model refinement. An unbiased control indicator R_{free} [Brünger, 1992] is then calculated from the corresponding amplitudes and compared to R_{work}, the value derived from the "working" set of intensities.

1.2. B-factor

Also called "atomic displacement parameter" (ADP), the B-factor is proportional to the mean square isotropic displacement $\langle u_{iso}^2 \rangle$ of an atom from its equilibrium or mean position:

$$B_{iso} = 8\pi^2 \langle u_{iso}^2 \rangle$$

Vibrational movement, stereochemical flexibility, and crystal lattice disorder all contribute to isotropic displacement [Rupp, 2009]. Since these properties can vary between different parts of a molecule, B-factors of individual atoms depend on their position in the unit cell. For example, atomic B-factors in tightly packed core regions of a protein will usually be much lower than those in surface-exposed side chains.

1.3. $(I/\sigma)^{asymptotic}$

$(I/\sigma)^{asymptotic}$, which will be referred to as ISa from here on, estimates the highest signal-to-noise ratio [I/σ(I)] that a given experimental setup (beam, crystal, spindle, detector, cooling, software,...) can produce, and may therefore be used to determine whether the multiplicity of data measured with said setup is sufficient for a given level of (merged) data precision, or whether it should be increased. It's value is proportional to the normalized difference between estimates for the overall variance of reflection intensities, and the variance due to counting statistics and detector properties [Diederichs, 2010].

1.4. $CC_{1/2}$

The $CC_{1/2}$ value estimates the level of signal available in a dataset. It can be computed by splitting unmerged data into random halves A and B, merging the intensities contained in each subset for unique reflections a_i and b_i, and calculating Pearson's correlation coefficient for said reflections [Pearson, 1895]:

$$CC_{1/2} = \frac{\sum (a_i - \bar{a})(b_i - \bar{b})}{[\sum (a_i - \bar{a})^2 \sum (b_i - \bar{b})^2]^{1/2}}$$

This indicator can be used to determine a reasonable resolution cut-off by estimating the signal content in high resolution shells [Karplus & Diederichs, 2012].

1.5. Wilson plot

In a Wilson plot, the natural logarithms of intensity means over narrow resolution ranges are mapped against the squared inverse mean resolution. The result for "good" data should be a relatively straight line at high resolution (< 3 Ångström). The gradient of this line can be used to estimate the overall isotropic B-factor for a protein, while the intercept provides a scale factor for bringing experimental intensity data onto an absolute scale [Cowtan, 2008; Rupp, 2009].

1.6. Rama-Z score

The torsion angles ψ and φ describe rotation around the atomic bonds connecting an amino acid's alpha carbon atom to the rest of the protein backbone. In a Ramachandran plot, these angles are mapped against each other to visualize energetically allowed regions for amino acid residues in a protein structure [Ramachandran et al., 1963]. The overall shape of the torsion angle distribution for a given structure can be summarized in a "Rama-Z" value. This indicator describes how "normal" a model is, compared to a reference set of high-resolution structures. The average score of the reference data set is 0, standard deviation is 1. Indicator values between −2 and 2 are expected for most structures, and anything beyond 3 indicates fairly improbable structure geometry [Sobolev et al., 2020].

1.7. Clashscore

As a result of an all-atom contact analysis performed by the **PROBE** program [Word et al., 1999], van der Waals surface overlap between atoms is measured. When two non-bonded atoms overlap by more than 0.4 Å, this is denoted as a serious clash. The value of the "Clashscore" indicator reported by **MolProbity** is the average number of serious clashes per 1000 atoms. [Chen et al., 2010]

1.8. MolProbity score

This indicator is "a log-weighted combination of the Clashscore, percentage Ramachandran not favoured and percentage bad side-chain rotamers, giving one number that reflects the crystallographic resolution at which those values would be expected. Therefore, a structure with a numerically lower MolProbity score than its actual crystallographic resolution is, quality-wise, better than the average structure at that resolution." [Chen et al., 2010]

2. Methods

To test the impact of different simulated parameters on various data quality indicators, several programs were used in a sequence that was automated using scripts (see appendix A) written in **GNU bash** [GNU, 2007] and **Python3.6** [Van Rossum et al., 1995], including the modules **NumPy** (version 1.19.1) [Harris et al., 2020], **Matplotlib** (version 3.3.0) [Hunter et al., 2007], **math**, **sys**, and **os** [Van Rossum, 2020].

2.1. Input models

2bn3 (insulin) [Nanao et al., 2005] and 1dpx (lysozyme) [Weiss et al., 2000] were chosen as starting models due to their small size, which allows for fast calculations. From each of these starting models, two new models were created. For both, alternative side chain conformations and anisotropic B-factors were removed. The first one was used for the calculation of structure factors to a resolution of 1.8 Å. To simulate crystal disorder, achieve more realistic R-factors, and thereby ensure that observed effects wouldn't be obscured by noise in real experiments, 20 copies of this model had their atomic positions "shaken" (see **create_ensemble.sh**, download link in appendix A) and were then combined into an ensemble model, as suggested in a recent publication on data simulation [Holton, 2019]. For

the second model, all isotropic B-factors were set to 20. This model was then used for refinement against the simulated structure factor data.

2.2. FULLSPHERE_FROM_P1

Performs symmetry expansion from a half sphere of P1 data (possibly with intensities of both members of Friedel pairs) to the full sphere. Expects an input file *fort.1* containing structure factor amplitudes and writes intensities to an output file *intensities.hkl*. Necessary for reformatting data from a structure factor calculation program for use in **SIM_MX**.

2.3. SIM_MX

Uses ray-tracing to simulate a diffraction experiment with multiple individual beams and the corresponding mosaic blocks (in the context of this program, each "mosaic block" is an infinite, perfect crystal which only scatters photons belonging to its corresponding beam) [Diederichs, 2009]. Expected inputs are a free-format file *intensities.hkl* containing (h, k, l) coordinates and intensity values, and a control file *SIMULATE.INP* containing different parameters describing the simulated experiment. All parameters can be found in table 1 of the method's original publication. The parameters relevant for this project are WAVELENGTH_STDDEV, BEAM_STDDEV, CELL_STDDEV, and ORIENTATION _STDDEV, which were incrementally increased to determine their specific impacts on data quality indicators and the resulting model. The parameter values are simulated up to a value at which $CC_{1/2}$ lies between zero and 40 (see figure 34).

2.3.1. WAVELENGTH_STDDEV

Standard deviation of the wavelength of simulated beams. Will be called "wavelength dispersion" from here on. Plotted values range from zero to 0.1726 Ångström.

2.3.2. BEAM_STDDEV

Standard deviation of the direction of individual beams along two axes perpendicular to the direction of the "primary" beam, which goes towards the centre of the detector. Will be called "beam divergence" from here on. Plotted values range from zero to 0.3959 degrees.

2.3.3. CELL_STDDEV

Standard deviation of all three unit cell axis lengths of simulated mosaic blocks. Will be called “cell axis variation” from here on. Plotted values range from zero to 3.3246 Ångström.

2.3.4. ORIENTATION_STDDEV

Standard deviation of the rotation of mosaic blocks around three orthogonal axes. Will be called “orientation variation” from here on. Plotted values range from zero to 3.6 degrees.

2.4. XDS

X-ray **D**etector **S**oftware [Kabsch, 2010] for processing single-crystal monochromatic diffraction data recorded by the rotation method or, in this case, simulated by **SIM_MX**. Expected inputs are images generated by diffraction experiment (or simulation), and a control file *XDS.INP* which for this project was linked to *SIMULATE.INP,* since **XDS** shares many control parameters with **SIM_MX**. MAXIMUM_ERROR_OF_SPOT_POSITION and MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT had to be adjusted to avoid early indexing failure due to reflections being broken up or not being recognized otherwise.

2.5. CCP4

The Collaborative Computational Project Number 4 supports an integrated suite of programs for experimental determination and analysis of protein structures. Version used: 7.0 [Brünger, 1992 & 1997; Krissinel et al., 2004; Winn et al., 2011]

Table 1 CCP4 programs in order of the scripted sequence in which they were used for this project. The reference page for each individual program is “http://legacy.ccp4.ac.uk/html/[Name].html”

Name	Function
mtzdump	Dump data from an MTZ reflection data file.
f2mtz	Convert a formatted reflection file to MTZ format.
cad	Collect and sort crystallographic reflection data from several files, to generate a single set.
unique	Generate a unique list of reflections.
freerflag	Tags each reflection in an MTZ file with a flag for cross-validation.
sftools	Reflection data file utility program including some density map handling.
wilson	Wilson plot, absolute scale and temperature factor.
mtzutils	Reflection data files utility program.
superpose	Structural alignment based on secondary structure matching.

2.6. PHENIX

PHENIX is a software suite for the automated determination of molecular structures using a variety of methods. Version used: dev-3965 [Afonine et al., 2012] [Liebschner et al., 2019] [Williams et al., 2018]

Table 2 PHENIX programs in order of the sequence in which they were used for this project. The reference page for each individual program is "https://www.phenix-online.org/documentation/reference/[Reference name].html"

Name	Function	Reference name
phenix.pdbtools	PDB model manipulations and statistics.	pdbtools
phenix.fmodel	Compute structure factors from PDB model file.	fmodel
phenix.reflection_file_converter	Conversion of many reflection file formats to MTZ, CNS, SCALEPACK or SHELX format.	reflection_file_tools
phenix.refine	General purpose crystallographic structure refinement program.	refinement
phenix.molprobity	PDB model geometry validation.	molprobity_tool

3. Results

All values plotted in the following figures were simulated for 2bn3 (insulin). Plots for 1dpx (lysozyme) show reasonably similar trends and are provided in appendix A.2.

3.1. Effects of different parameters on reflection spot shape

The following images were created using **XDS-Viewer** [Hoffer, 2013], **GIMP** (GNU Image Manipulation Program) [GIMP Development Team, 2019], and **ImageMagick** [ImageMagick Development Team, 2021]. They show diffraction frames simulated at different parameter values, to illustrate the respective effects on reflection spot shapes. To show the full symmetry of the crystal while keeping reflection size at an interpretable scale, figures 1, 2, 4, 6, 8, and 10 only show resolutions up to ~3.0 Ångström (image corners). Figures 3, 5, 7, and 9 then serve to show the effect of each parameter over the full resolution range (upper right corner: 1.8 Ångström)

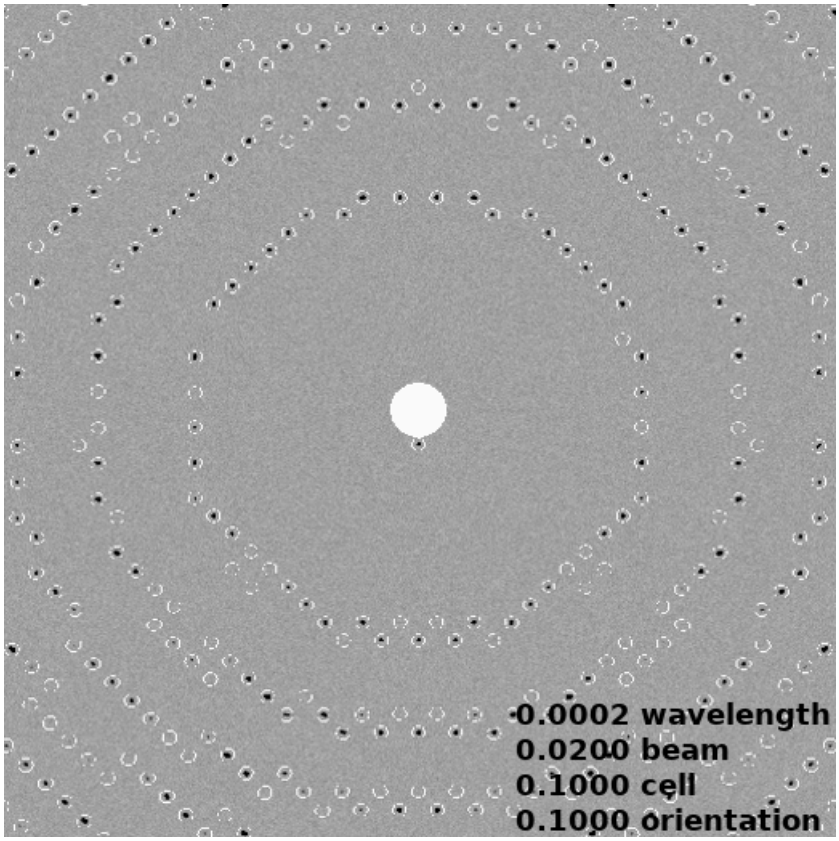


Figure 1 Reflection shapes on detector frames simulated with the (very low) default values of **SIM_MX** for wavelength dispersion, beam divergence, cell axis variation, and orientation variation. Reflection spots (black shapes) are very small and sharp, all pixels are entirely contained in the areas (white outlines) predicted by **XDS** for individual reflection positions. The white area in the centre is caused by the beamstop.

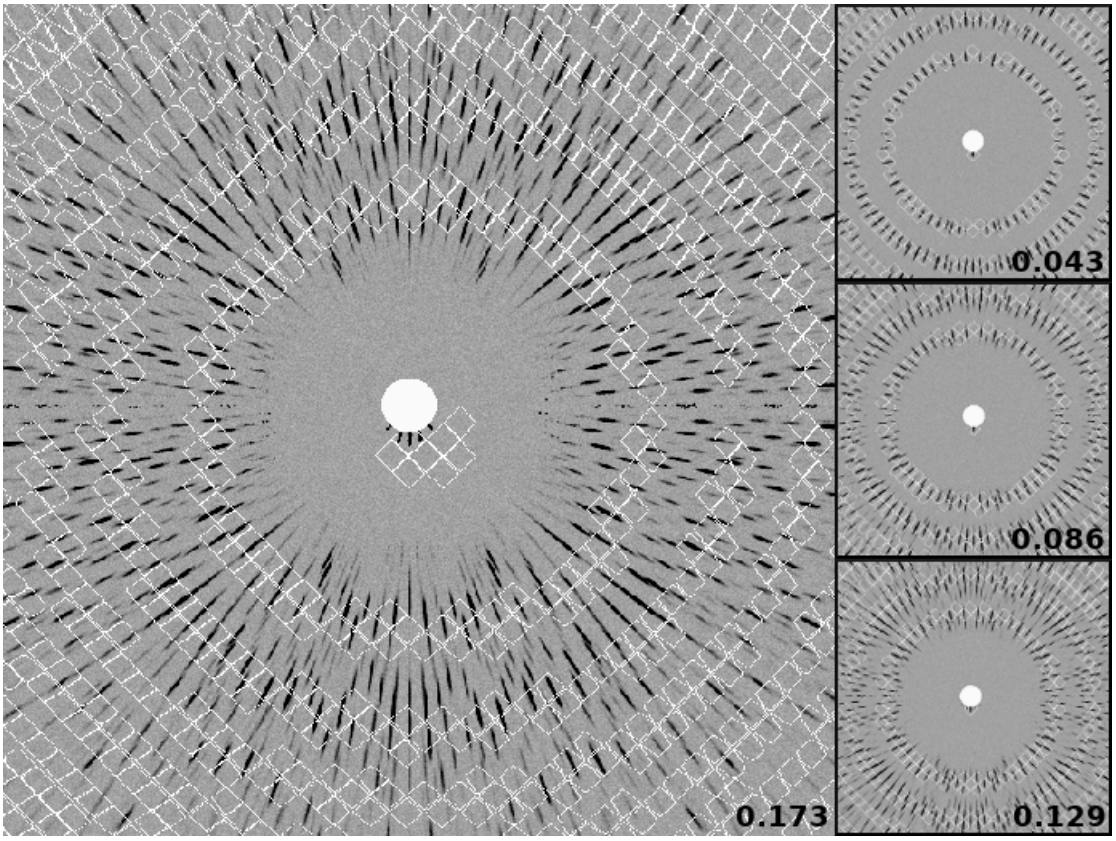


Figure 2 Reflection shapes on simulated detector frames at increasing wavelength dispersion values. Reflection spots are radially elongated and do not fit into the predicted areas at any resolution. Many reflections lie outside the predicted resolution shell areas.

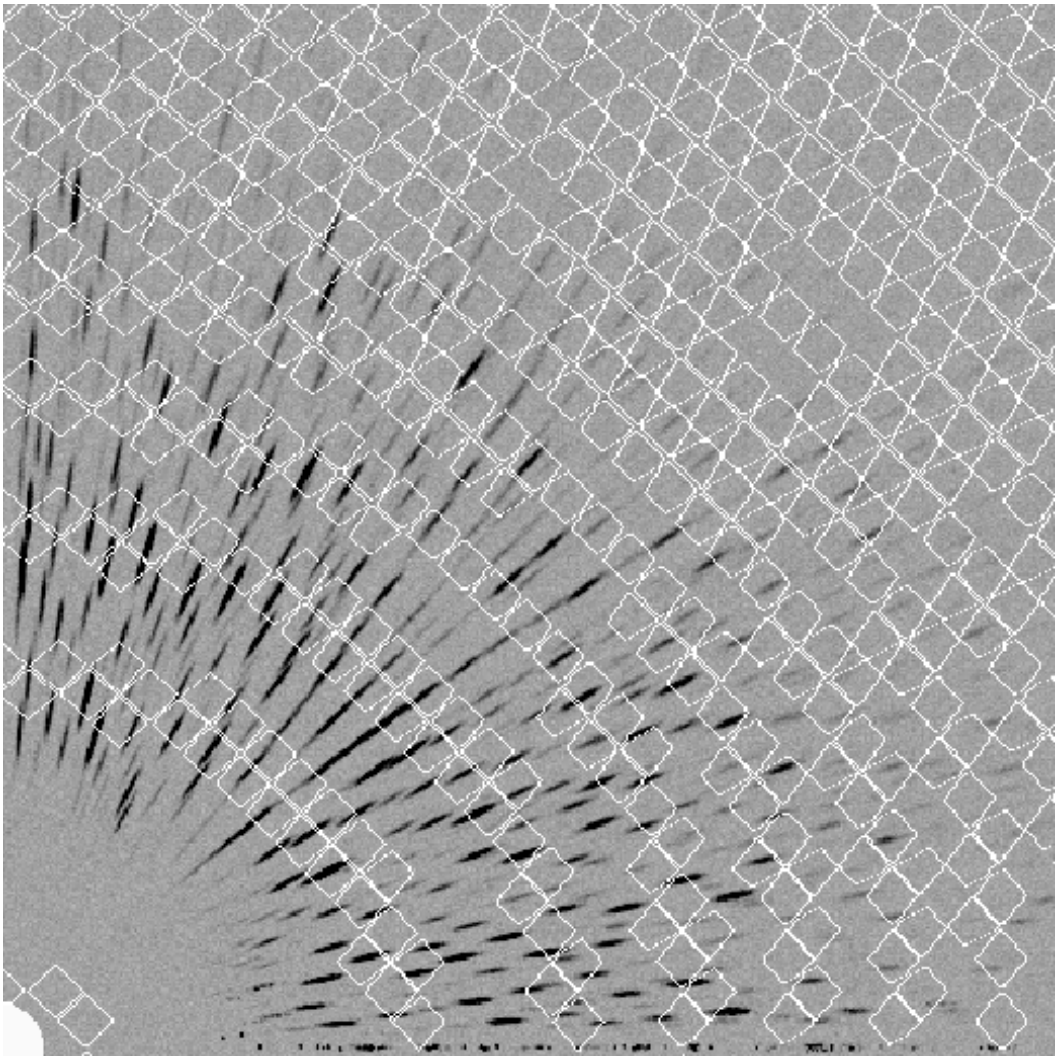


Figure 3 Reflection shapes at largest simulated wavelength dispersion value (0.1726 Å) up to the highest recorded resolution (1.8 Å, upper right image corner).

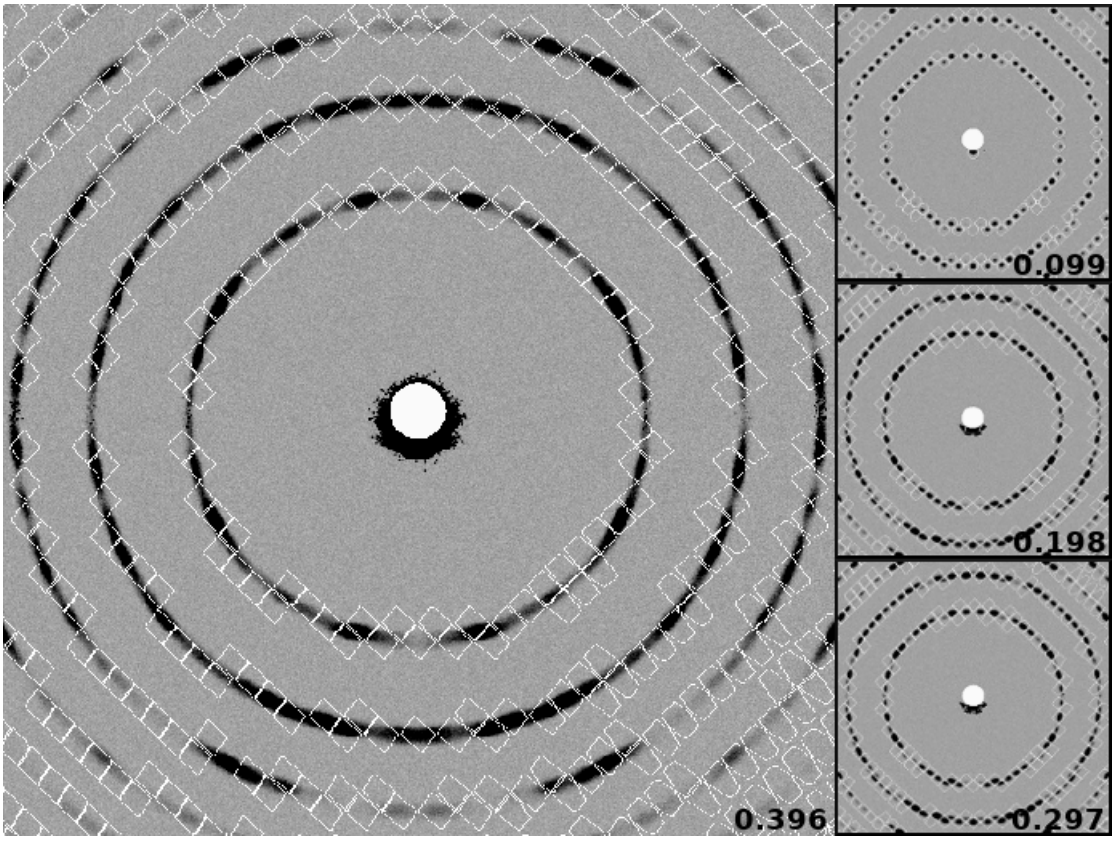


Figure 4 Reflection shapes at increasing simulated beam divergence values. Spots are broadened and difficult to separate, but most pixels are still contained within the predicted areas.

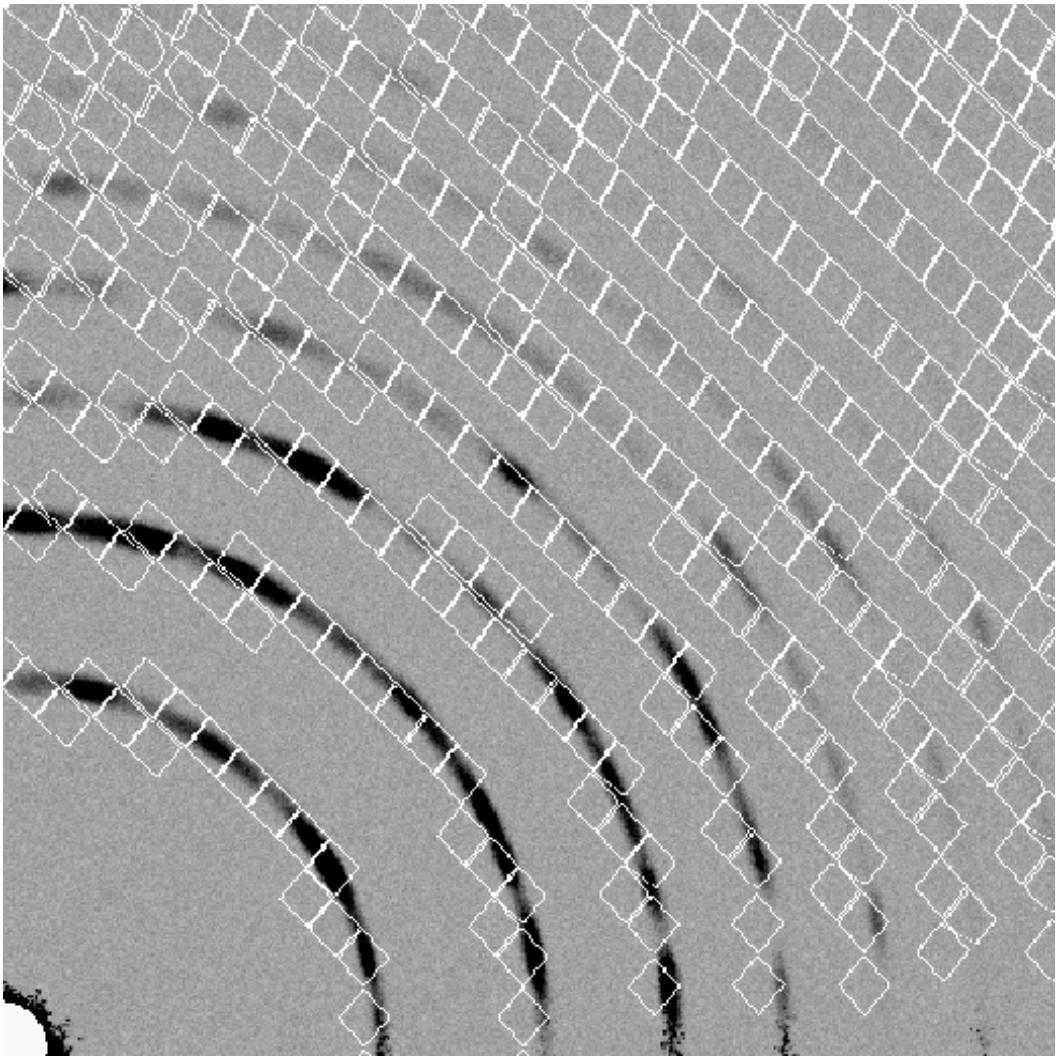


Figure 5 Reflection shapes at largest simulated beam divergence value (0.3959 Å) up to the highest recorded resolution (1.8 Å, upper right image corner).

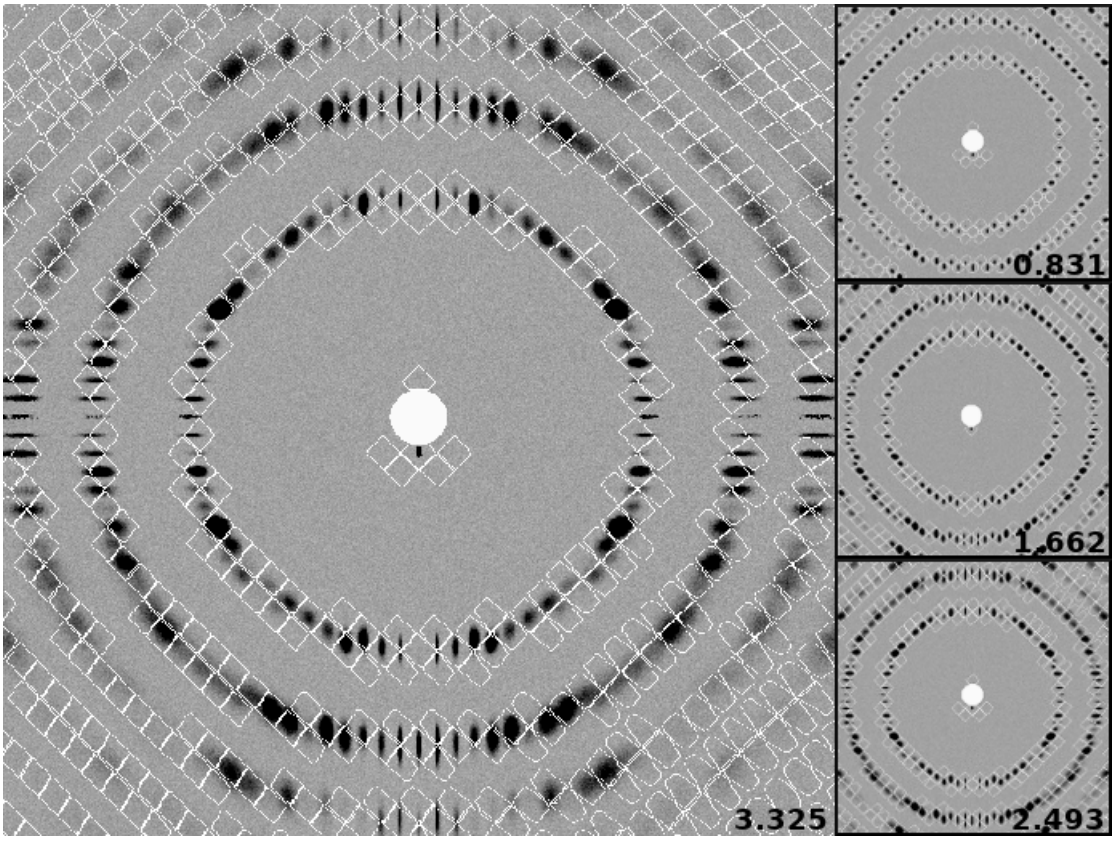


Figure 6 Reflection shapes at increasing simulated cell axis variation values. Distortion of spots increases with resolution. Spot shapes differ greatly between detector areas.

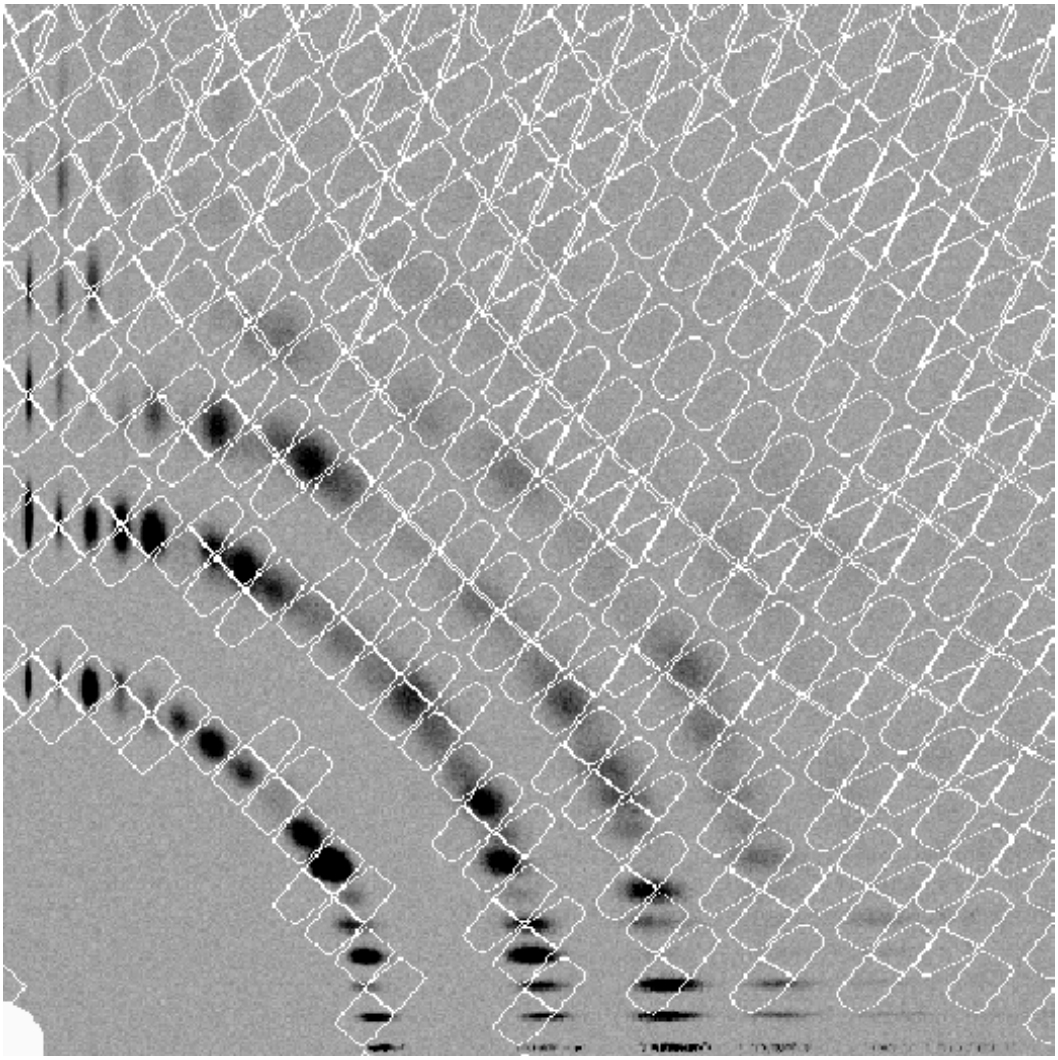


Figure 7 Reflection shapes at largest simulated cell axis variation value (3.3246 Å) up to the highest recorded resolution (1.8 Å, upper right image corner).

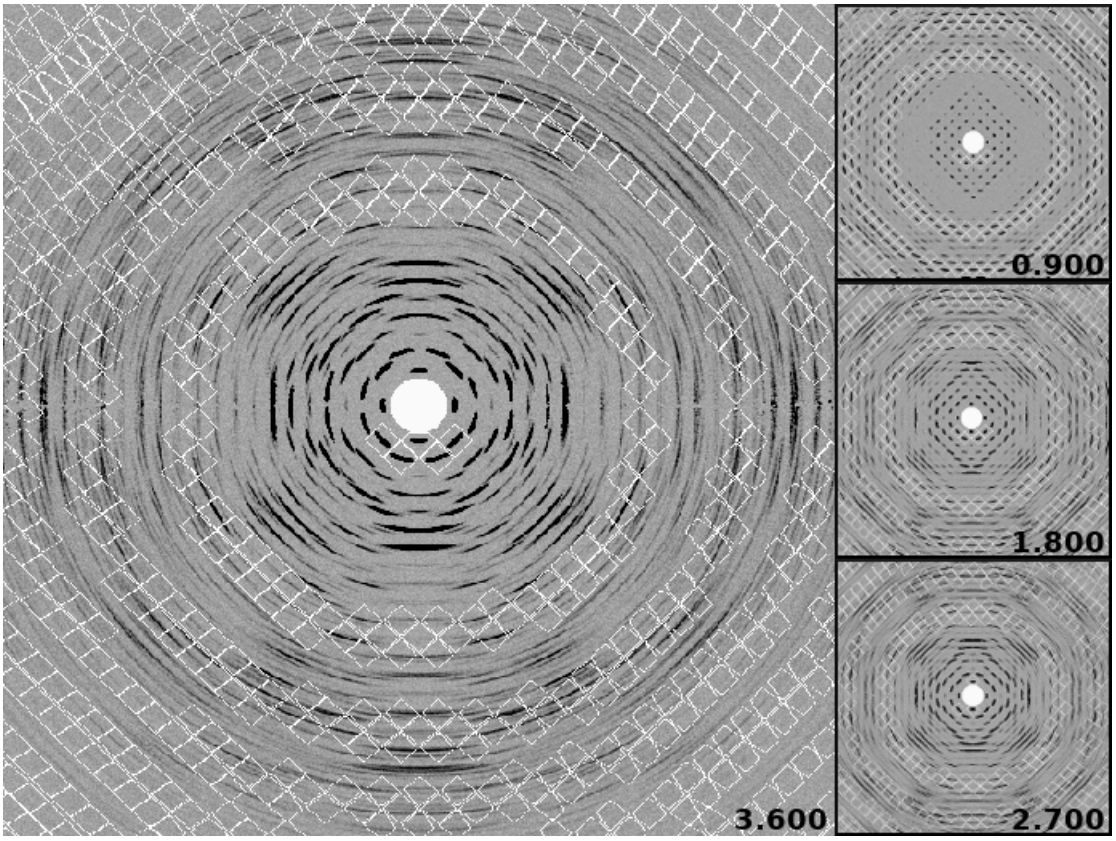


Figure 8 Reflection shapes at increasing simulated orientation variation values. Spots are elongated along circles around the beamstop and do not fit the predicted areas well at any resolution. Additional reflections can be seen outside the resolution shells, and overlap between reflections, particularly at higher resolutions, looks to be very strong compared to what is caused by other parameters.

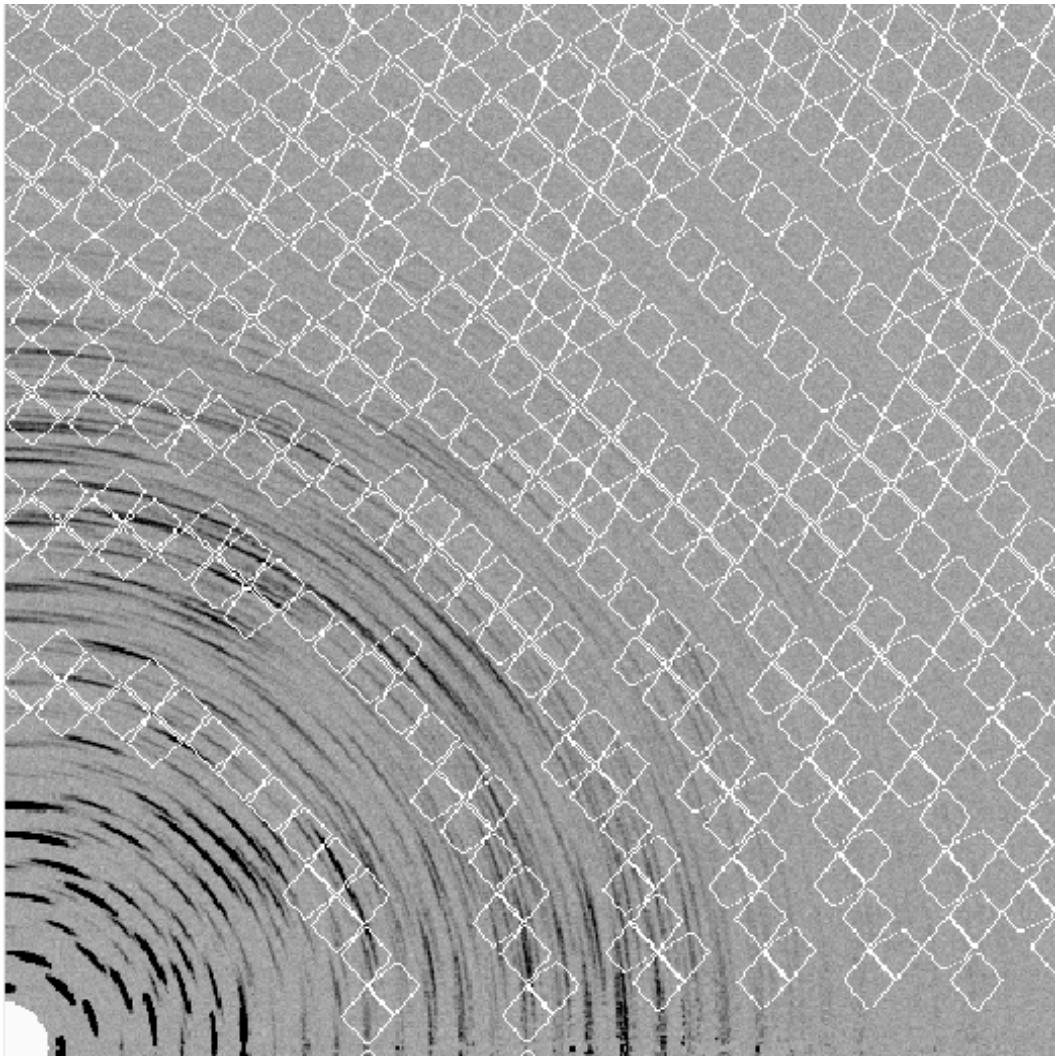


Figure 9 Reflection shapes at largest simulated orientation variation value (3.6 Å) up to the highest recorded resolution (1.8 Å, upper right image corner).

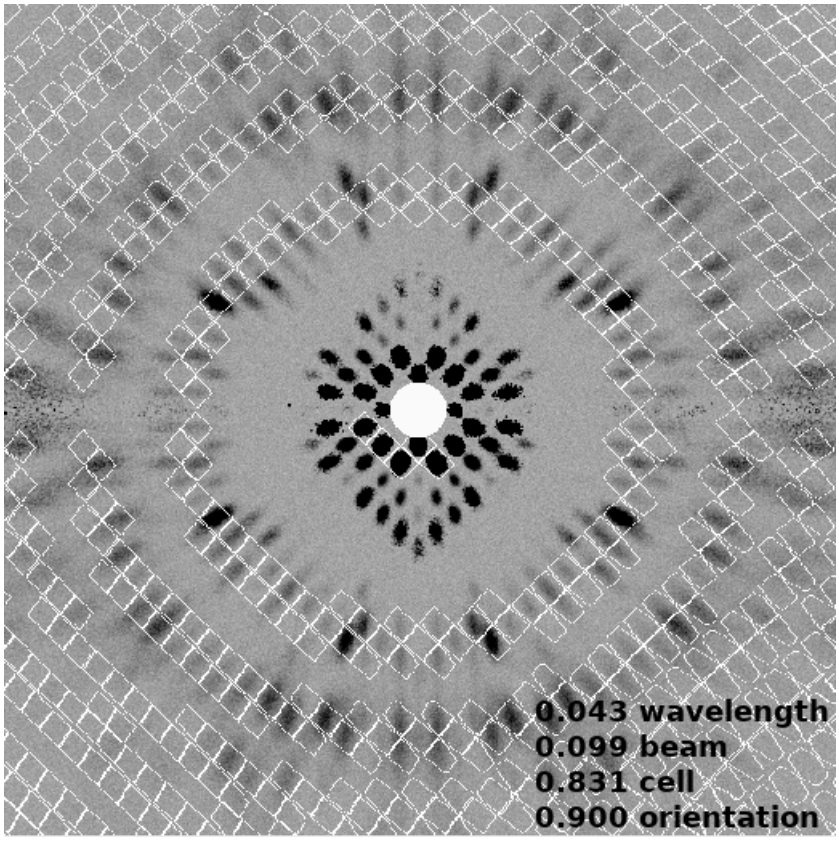


Figure 10 Reflection shapes with all simulated parameter values at one fourth of their respective maxima. Spot shapes show effects of all parameters (elongation, broadening, additional reflections outside the predicted resolution shells, disappearance at high resolution) as expected. No additional effects seem to result from the combination of parameters.

3.2. Number of overlapping / contributing pixels per reflection

For each simulation, **SIM_MX** reports the average number of contributing and overlapping pixels per reflection. All pixels that obtain intensity contributions from only a single reflection are counted as “contributing”, while all those receiving intensity contributions from more than one reflection are counted as “overlapping”. For this reason, increase of the latter number decreases the former, which can be observed in all figures belonging to this section. Reflection overlap also increases errors in intensity measurements, and is therefore expected to correlate with deterioration of data quality.

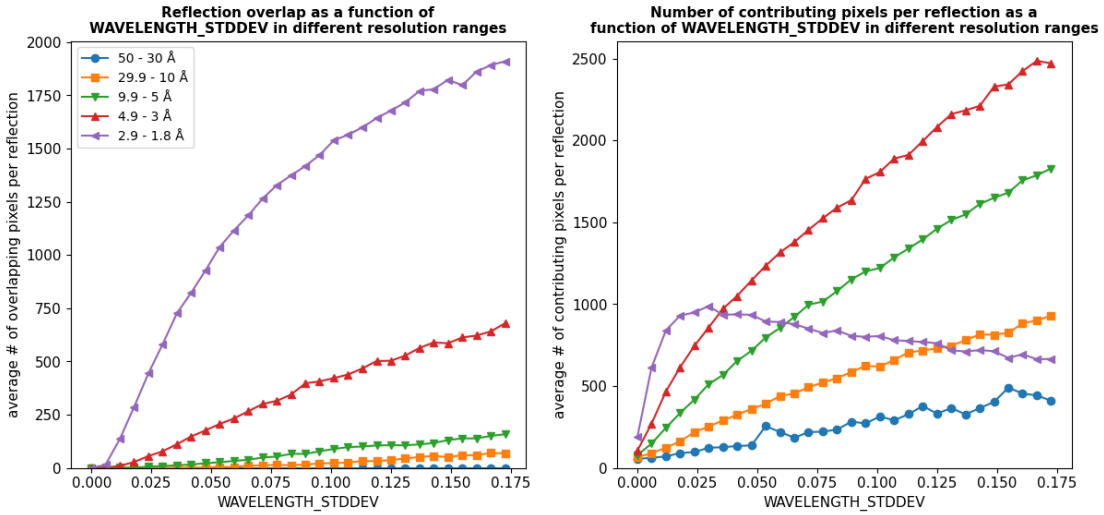


Figure 11 Average number of contributing and overlapping pixels per reflection with increasing values of wavelength dispersion (max = 0.1726 Å) in different resolution ranges.

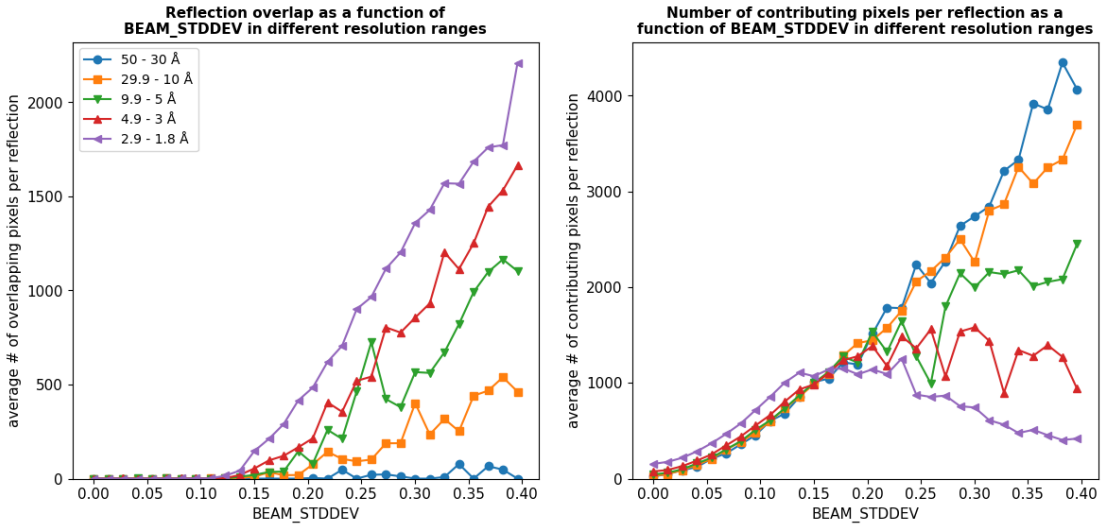


Figure 12 Average number of contributing and overlapping pixels per reflection with increasing values of beam divergence (max = 0.3959 Å).

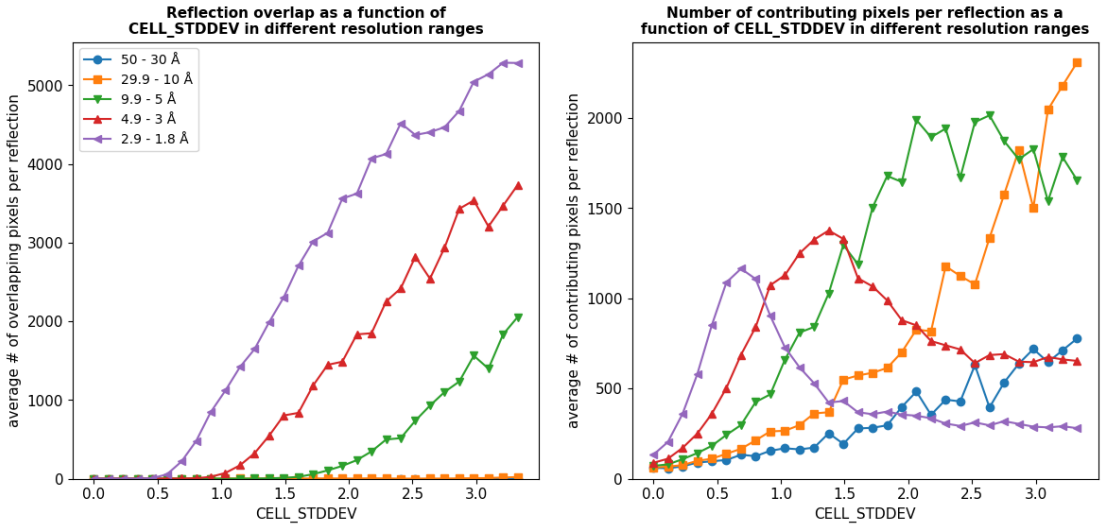


Figure 13 Average number of contributing and overlapping pixels per reflection with increasing values of cell axis variation (max = 3.3245 Å).

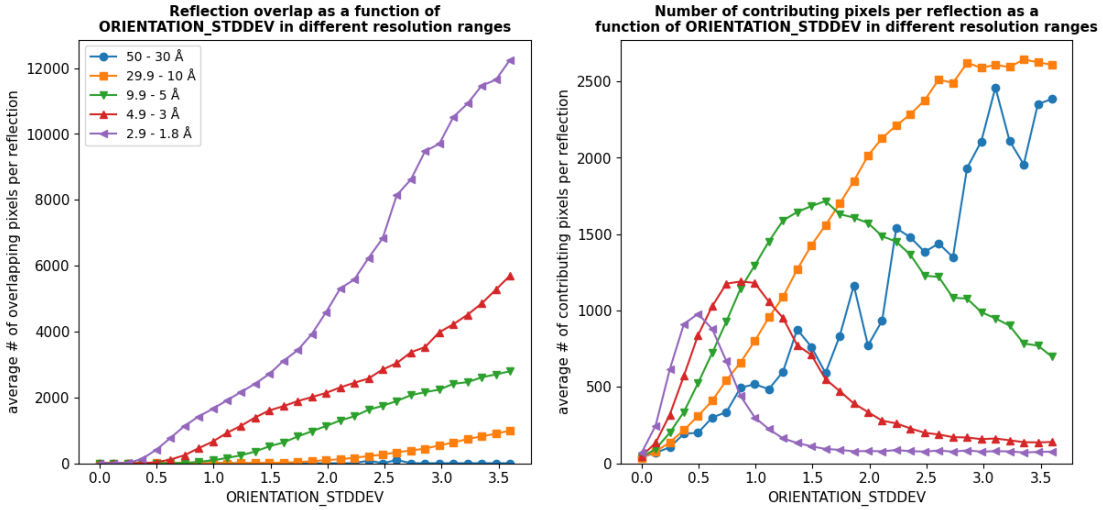


Figure 14 Average number of contributing and overlapping pixels per reflection with increasing values of orientation variation (max = 3.6 Å). The overlap at high resolution increases far more with this parameter compared to any of the others.

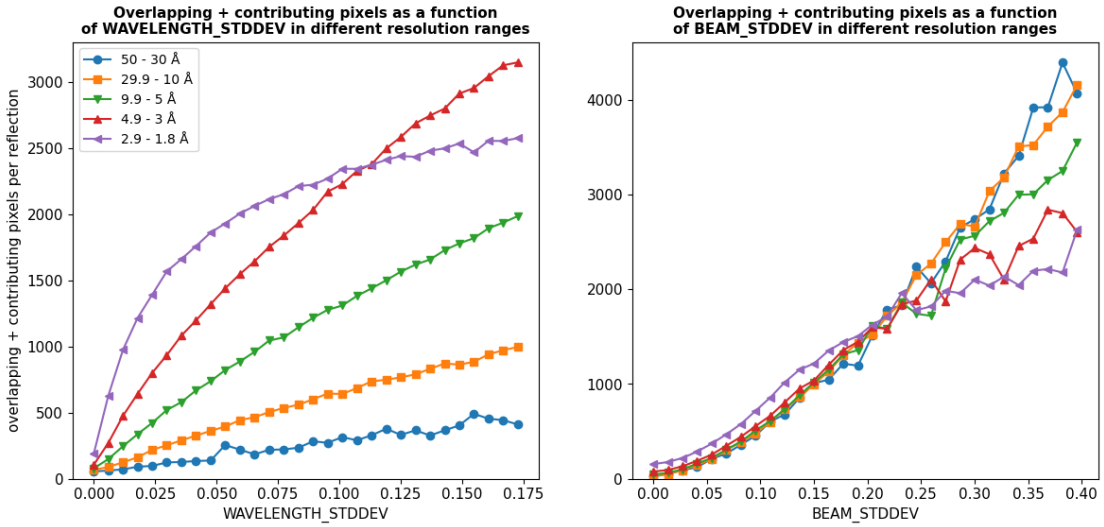


Figure 15 The sum of all overlapping and contributing pixels per reflection gives the overall spot size, which can be seen going up in all resolution ranges as parameter values increase. While the effect of wavelength dispersion sets in much faster than that of beam divergence, it also starts slowing down and seemingly reaching a limit much earlier at high resolution.

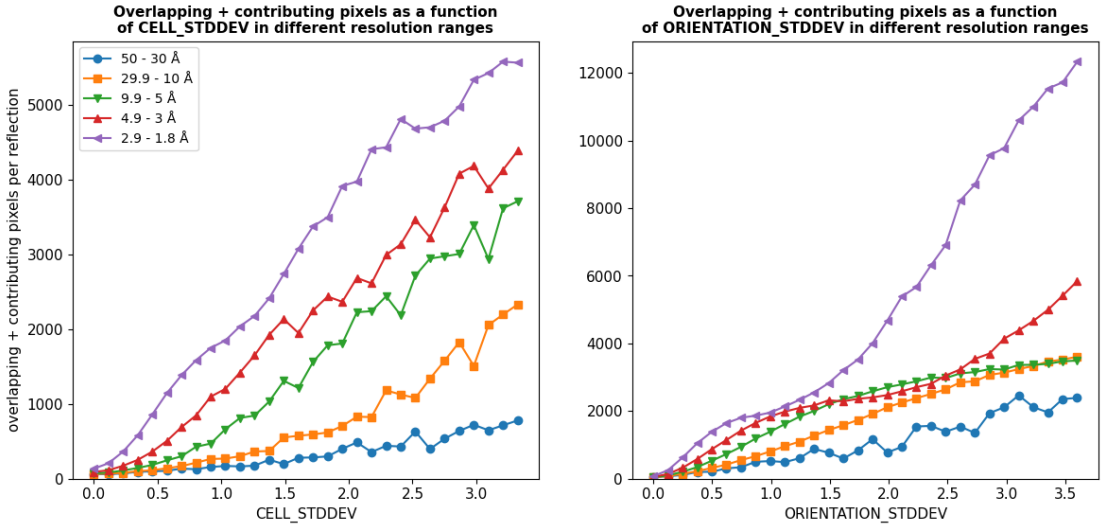


Figure 16 Pixel size for cell axis variation increases as expected. For orientation variation, the extreme overlap at high resolution also causes a wide gap in pixels sizes between the highest resolution shell and all others.

3.3. R-factors, overall B-factors, and ISa

The plots in this section illustrate the effects wavelength dispersion, beam divergence, cell axis variation, and orientation variation on R-factors, B-factors, and ISa.

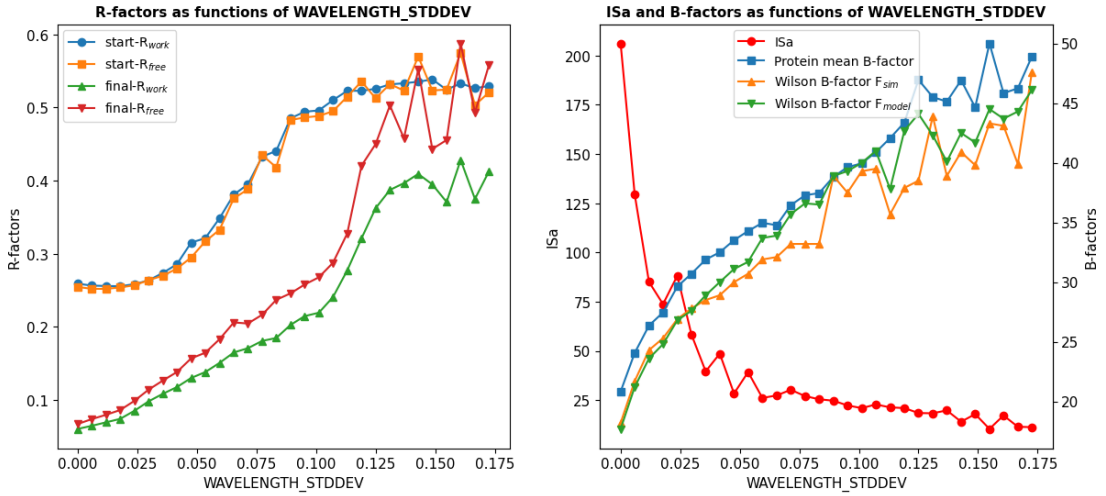


Figure 17 Left side: Change in R-values before (start-R_{work}, start-R_{free}) and after (final-R_{work}, final-R_{free}) refinement with increasing wavelength dispersion. Right side: Change in ISa, mean B-factor for protein atoms, and overall Wilson B-factors corresponding to simulated (F_{sim}) and modelled (F_{model}) structure factor amplitudes. All indicators are visibly affected by wavelength dispersion even at low values, with final-R_{free} reaching higher numbers than with any other parameter.

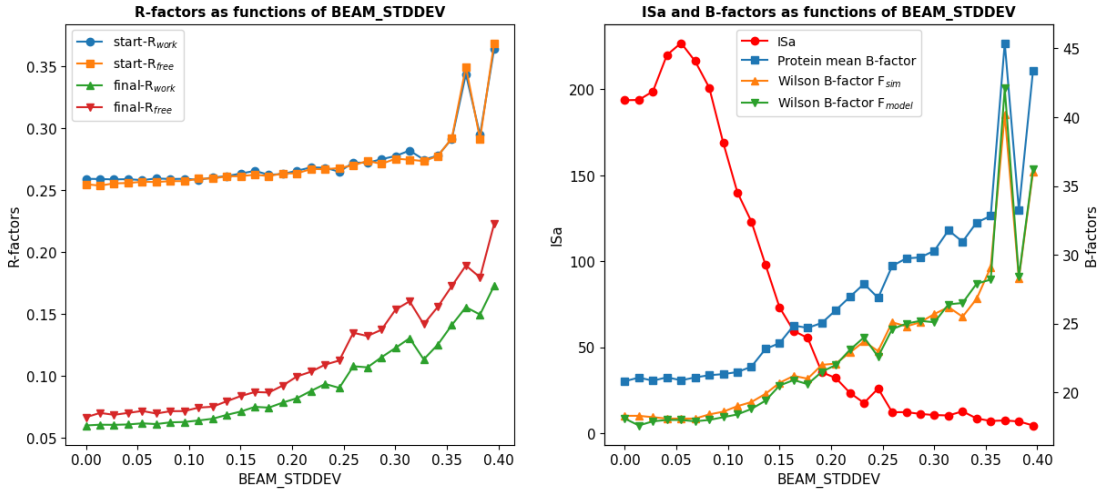


Figure 18 Left side: Change in R-values before (start-R_{work}, start-R_{free}) and after (final-R_{work}, final-R_{free}) refinement with increasing beam divergence. Right side: Change in ISa, mean B-factor for protein atoms, and overall Wilson B-factors corresponding to simulated (F_{sim}) and modelled (F_{model}) structure factor amplitudes. Beam divergence shows the weakest effect on all data quality indicators, and initially even causes an increase in ISa. At high parameter values, B-factors start fluctuating.

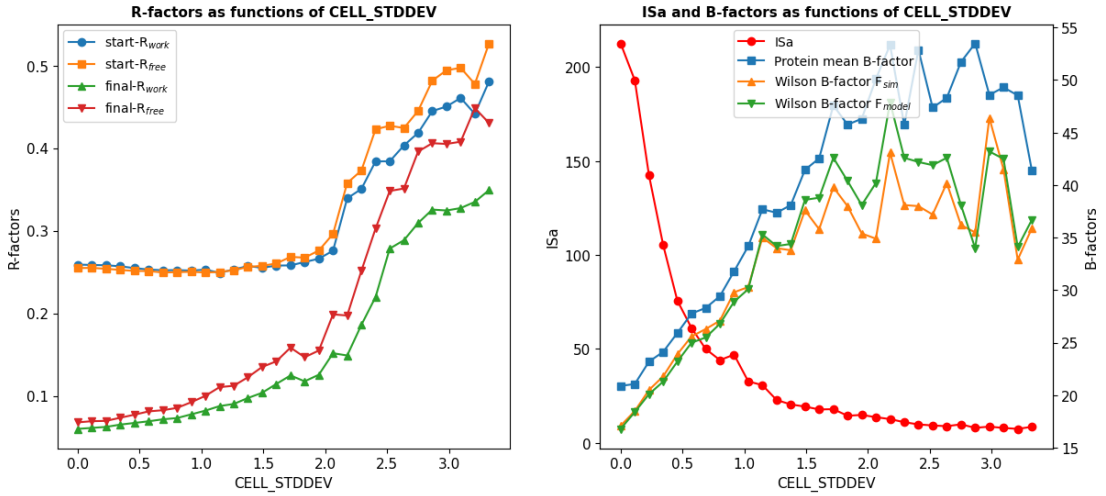


Figure 19 Left side: Change in R-values before (start-R_{work}, start-R_{free}) and after (final-R_{work}, final-R_{free}) refinement with increasing cell axis variation. Right side: Change in ISa, mean B-factor for protein atoms, and overall Wilson B-factors corresponding to simulated (F_{sim}) and modelled (F_{model}) structure factor amplitudes. Among the four parameters, cell axis variation causes the strongest absolute increase in B-factors, as well as the second strongest increase in R-values.

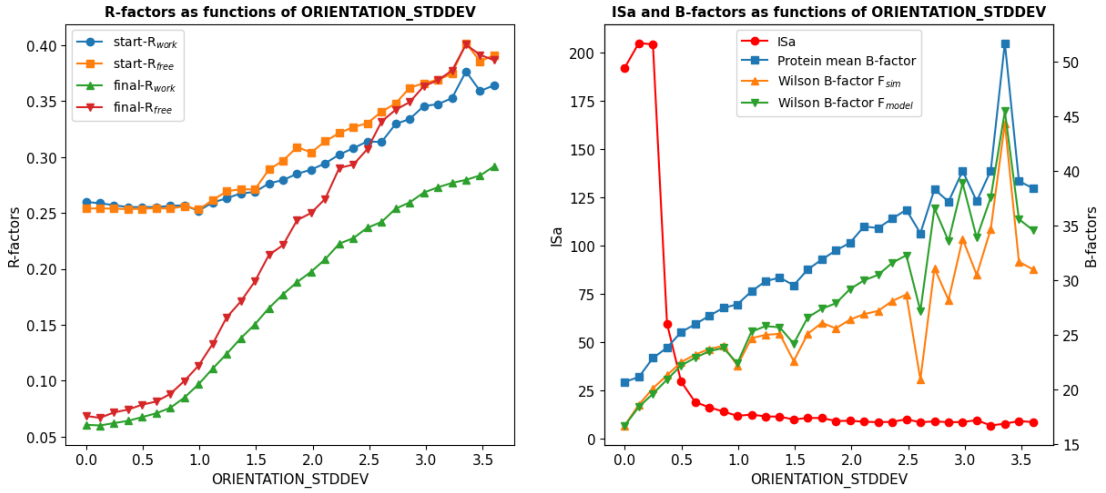


Figure 20 Left side: Change in R-values before (start-R_{work}, start-R_{free}) and after (final-R_{work}, final-R_{free}) refinement with increasing orientation variation. Right side: Change in ISa, mean B-factor for protein atoms, and overall Wilson B-factors corresponding to simulated (F_{sim}) and modelled (F_{model}) structure factor amplitudes. The highest R-values caused by this parameter are below those caused by wavelength dispersion and cell axis variation. Like with beam divergence, low values of orientation variation initially cause a small increase in ISa, although the effect is less noticeable.

3.4. Resolution-dependent R-factors

The plots in this section show the change of R_{work} and R_{free} in different resolution bins with increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation. Resolution range (Å) per bin: **(1)** 31.80 – 3.87, **(2)** 3.86 – 3.07, **(3)** 3.07 – 2.68, **(4)** 2.68 – 2.43, **(5)** 2.43 – 2.26, **(6)** 2.26 – 2.13, **(7)** 2.13 – 2.02, **(8)** 2.02 – 1.93, **(9)** 1.93 – 1.86, **(10)** 1.86 – 1.8; Trends of R-values beyond ~0.42 can be ignored, as such values could also be achieved with refinement against random data [Evans & Murshudov, 2013]. They are included nonetheless because they show which resolution bins still contain information at the corresponding parameter values. Interpretable R-values (< 0.42) increase with resolution, as one would expect. R_{free} fluctuates more than R_{work} due to the smaller number of reflections used to determine it, and the resulting statistical noise.

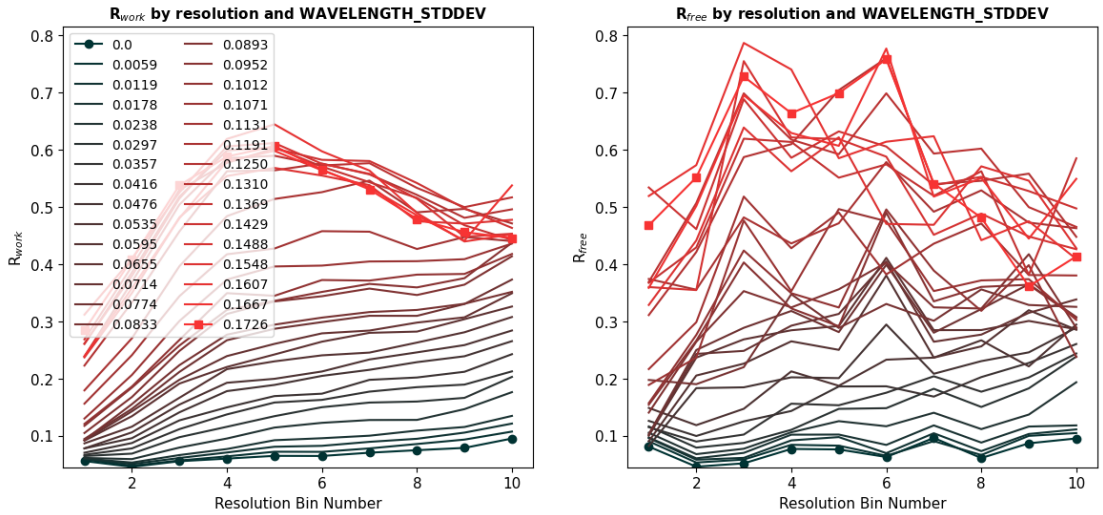


Figure 21 Change in R-values at different wavelength dispersion values (see legend) with increasing resolution. At resolutions beyond ~3.0 Å, data becomes unusable with large wavelength dispersion values, since R-values rise above 0.42.

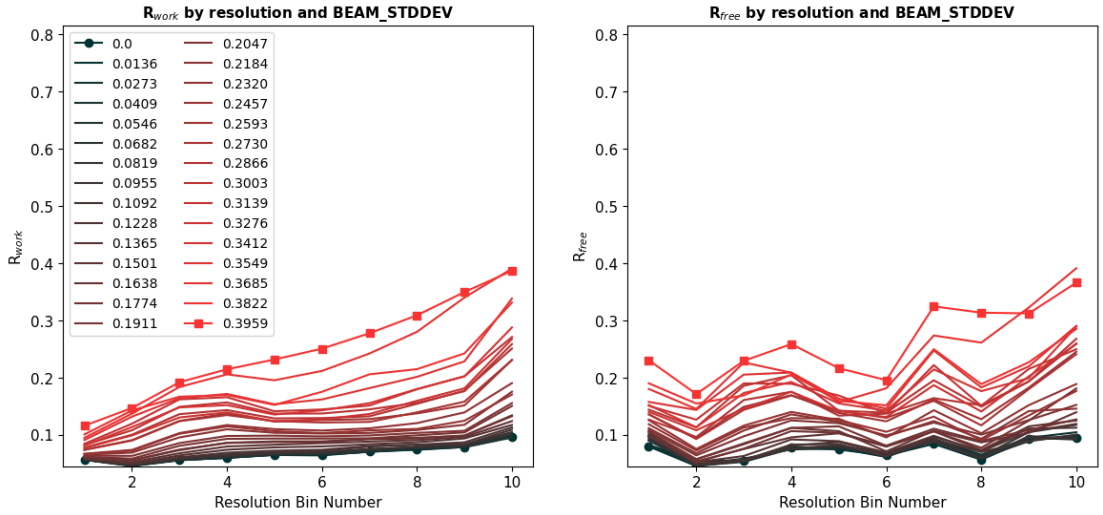


Figure 22 Change in R-values at different beam divergence values (see legend) with increasing resolution. The impact of beam divergence on R-values is again the weakest overall out of the four parameters.

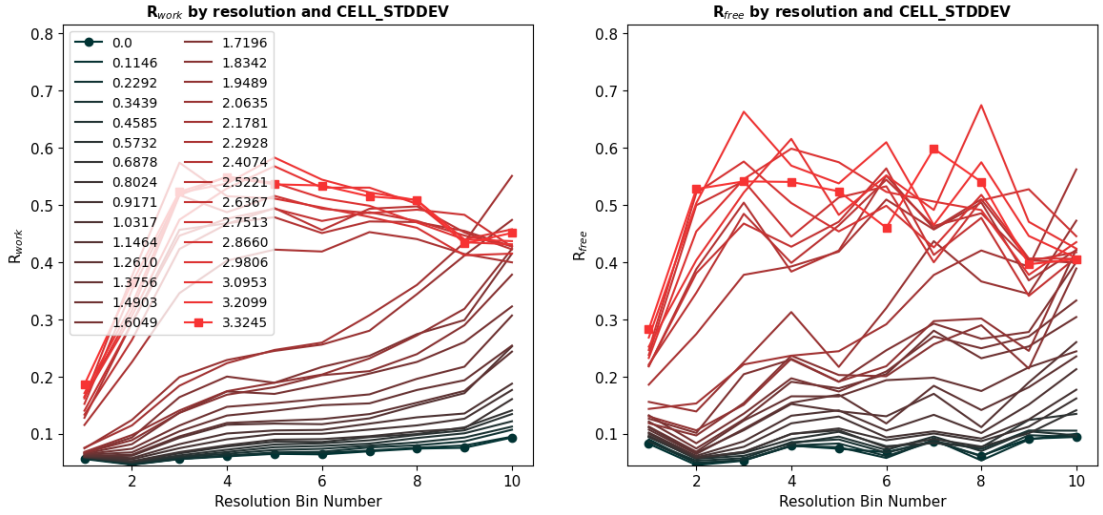


Figure 23 Change in R-values at different cell axis variation values (see legend) with increasing resolution. At resolutions beyond ~3.0 Å, data becomes unusable with large cell axis variation values, since R-values rise above 0.42.

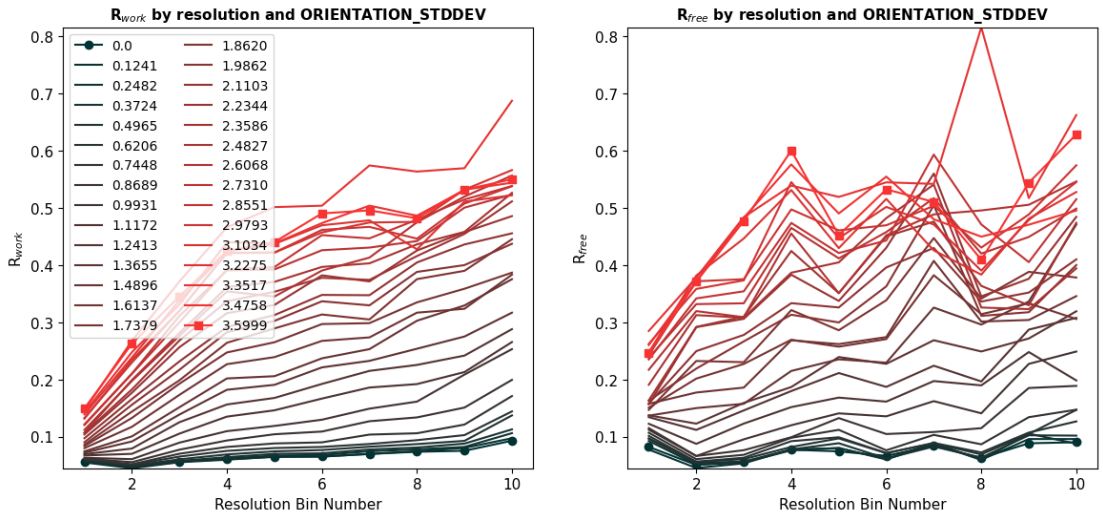


Figure 24 Change in R-values at different orientation variation values (see legend) with increasing resolution. At resolutions beyond ~3.0 Å, data becomes unusable with large orientation variation values, since R-values rise above 0.42.

3.5. Atomic B-factors

In this section, the B-factors for each atom in the unit cell are compared between the ideal starting model, and models refined against data simulated with the maximum values for each parameter. If each B-factor was increased equally, the slope of the regression line should be around one. This is not the case however, as the slopes for all parameters are closer to zero than one, indicating very low correlation between the atomic B-factors.

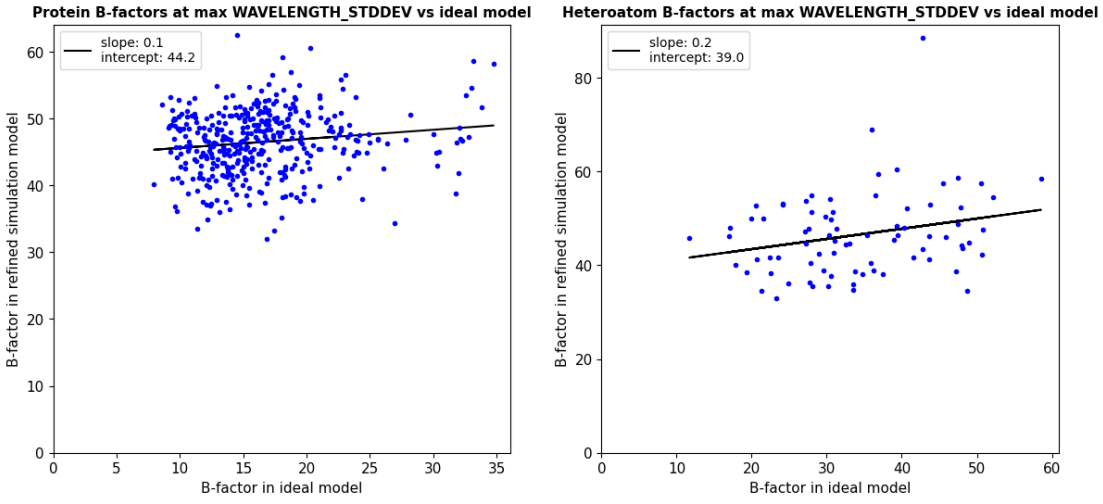


Figure 25 Comparison of B-factors for each protein and water atom between the ideal starting model and the model refined from frames with the largest simulated wavelength dispersion value (0.1726 Å). The regression slopes are 0.1 for atomic B-factors belonging to the protein, and 0.2 for atomic B-

factors of water molecules. Absolute values are the second largest among those recorded for the maximum values of each parameter.

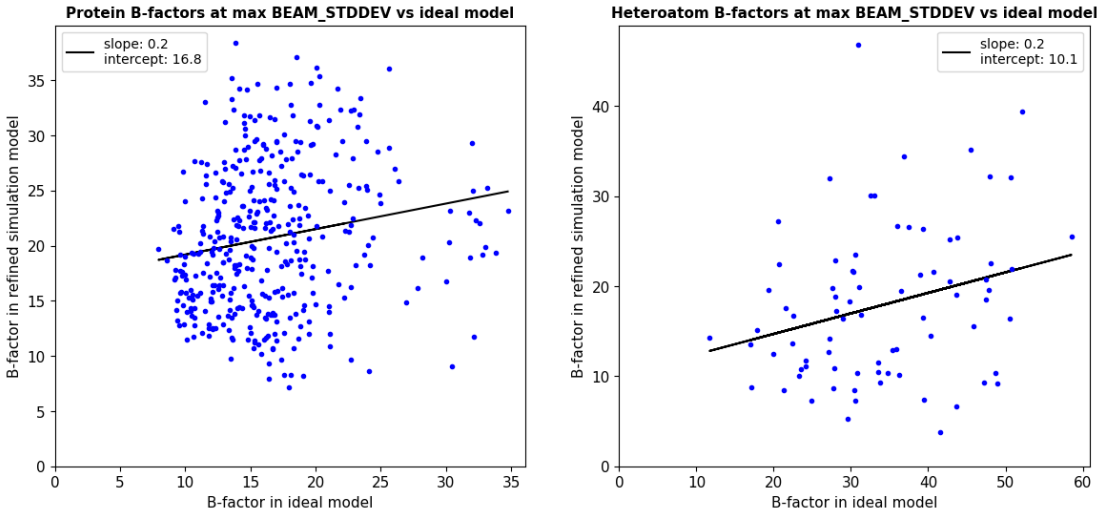


Figure 26 Comparison of B-factors for each protein and water atom between the ideal starting model and the model refined from frames with the largest simulated beam divergence value (0.3959 Å). The regression slopes are 0.2 for atomic B-factors of both protein and water molecules. Absolute values are the smallest among those recorded for the maxima of each parameter.

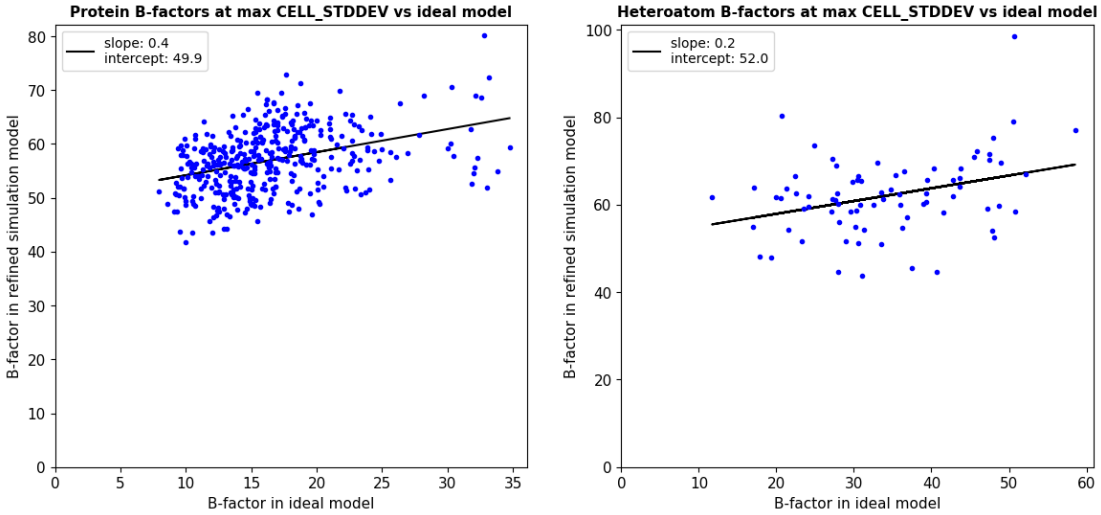


Figure 27 Comparison of B-factors for each protein and water atom between the ideal starting model and the model refined from frames with the largest simulated cell axis variation value (3.3245 Å). The regression slopes are 0.4 for B-factors of protein atoms, and 0.2 for atomic B-factors of water molecules. Absolute values are the largest among those recorded for the maxima of each parameter.

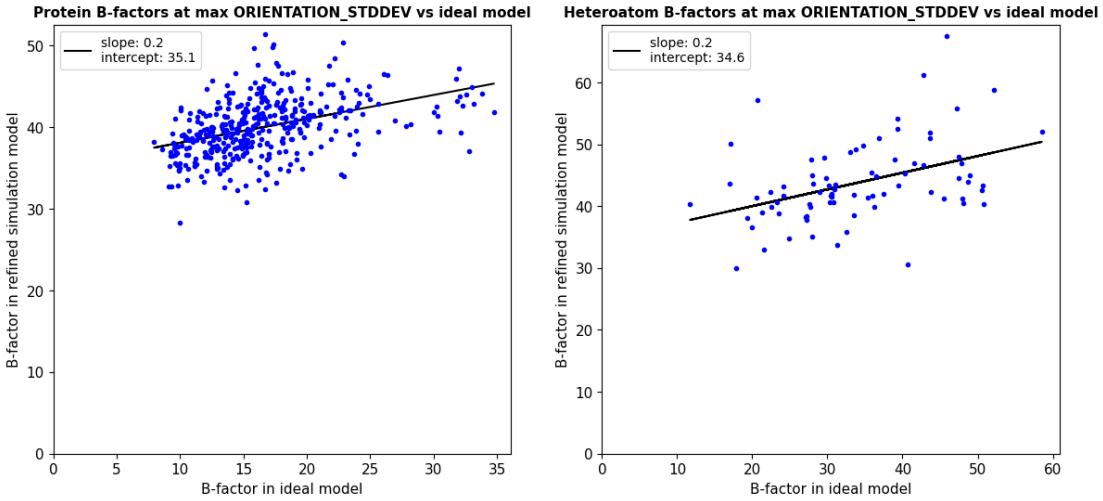


Figure 28 Comparison of B-factors for each protein and water atom between the ideal starting model and the model refined from frames with the largest simulated orientation variation value (3.6 Å). The regression slopes are 0.2 for atomic B-factors of both protein and water molecules. Overall values are the second smallest among those recorded for the maxima of each parameter.

3.6. Wilson plots

The plots in this section show the attenuation of intensities with increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation at different resolutions. While the x-axis ticks display corresponding resolution values in Ångström for ease of interpretation, what is actually plotted are $\mathbf{s}^2$-values (s = 1/resolution), as is standard for Wilson plots.

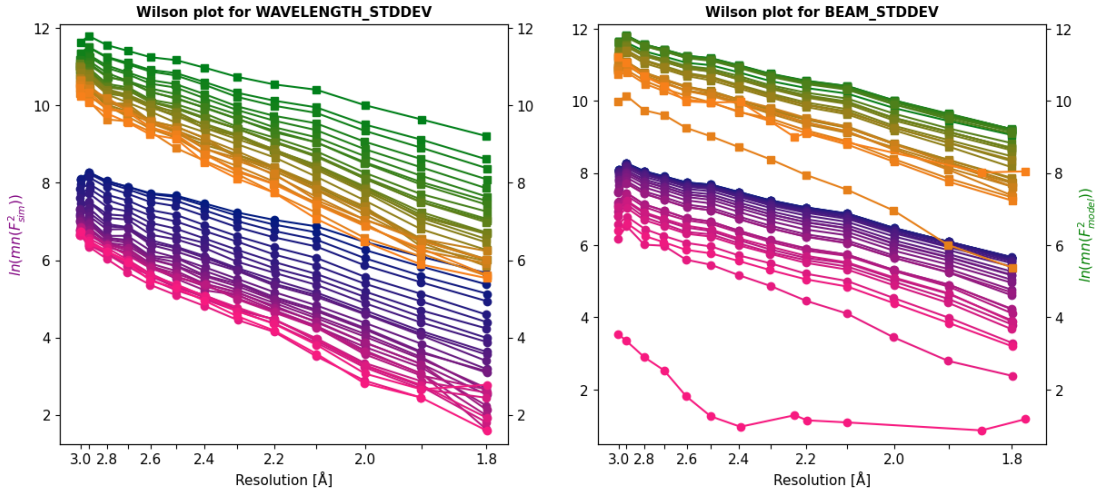


Figure 29 Natural logarithm of simulated (F^2_{sim}) and modelled (F^2_{model}) mean intensities with increasing resolution at different wavelength dispersion / beam divergence values, starting at 0 (green

squares / blue circles) and increasing to the respective maxima (orange squares / pink circles, see 2.3). Wavelength dispersion causes greater overall attenuation of intensities than beam divergence, but the largest simulated beam divergence value causes an extreme reduction in intensities (compared to the trend up to that point), which noticeably changes the shape of the plotted line relative to those at smaller values.

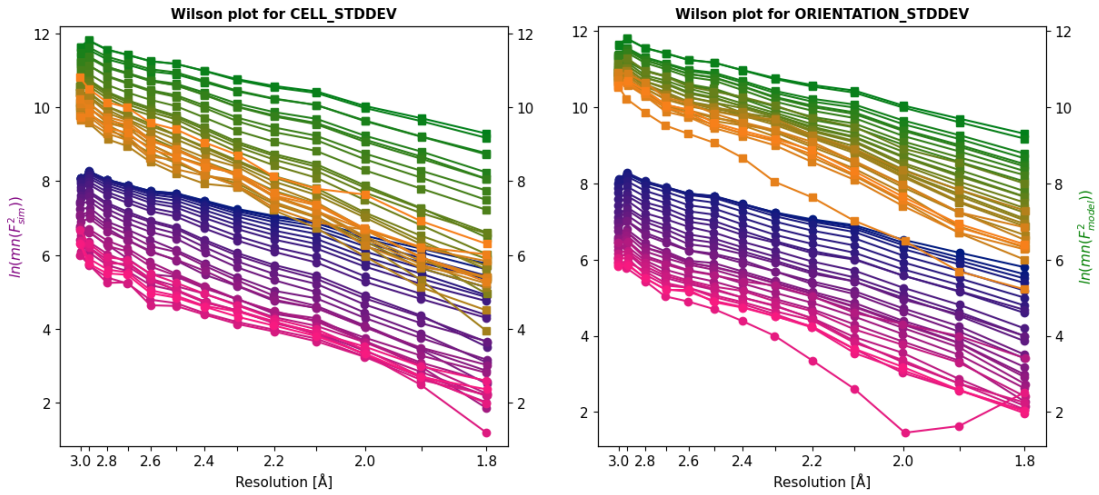


Figure 30 Natural logarithm of simulated (F^2_{sim}) and modelled (F^2_{model}) mean intensities with increasing resolution at different cell axis / orientation variation values, starting at 0 and increasing to the respective maxima (see 2.3). Cell axis variation causes the greatest overall intensity attenuation out of the four simulated parameters, while the largest simulated value for orientation variation causes a noticeable change in the shape of the plotted line compared to those at smaller values (similar to beam divergence).

3.7. Estimated standard deviation of beam divergence and reflecting range

The plots in this section illustrate the effects of beam divergence, wavelength dispersion, cell axis variation, and orientation variation on two estimates made by **XDS**. While "beam divergence" is one of the parameters simulated by **SIM_MX**, "beam divergence e.s.d" is the estimate which **XDS** makes for the same physical effect.

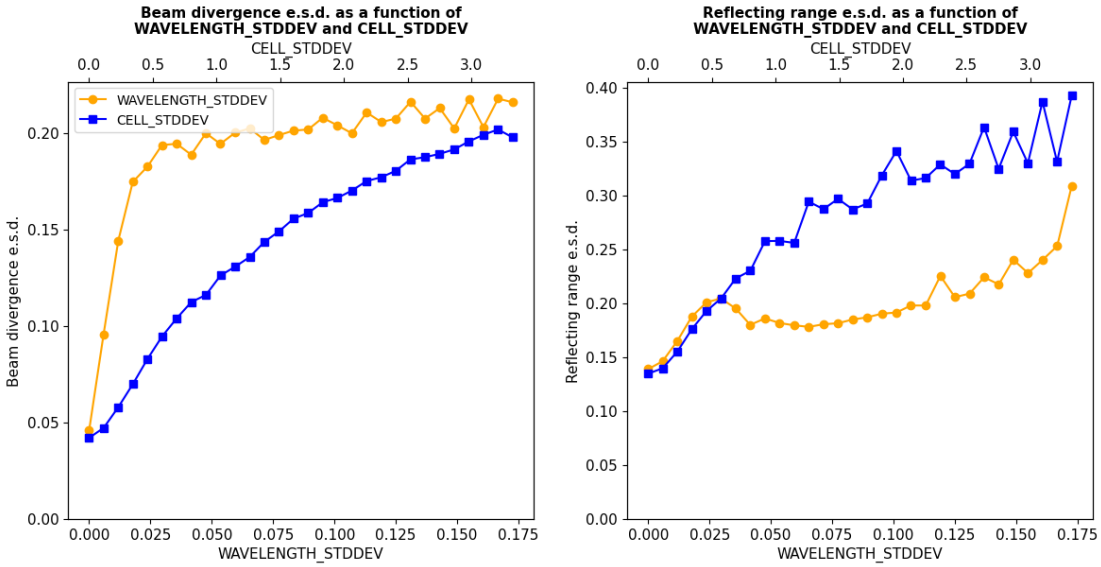


Figure 31 Estimated standard deviation (e.s.d.) of beam divergence and reflecting range, determined by **XDS**, with increasing values of wavelength dispersion and cell axis variation. Simulated STDDEV parameters start at zero and go up to the respective maxima (see 2.3). Both parameters affect beam divergence e.s.d., with wavelength dispersion causing a particularly strong increase even at small values. Cell axis variation has a stronger impact than wavelength dispersion on the estimated standard deviation of reflecting range.

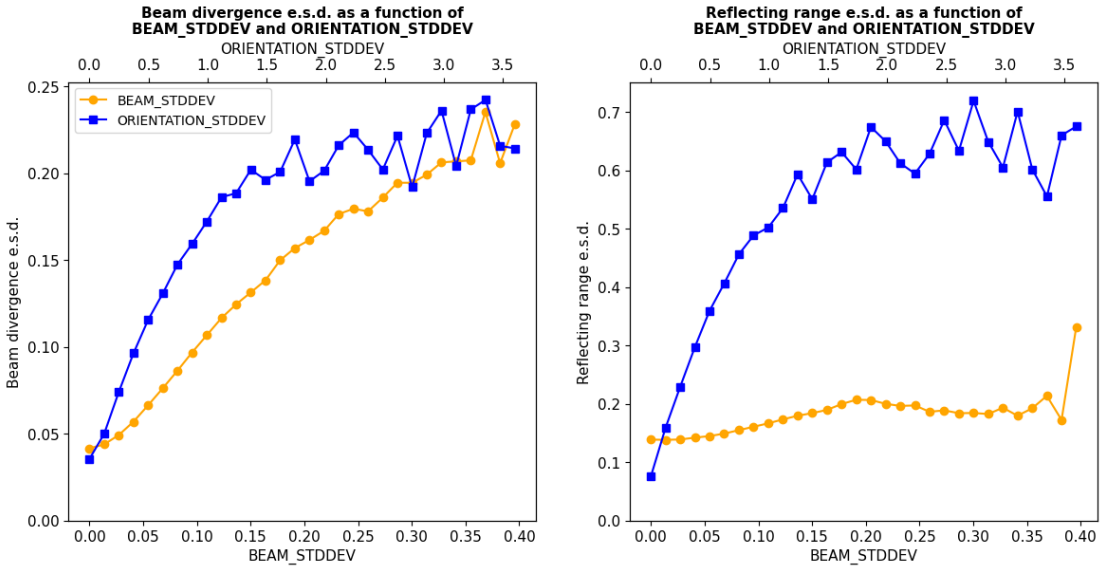


Figure 32 Estimated standard deviation (e.s.d.) of beam divergence and reflecting range, determined by **XDS**, with increasing values of beam divergence and orientation variation. Simulated STDDEV parameters start at zero and go up to the respective maxima (see 2.3). Both parameters increase beam divergence e.s.d., with the effect of simulated beam divergence being approximately linear, while

orientation variation causes an even faster increase in the indicator. Orientation variation also has the strongest effect out of all parameters on reflecting range e.s.d., which is barely affected by simulated beam divergence.

3.8. $CC_{1/2}$

This plot shows $CC_{1/2}$ in the resolution range 1.85 – 1.8 Å with increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation. $CC_{1/2}$ was used to determine the maximum values for each parameter, which is why the end points of all four lines lie so close together.

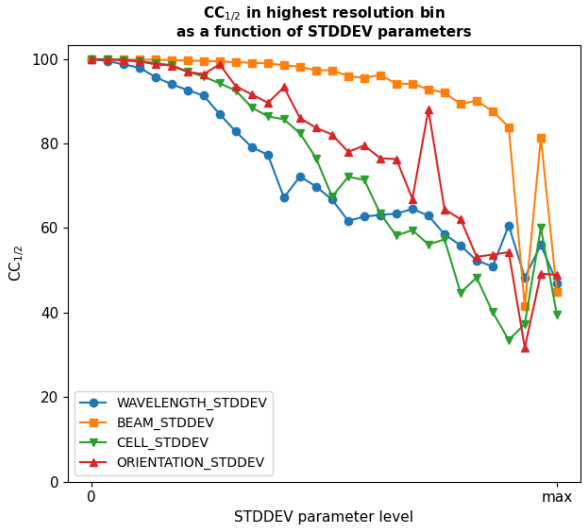


Figure 33 $CC_{1/2}$ in the highest resolution bin (1.85 – 1.8 Å) with increasing values of simulated wavelength dispersion, beam divergence, cell axis variation, and orientation variation. All parameters start at zero and go up to their respective maxima (see 2.3). Beam divergence in particular has a rather weak effect on this indicator until very large simulated values, at which the data becomes so difficult to interpret that most quality indicators, including $CC_{1/2}$, start fluctuating quite strongly.

3.9. Root-mean-square deviation of atomic positions

This plot illustrates the extent to which increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation affect the atomic coordinates of the refined model.

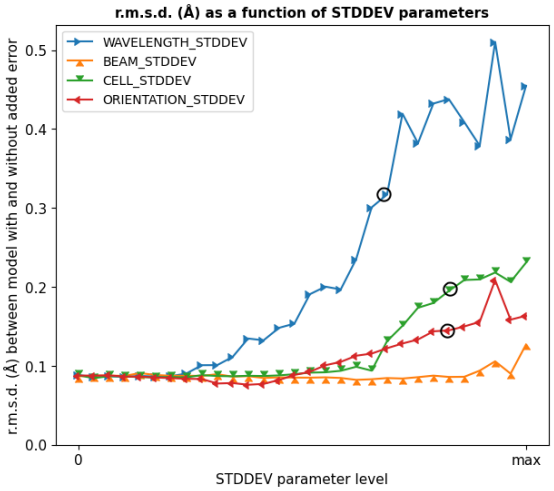


Figure 34 Root-mean-square deviation (r.m.s.d.) of atomic coordinates between the ideal reference model and four models refined against data with increasing values of simulated wavelength dispersion, beam divergence, cell axis variation, and orientation variation. Simulated STDDEV parameters start at zero and go up to the respective maxima (see 2.3). Wavelength dispersion has the strongest impact on this indicator, while beam divergence barely affects it. The black circles mark the first value for each parameter at which final-R_{free} increases beyond 35 (see figures 17 – 20).

3.10. Geometric quality indicators

The plots in this section show the changes in geometric quality indicators reported by MolProbity with increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation.

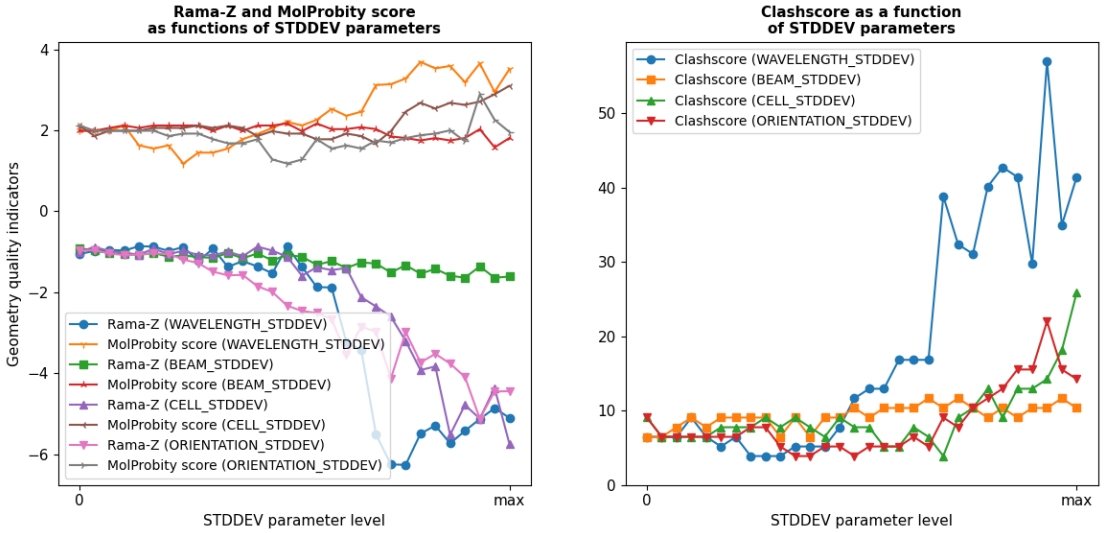


Figure 35 Change of Rama-Z, Clashscore, and MolProbity score (all determined by **phenix.molprobity**) with increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation. Simulated STDDEV parameters start at zero and go up to the respective maxima (see 2.3). MolProbity score increases most due to wavelength dispersion, followed by cell axis variation, with beam divergence again having the lowest impact and even causing the score to improve at large parameter values. Rama-Z is strongly affected by large values of all parameters except beam divergence, with the latter only causing a very slight decrease. Clashscore is most affected by increased wavelength dispersion, while beam divergence again has the weakest impact.

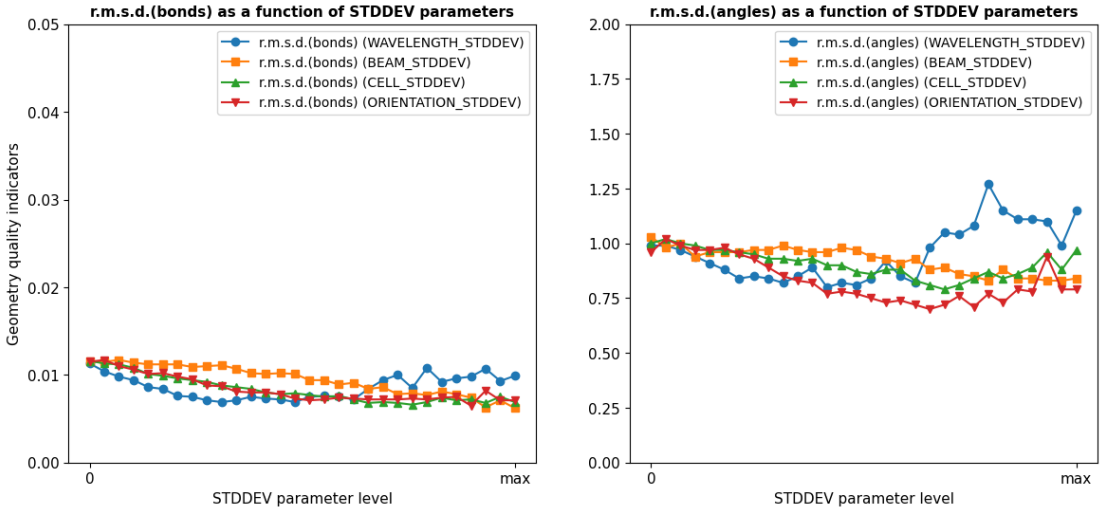


Figure 36 Change of root mean square deviation (r.m.s.d.) of atomic bond lengths and torsion angles (determined by **phenix.molprobity**) with increasing values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation. Simulated STDDEV parameters start at zero and go up to the respective maxima (see 2.3). Apart from the slight increase of torsion angle r.m.s.d. caused by large values of wavelength dispersion, neither indicator is significantly affected by the simulated parameters. In fact, the r.m.s.d. values slightly decrease as parameter values increase, indicating an improvement in structure geometry.

3.11. Refined models with electron density difference maps

This figure shows models refined against data simulated with the largest values for wavelength dispersion, beam divergence, cell axis variation, and orientation variation. Image was created with **Coot** (Crystallographic Object-Oriented Toolkit) [Emsley et al., 2010].

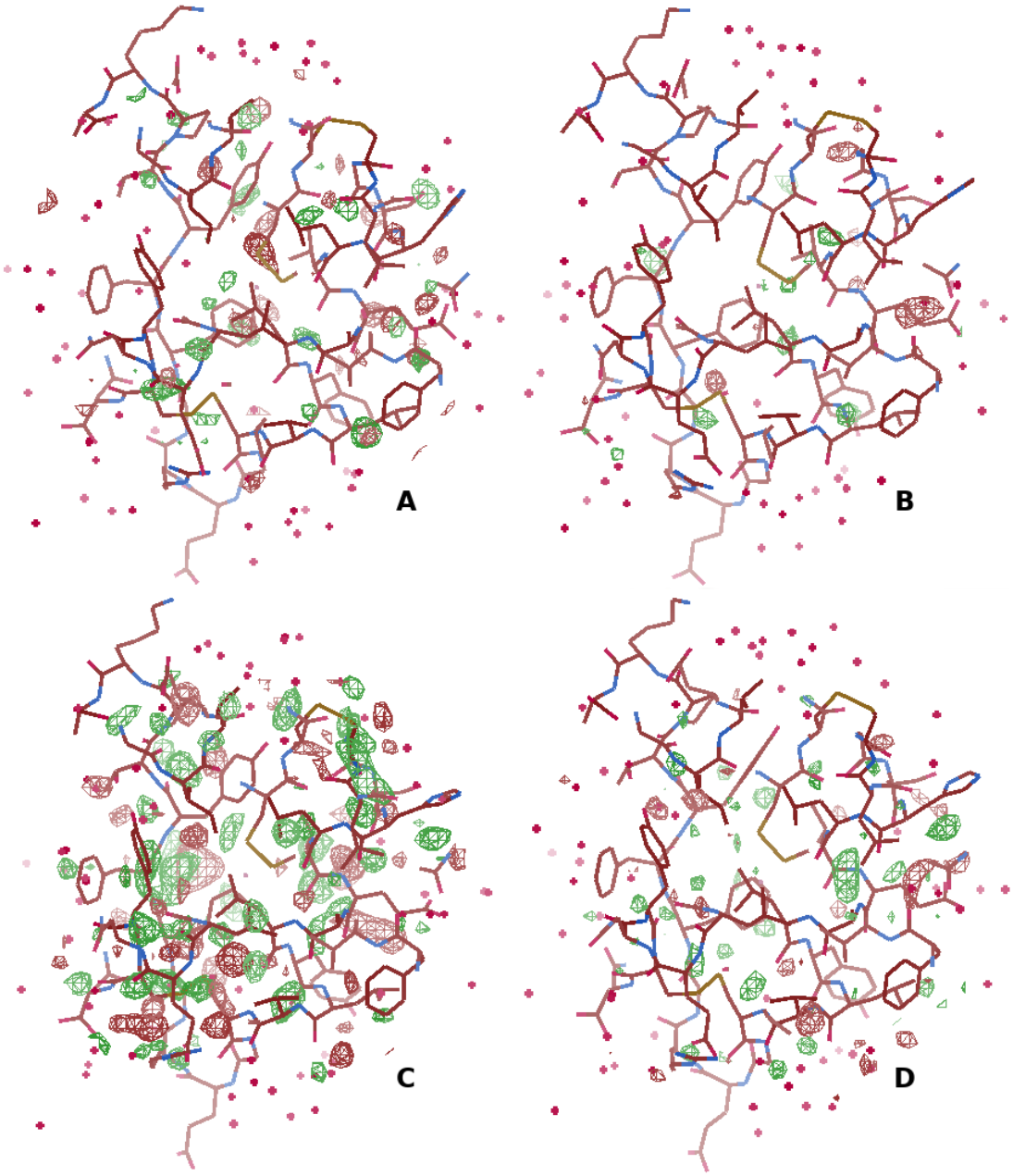


Figure 37 Comparison of refined models and electron density difference maps at the respective maximum values (see 2.3) for **A** wavelength dispersion, **B** beam divergence, **C** cell axis variation, and **D** orientation variation. Red crosses represent water molecules; green grid areas represent positive electron density difference; red grid areas represent negative electron density difference. While the differences between the refined models are relatively minor and mostly concern the torsion angles, the electron density difference maps show that cell axis variation causes the strongest disagreement between data and model. The second strongest disagreement appears to be caused by wavelength dispersion, followed by orientation variation. Beam divergence once again appears to have the smallest impact on data quality.

4. Discussion

In this section, possible explanations for the specific impacts of each simulated parameter on reflection shapes, data quality indicators, and model properties will be presented. In addition, the reliability and possible interpretations of certain indicators will be discussed, based on the different ways in which they are affected by each parameter. The discussion will focus on the results for 2bn3, with occasional mentions of significant differences to 1dpx. Whenever there is mention of a trend correlating with "increasing" resolution, what is actually meant is a correlation with s^2 (s = 1 / resolution), because "higher" resolution corresponds to smaller Ångström values.

4.1. Reflection shape

Wavelength dispersion causes radial elongation of reflection spots and appearance of additional reflections outside the predicted resolution shell areas (see figure 2 and 3). Since the wavelength of a beam determines the radius of the Ewald sphere, one may imagine the sphere becoming "thicker", as multiple spheres layer on top of each other, when multiple beams covering a range of wavelengths are diffracted at once. This leads to reciprocal lattice points (RLPs) intersecting with multiple spheres during the rotation range for a single frame, and to additional RLPs (which would not usually appear on the frame) intersecting with at least some of the spheres.

Simulated beams have a "primary" direction towards the detector centre. Beam divergence is simulated by adding random angles along two axes (orthogonal to each other and to the primary beam) from a Gaussian distribution around zero to each beam's direction. Depending on the rotation range that is measured for each frame, most of these beams still satisfy the Bragg condition. This leads to a Gaussian distribution of spots around the "true" position of each reflection on the detector. Since all reflections are simulated to have the same distance from the crystal, as would be the case with a curved detector, the area which is covered by this distribution does not increase with resolution. While no additional reflections outside the predicted areas are visible in figures 4 and 5, they do appear on some frames.

Orientation variation causes circular elongation of reflection shapes and appearance of additional reflections (see figures 8 and 9). Since this parameter simulates rotation around three orthogonal axes, it might seem surprising that the reflection shapes appear to be

distorted in mainly one dimension. However, this can be explained with another thought experiment. Individual mosaic blocks are simulated as perfect, infinite crystals which share the same reciprocal lattice origin. Rotation of these blocks means rotation of the corresponding reciprocal lattices. Rotation of a reciprocal lattice will displace RLPs on the surface of spheres around the origin, since their distances from the origin do not change. Random, normally distributed rotations of multiple lattices along three axes will therefore result in each RLP being replaced by a partial sphere. When two spheres intersect, the intersecting points form a circle. Thus, a partial sphere intersecting with the Ewald sphere results in a reflection shaped like a partial circle. It also follows that additional reflections would appear on a frame, since a partial sphere covers a greater area than a single point.

Finally, the irregular distortion of reflection spots due to cell axis variation may be explained as follows: Changing the length of a unit cell axis will change the spacing of different sets of lattice planes, as well as the angles of their orientation, to varying degrees. Since the distance of each RLP from the reciprocal lattice origin depends on the spacing of the corresponding set of planes, and its rotational displacement depends on the orientation of the crystal, changes in the length of all three unit cell axes can move an RLP in any direction, as opposed to the limited movement caused by orientation variation. In addition, the magnitude and direction of displacement will differ between RLPs, which explains the difference in spot shape between detector regions (see figure 6).

4.2. Number of contributing and overlapping pixels per reflection

With increasing values of all parameters, reflection overlap increases more steeply and to larger values at higher resolutions. This is likely due to the greater number of reflections at higher resolution (the reciprocal space volume which can intersect with the Ewald sphere increases with the third power of the resolution sphere radius).

The increase in the sum of contributing and overlapping pixels per reflection (overall reflection size) with increasing wavelength dispersion values is initially steeper than for the other parameters, but flattens at larger parameter values in the highest resolution range (figure 15). A possible reason may be that the elongated reflection shape leads to significant parts of the high resolution intensities lying outside of the detector area.

Beam divergence causes a slower but also steadier increase in overall reflection size, with a weaker effect than wavelength dispersion at small values, but a stronger effect at large values (figure 15). Again, the increase flattens at large parameter values in the highest resolution range. The reason may be the same as for wavelength dispersion, since the change in reflection shape caused by beam divergence also has a radial component.

Cell axis variation causes the second strongest reflection overlap at high resolution, and no overlap at all below 10 Ångström. Looking at figure 7, it can be seen that high resolution reflections are distorted more than those close to the beamstop (this also appears to be the case for orientation variation). A possible explanation is the increase in radial and rotational displacement of RLPs which results from greater distance to the reciprocal lattice origin.

Orientation variation causes the greatest overall reflection size out of all parameters at its largest simulated value, mainly due to extreme (compared to what results from the other parameters) overlap between reflections in the highest resolution range. Said overlap probably results from the increase in RLP size with resolution. Additionally, orientation variation is the only parameter that does not have a radial component in the distortion it causes, meaning that a smaller part of the high resolution intensities ends up outside the detector area. In the simulations for 1dpx, overlap as well as overall reflection size for this parameter do not increase nearly as much, mainly because $CC_{1/2}$ sinks below 40 at much smaller values than for 2bn3, leading to a smaller maximum value.

4.3. R-factors

According to the R-values, agreement between the ideal model and the simulated data is most affected by increased wavelength dispersion, followed by orientation variation and then cell axis variation (figures 17 – 20). This is likely due to the bad fit of the elongated reflection shapes caused by wavelength dispersion with the areas predicted by **XDS** (figures 2 and 3). In contrast, the broadened reflection shapes (figure 4) caused by increased beam divergence are mostly contained in the predicted areas, and only have a relatively weak effect on the R-values (figure 18). While the largest simulated wavelength dispersion value increases R_{free} above 50 after refinement, it should be noted that said value is way beyond any level of wavelength dispersion that would be encountered in a real experiment. For polychromatic beams used in pink-beam serial crystallography, commonly used wavelength dispersion

values are about an order of magnitude below the maximum value which was simulated for this project [Tolstikova, 2020]. For 1dpx, the order of the degrees to which parameters affect R-values is different, with cell axis variation having the strongest impact (see appendix A.2). This shows that the relative impact of specific parameters depends on unit cell properties.

4.4. B-factors

Overall B-factors are most affected by cell axis variation, followed by wavelength dispersion, then orientation variation, with beam divergence again having the weakest impact (figures 17 – 20, 25 – 29). This illustrates the extent to which each parameter causes reflection intensity to decrease with rising resolution, since B-factors are proportional to the slope of that decrease. In terms of the physical reality of the crystal, these trends can be interpreted as follows: Since cell axis variation changes the shape of the unit cell, and therefore individual atomic positions, it makes sense that it would increase crystal disorder. Similarly, rotation of mosaic blocks will change atomic positions, which should be reflected in the B-factors. Wavelength dispersion on the other hand is not related to crystal disorder, and yet it causes a strong increase in B-factors. For this reason, one may want to adjust B-factors in the final model when working with polychromatic radiation. The extent of the necessary corrections could be empirically derived from simulated data for the model in question.

From the regression slopes of the atomic B-factor comparisons, one can discern that all simulated parameters increase the displacement of individual atoms more strongly for initially well-ordered atoms, because the regression slopes are far smaller than one. This seems plausible, as atoms that already partially occupy multiple positions may require less adjustment when fitting a model to the simulated data during refinement.

4.5. Wilson plots

Attenuation of all reflection intensities is strongest at high values of cell axis variation, closely followed by wavelength dispersion and orientation variation. The same trend holds for the increase in slope of the plotted lines, which again shows the effect on the B-factors. While beam divergence has the weakest effect, figure 29 still shows that the overall data quality is starting to degrade quite severely at the largest simulated beam divergence value, indicating that there would be little value in simulating even higher values.

4.6. ISa

ISa is supposed to provide information about the quality of an experimental setup. Since all simulated parameters represent deviations from an "ideal" experiment, they should all decrease this indicator. For wavelength dispersion and cell axis variation, even low values show the expected impact (figures 17 and 19). Surprisingly, increasing levels of beam divergence and orientation variation initially increase ISa, indicating an improvement in the highest achievable signal-to-noise ratio (figures 18 and 20). This is because **XDS** expects a Gaussian distribution of pixel intensities around the maximum for each reflection, which is not the case for extremely sharp reflection spots like those simulated at very low parameter values. Beam divergence, and to some extent orientation variation, cause those reflection shapes to become more Gaussian, which is interpreted by **XDS** as a reduction in systematic error. In real experiments this should not be an issue, since reflection spots will usually not be as sharp.

4.7. Estimated standard deviation of beam divergence and reflecting range

The relationship between standard deviation of beam divergence estimated by **XDS**, and beam divergence simulated by **SIM_MX** is approximately linear as one would expect. However, all other parameters have a similar or even stronger impact on this indicator, probably because **XDS** estimates beam divergence solely based on the size of reflections, rather than their shape (figures 31 and 32). Less surprisingly, estimated standard deviation of reflecting range (angular range that a crystal rotates through during the recording of a single diffraction image) is most affected by orientation variation, followed by cell axis variation. The former parameter effectively simulates increased reflecting range along three axes. For the latter, it makes sense that the change in atomic positions, caused by varying unit cell shapes, would be partially interpreted as rotation of the crystal. One conclusion that may be drawn from this is that, in the case of the actual reflecting range variation being known or negligible, this parameter could be used to estimate crystal mosaicity.

4.8. $CC_{1/2}$

This indicator estimates the signal content in the highest resolution shell, and was used to determine the maximum value for each parameter. As can be seen in figure 33, beam

divergence barely affects $CC_{1/2}$ until very large values, where high resolution intensities become weak enough to lead to jumps and fluctuations in most data quality indicators. In simulations with 1dpx (appendix A.2), much lower parameter values are sufficient for reducing this indicator below 40. This is most likely due the larger unit cell for 1dpx, which leads to more reflections and therefore more reflection overlap, which increases fastest at high resolutions (see 4.2). Overlap causes **XDS** to exclude larger numbers of pixels from the intensity summation for each reflection, and $CC_{1/2}$ depends on the accuracy and precision of intensity measurements.

4.9. Root mean square deviation of atomic positions

This indicator describes another aspect (accuracy of coordinates) of the overall agreement between ideal model and simulated data. One might therefore expect the effects of the simulated parameters to be qualitatively similar to their effects on the R-factors. This is the case to some extent, but the relative impact of each indicator is different. At the lowest simulated values of wavelength dispersion, beam divergence, cell axis variation, and orientation variation that cause R_{free} to increase beyond 35% (the point after which model phases tend to become less useful [Kay Diederichs, personal communication]), wavelength dispersion increases coordinate r.m.s.d. by ~200%, cell axis variation increases it by ~100%, and orientation variation increases it by ~50% (figure 34). Beam divergence barely affects the coordinates, which is unsurprising. The strong effect of wavelength dispersion only appears at values which are much larger than what is to be expected in a real experiment. For “realistic” parameter values, B-factors are affected more than coordinate r.m.s.d., showing that they act as a “sink” for errors arising from deviations from an ideal experimental situation. For the simulated crystal data from lysozyme, cell axis variation has a stronger impact on coordinate r.m.s.d. than wavelength dispersion does.

4.10. Geometric validation with MolProbity

The MolProbity score already starts at 1.99 for nearly perfect data (simulated with default parameter values, see figure 1). This is worse than the simulated resolution of 1.8 and therefore indicates below average structure quality, which doesn’t make much sense. Increased parameter values actually lower this indicator in some cases, possibly because it is based on comparison with real reference structures that may be more similar to slightly faulty

models than to perfect ones. Only large values of wavelength dispersion, cell axis variation, and orientation variation cause an increase in this score (see figure 35).

The Rama-Z value decreases significantly with large values of all parameters except for beam divergence, which barely affects it.

On the Clashscore, wavelength dispersion (cell axis variation for 1dpx) appears to have a strong effect, which matches up with its impact on the r.m.s.d. of atomic positions (see figure 33).

The r.m.s.d. of bond lengths and angles surprisingly decreases with increasing values of all parameters except wavelength dispersion (see figure 36). This is most likely due to the fact that geometry restraints get weighted more strongly as overall data quality decreases.

4.11. Refined models and electron density difference maps

Unlike the R-values and the r.m.s.d. of atomic positions, the electron density difference maps indicate that the agreement between model and data is decreased most by the largest simulated value for cell axis variation, rather than wavelength dispersion, which comes in second place (see figure 37).

The models resulting from refinement against data simulated with each parameter still look quite similar however, which is likely because the starting models already represent an energetic minimum for their respective proteins, and refinement therefore keeps converging on the same result.

4.12. Conclusion and outlook

This project demonstrates how the effects of different experimental parameters can be explored through data simulation to better understand their impact on the operation of data reduction and analysis software, as well as the reliability of refined structures. In addition to expanding the proof of concept given in the original publication for **SIM_MX**, it also suggests additional interpretations for certain data quality indicators. For B-factors in particular, details like the unwanted effect of wavelength dispersion, and the loss of information about local disorder with increase of certain error types are illuminated. As for the specific parameters chosen in this project, it should be kept in mind that wavelength dispersion and beam divergence are negligible factors in most diffraction experiments, which

means that practical application of the illustrated trends for these parameters is limited. The differences in the impact of each parameter between insulin (2bn3) and lysozyme (1dpx) demonstrate that the same aspect of an experimental setup can affect data from different crystals in different ways, depending on the properties of said crystals. The exact nature of the interactions between various experimental parameters and crystal attributes may be a worthwhile subject for further research. Another possible continuation of this project could compare the effects of simulated parameters on the operation of other data reduction software, and how those effects differ from the ones described here for **XDS**.

Appendix A. Automated simulation

To reproduce the plots shown in this thesis, download the "models" archive from https://github.com/jadler-13/simulations.git and extract it into an empty directory with at least 2 gigabytes of available space, then follow the instructions in the README file.

A.1. *SIMULATE.INP* for 2bn3

```
! lines starting with "!simulate" are inputs to sim_mx.
! sim_mx does not consider AIR= and FRACTION_OF_POLARIZATION=; it implies values
! of 0 and 0.5
! inputs for sim_mx (values with defaults do not have to be provided):
!simulate UNIT_CELL_A-AXIS= 77.900 0 0
!simulate UNIT_CELL_B-AXIS= 0 77.900 0
!simulate UNIT_CELL_C-AXIS= 0 0 77.900
!!simulate BIG_CRYSTAL=TRUE  ! default=TRUE; FALSE for infinitesimally small reflections
(affecting a single pixel)
!  default=TRUE which distributes counts to four adjacent pixels proportional to distance
!simulate EXPOSURE_FACTOR= 0.035 ! factor to multiply input intensities with; 1 is the default
!(don't)simulate DEBUG= -33   -13    -8 ! monitor contributions to a reflection
!(don't)simulate BACKGROUND= 30.
!simulate GAIN= 1
!(don't)simulate WAVELENGTH_STDDEV= 0.1726
!(don't)simulate BEAM_STDDEV= 0.3959 0.3959
!(don't)simulate CELL_STDDEV= 3.3246 3.3246 3.3246
!(don't)simulate ORIENTATION_STDDEV= 3.6 3.6 3.6
!simulate SENSITIVITY_IN_PHI_SUBRANGE= 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
!(don't)simulate SENSITIVITY_IN_PHI_SUBRANGE= 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
!(don't)simulate SENSITIVITY_IN_PHI_SUBRANGE= 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1
!(don't)simulate MODULATION_IN_PHI= 0.5 10 0
!
! The following inputs are for XDS; some (namely
NAME_TEMPLATE_OF_DATA_FRAMES,OSCILLATION_RANGE,X-RAY_WAVELENGTH
! DIRECTION_OF_DETECTOR_X-AXIS,DIRECTION_OF_DETECTOR_Y-AXIS,ROTATION_AXIS)
! are evaluated by sim_mx as well
!
JOB=XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE CORRECT
!JOB=CORRECT
DELPHI= 16
REFINE(INTEGRATE)= !
REFINE(CORRECT)=CELL BEAM ORIENTATION
CORRECTIONS=
REFERENCE_DATA_SET= ../../indexing_reference.hkl
DATA_RANGE= 1 180
SPOT_RANGE= 1 180
OSCILLATION_RANGE= 0.5
INCLUDE_RESOLUTION_RANGE= 50 1.79
 !======================GEOMETRICAL PARAMETERS============================
 ORGX=600  ORGY=610  !Detector origin (pixels).  ORGX=NX/2; ORGY=NY/2
 DETECTOR_DISTANCE=150.0   !(mm)
 X-RAY_WAVELENGTH= 1.0          !Angstroem
 SPACE_GROUP_NUMBER= 199 ! zero for unknown crystals; cell constants are ignored.
UNIT_CELL_CONSTANTS= 77.900 77.900 77.900 90.00 90.00 90.00
 NAME_TEMPLATE_OF_DATA_FRAMES= /home/models/2bn3/new/xds/image/insu_???.cbf
 NX=1200 NY=1200 QX=0.15 QY=0.15  !images  ???.mar1200 or ???.pck1200
 MINIMUM_VALID_PIXEL_VALUE=0  OVERLOAD=100000
! SENSOR_THICKNESS=0.0
SILICON=0.0       !Commenting this line in means no silicon efficiency correction
! DETECTOR= MAR345  MINIMUM_VALID_PIXEL_VALUE=0  OVERLOAD=130000
 DETECTOR= PILATUS  MINIMUM_VALID_PIXEL_VALUE=0  OVERLOAD=1000000
 DIRECTION_OF_DETECTOR_X-AXIS= 1.0 0.0 0.0
 DIRECTION_OF_DETECTOR_Y-AXIS= 0.0 1.0 0.0 ! 0.0 cos(2theta) sin(2theta)
 TRUSTED_REGION=0.0 1.42 !Relative radii limiting trusted detector region
!====================== GEOMETRICAL PARAMETERS============================
 ROTATION_AXIS= 1.0 0.0 0.0
```

INCIDENT_BEAM_DIRECTION=0.0 0.0 1.0

! FRACTION_OF_POLARIZATION=0.99 !default=0.5 for unpolarized beam;0.90 at DESY;

! POLARIZATION_PLANE_NORMAL= 0.0 1.0 0.0

AIR= 0.000001 ! 0.0005 !Air absorption coefficient of x-rays

!FRIEDEL'S_LAW=FALSE !Default is TRUE.

STRICT_ABSORPTION_CORRECTION=TRUE

!STARTING_ANGLE= 0.0 STARTING_FRAME=1

!used to define the angular origin about the rotation axis.

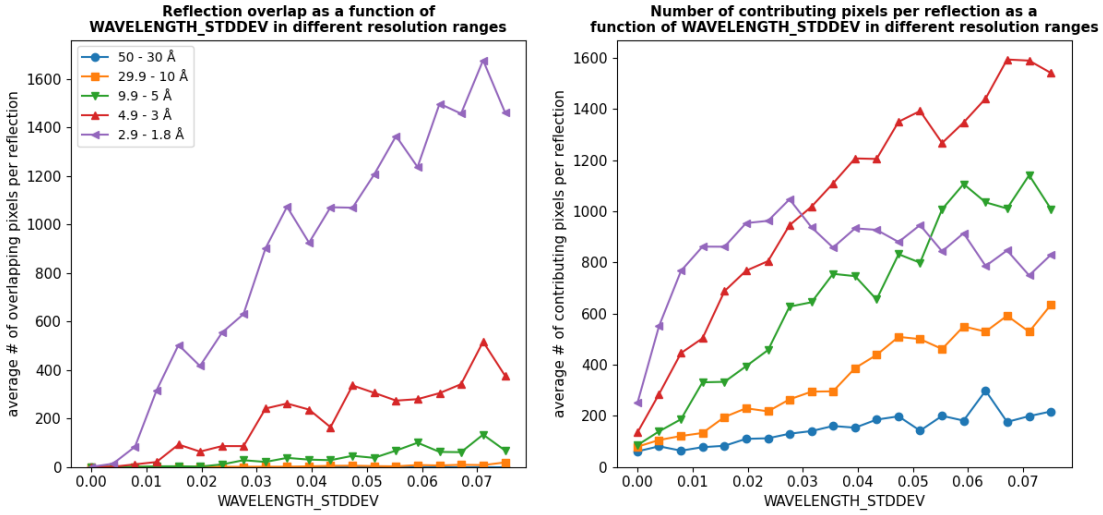
!Default: STARTING_ANGLE= 0 at STARTING_FRAME=first data image

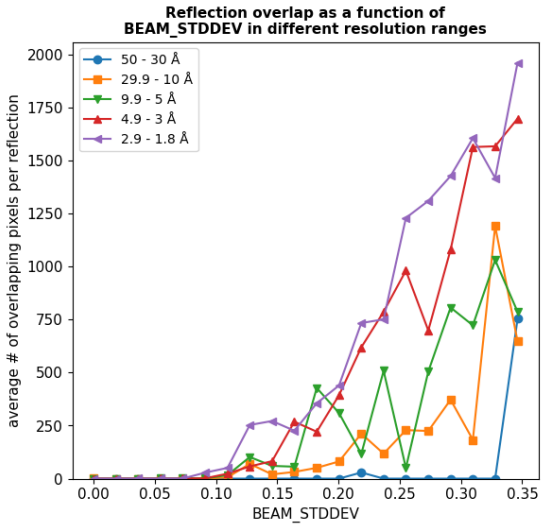
!MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT=20 !used by: COLSPOT

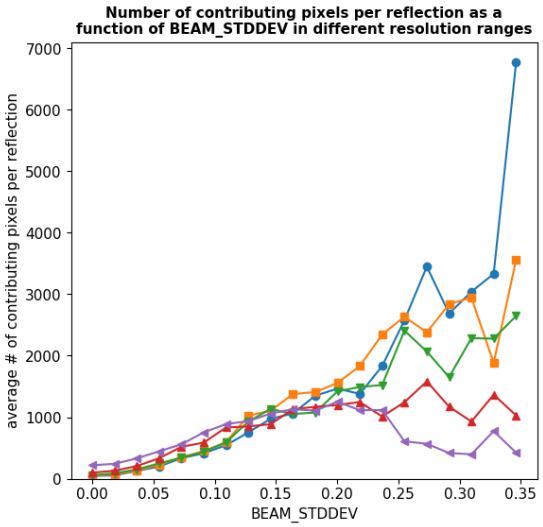
MAXIMUM_ERROR_OF_SPOT_POSITION=9.0

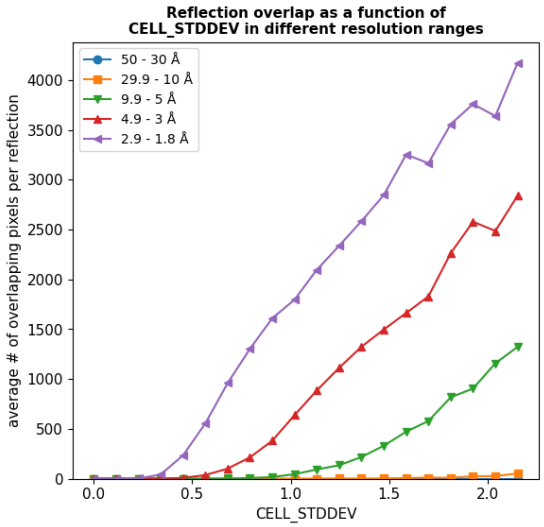
MINIMUM_FRACTION_OF_INDEXED_SPOTS=0.3

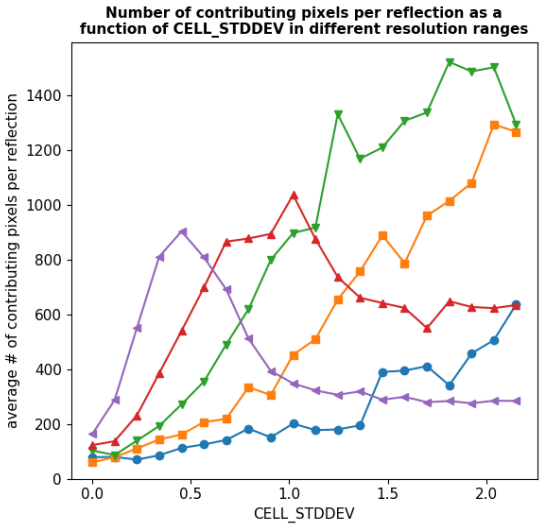
A.2. Plots for 1dpx (lysozyme)

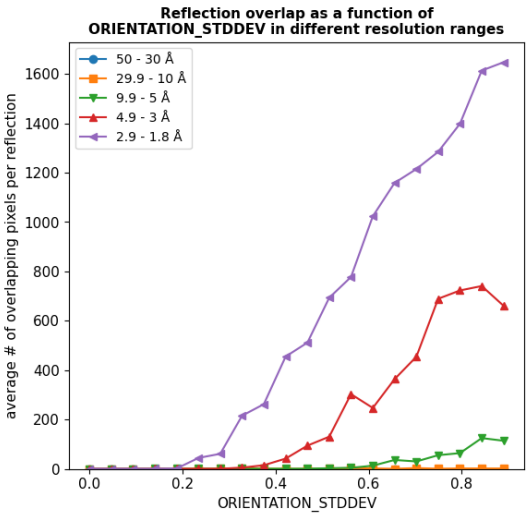


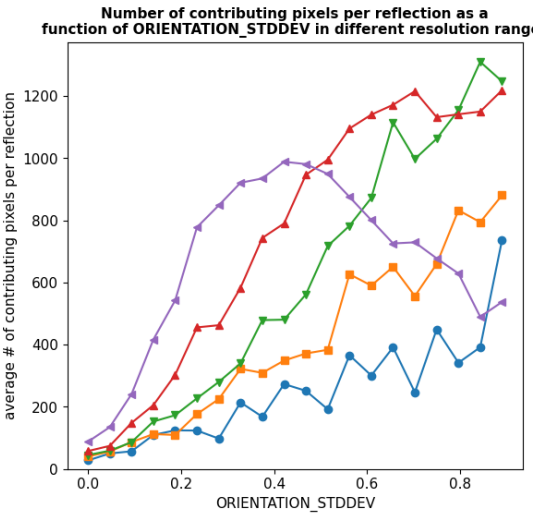

Reflection overlap as a function of
BEAM_STDDEV in different resolution ranges
50 - 30 Å
29.9 - 10 Å
9.9 - 5 Å
4.9 - 3 Å
2.9 - 1.8 Å
average # of overlapping pixels per reflection
BEAM_STDDEV

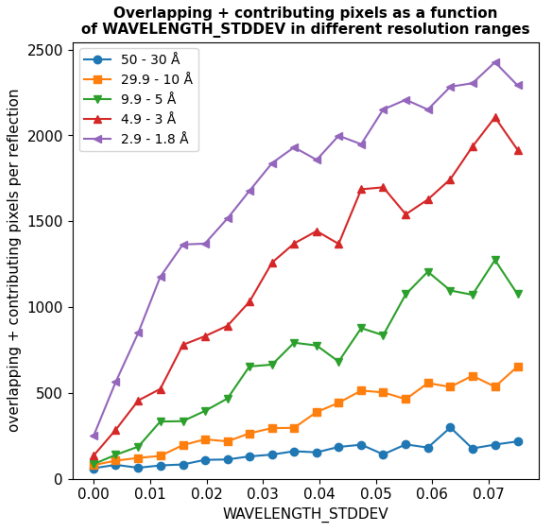

Number of contributing pixels per reflection as a
function of BEAM_STDDEV in different resolution ranges
average # of contributing pixels per reflection
BEAM_STDDEV

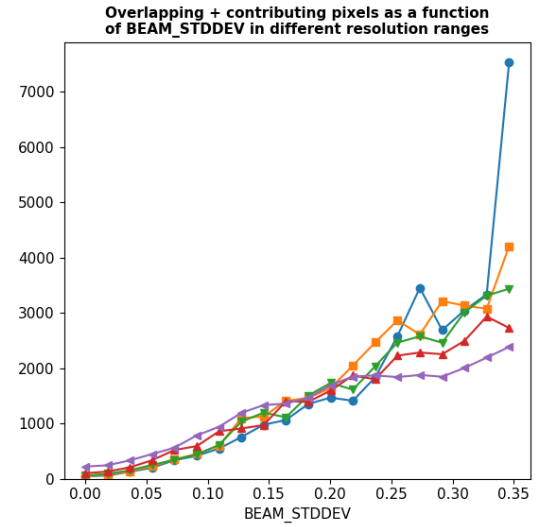

Reflection overlap as a function of
CELL_STDDEV in different resolution ranges
50 - 30 Å
29.9 - 10 Å
9.9 - 5 Å
4.9 - 3 Å
2.9 - 1.8 Å
average # of overlapping pixels per reflection
CELL_STDDEV

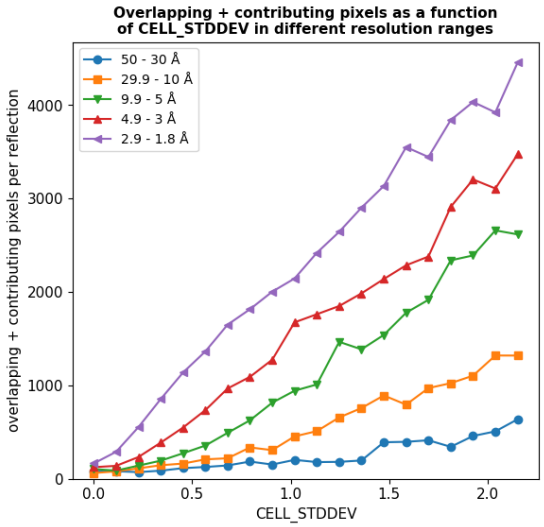

Number of contributing pixels per reflection as a
function of CELL_STDDEV in different resolution ranges
average # of contributing pixels per reflection
CELL_STDDEV

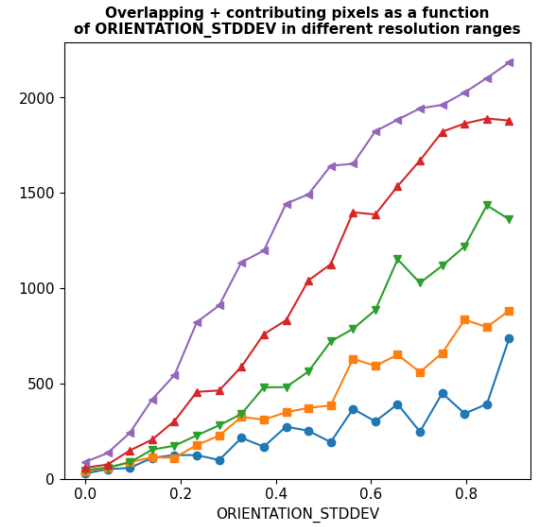

Reflection overlap as a function of
ORIENTATION_STDDEV in different resolution ranges
50 - 30 Å
29.9 - 10 Å
9.9 - 5 Å
4.9 - 3 Å
2.9 - 1.8 Å
average # of overlapping pixels per reflection
ORIENTATION_STDDEV

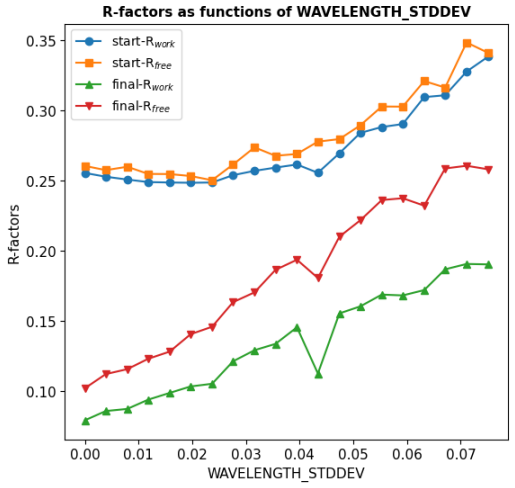

Number of contributing pixels per reflection as a
function of ORIENTATION_STDDEV in different resolution ranges
average # of contributing pixels per reflection
ORIENTATION_STDDEV

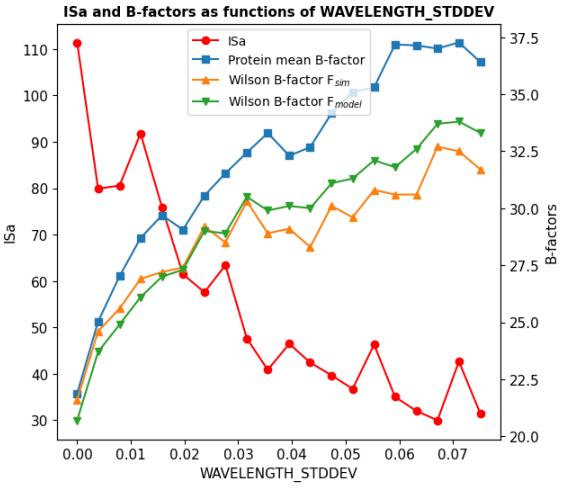

Overlapping + contributing pixels as a function of WAVELENGTH_STDDEV in different resolution ranges
50 - 30 Å
29.9 - 10 Å
9.9 - 5 Å
4.9 - 3 Å
2.9 - 1.8 Å
overlapping + contributing pixels per reflection
WAVELENGTH_STDDEV

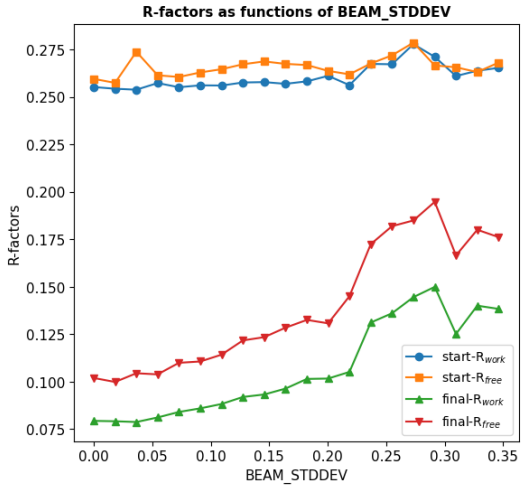

Overlapping + contributing pixels as a function of BEAM_STDDEV in different resolution ranges
BEAM_STDDEV

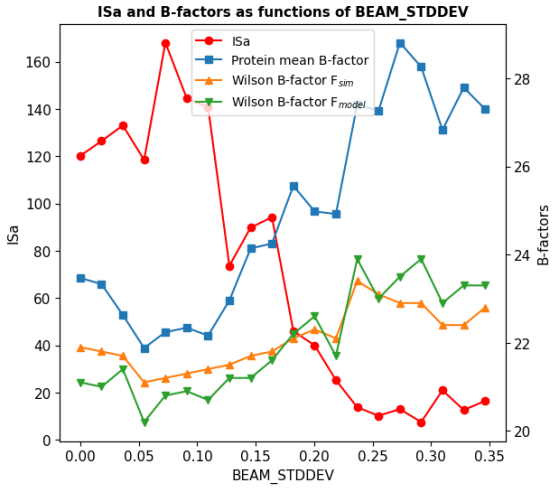

Overlapping + contributing pixels as a function of CELL_STDDEV in different resolution ranges
50 - 30 Å
29.9 - 10 Å
9.9 - 5 Å
4.9 - 3 Å
2.9 - 1.8 Å
overlapping + contributing pixels per reflection
CELL_STDDEV

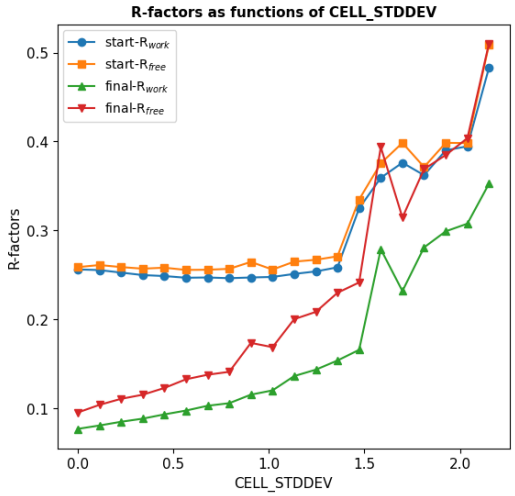

Overlapping + contributing pixels as a function of ORIENTATION_STDDEV in different resolution ranges
ORIENTATION_STDDEV

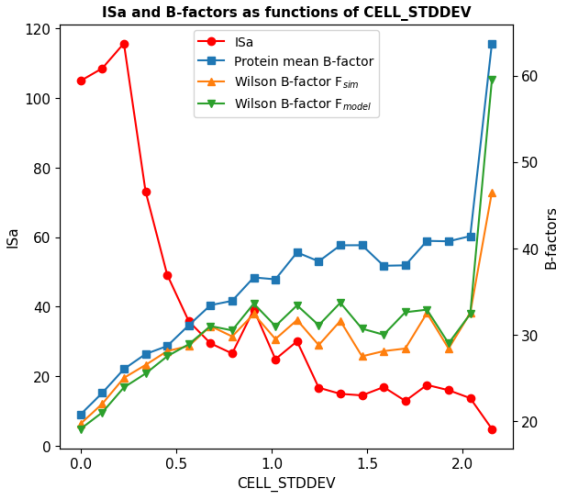

R-factors as functions of WAVELENGTH_STDDEV
start-R$_{work}$
start-R$_{free}$
final-R$_{work}$
final-R$_{free}$
R-factors
WAVELENGTH_STDDEV

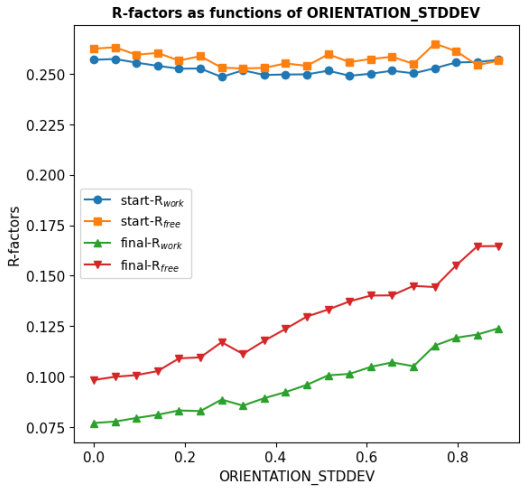

ISa and B-factors as functions of WAVELENGTH_STDDEV
ISa
Protein mean B-factor
Wilson B-factor F$_{sim}$
Wilson B-factor F$_{model}$
ISa
B-factors
WAVELENGTH_STDDEV

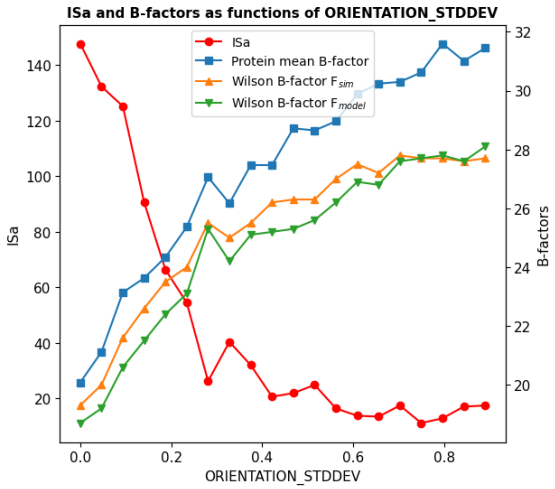

R-factors as functions of BEAM_STDDEV
R-factors
BEAM_STDDEV
start-R$_{work}$
start-R$_{free}$
final-R$_{work}$
final-R$_{free}$

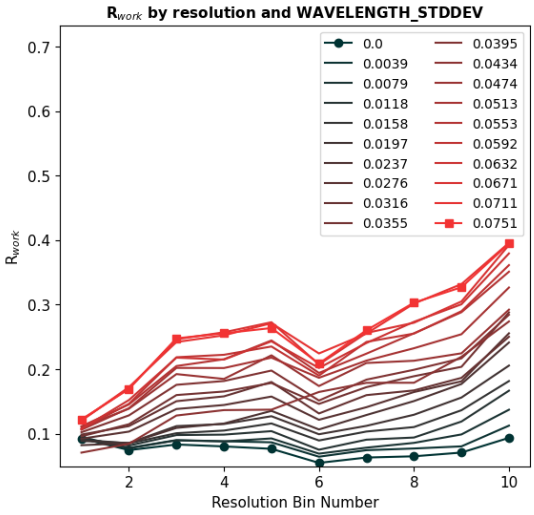

ISa and B-factors as functions of BEAM_STDDEV
ISa
B-factors
BEAM_STDDEV
ISa
Protein mean B-factor
Wilson B-factor F$_{sim}$
Wilson B-factor F$_{model}$

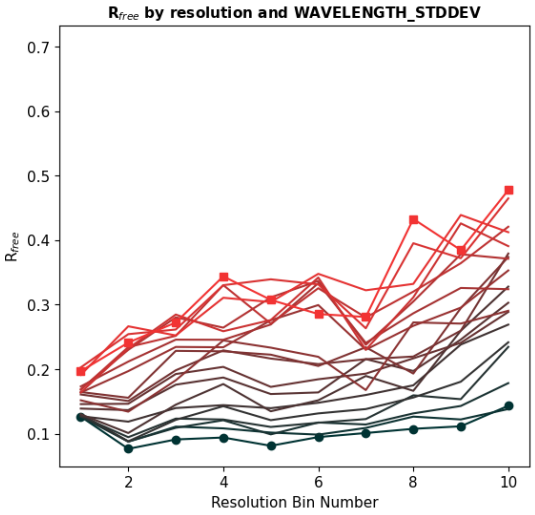

R-factors as functions of CELL_STDDEV
R-factors
CELL_STDDEV
start-R$_{work}$
start-R$_{free}$
final-R$_{work}$
final-R$_{free}$

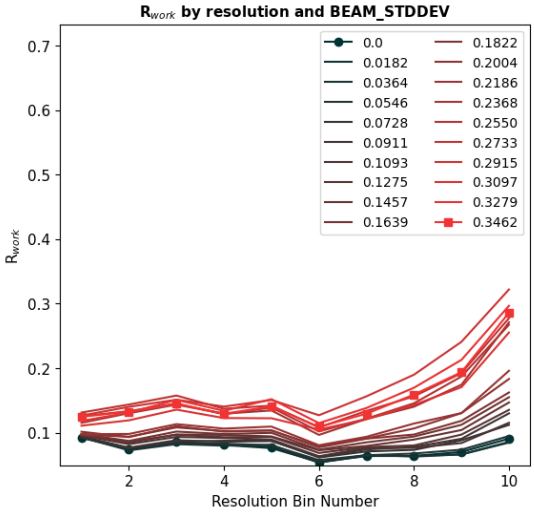

ISa and B-factors as functions of CELL_STDDEV
ISa
B-factors
CELL_STDDEV
ISa
Protein mean B-factor
Wilson B-factor F$_{sim}$
Wilson B-factor F$_{model}$

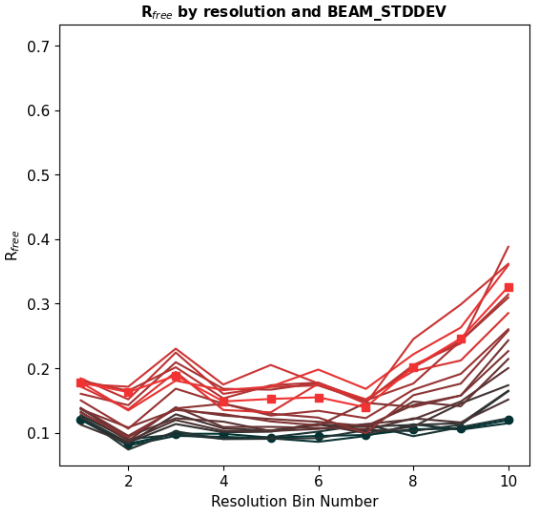

R-factors as functions of ORIENTATION_STDDEV
R-factors
ORIENTATION_STDDEV
start-R$_{work}$
start-R$_{free}$
final-R$_{work}$
final-R$_{free}$

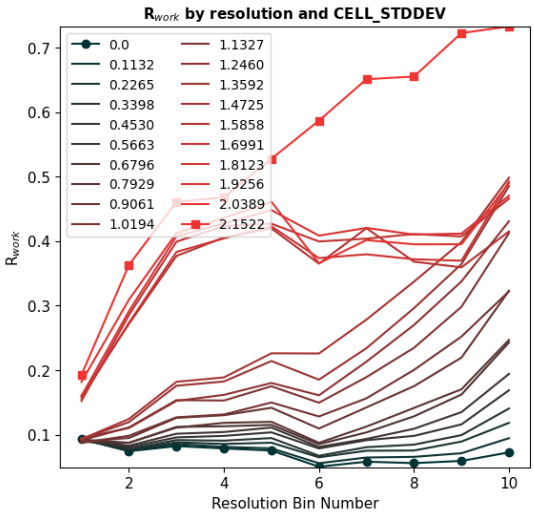

ISa and B-factors as functions of ORIENTATION_STDDEV
ISa
B-factors
ORIENTATION_STDDEV
ISa
Protein mean B-factor
Wilson B-factor F$_{sim}$
Wilson B-factor F$_{model}$

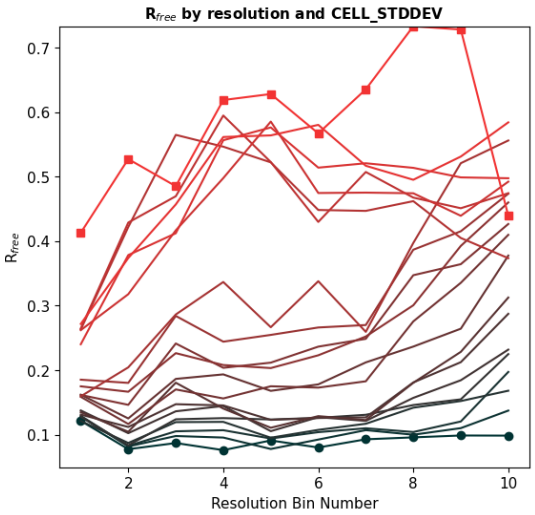

R_{work} by resolution and WAVELENGTH_STDDEV
0.0
0.0039
0.0079
0.0118
0.0158
0.0197
0.0237
0.0276
0.0316
0.0355
0.0395
0.0434
0.0474
0.0513
0.0553
0.0592
0.0632
0.0671
0.0711
0.0751
R_{work}
Resolution Bin Number

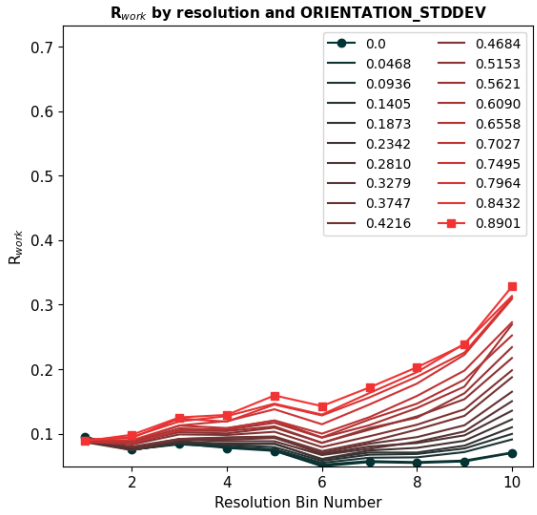

R_{free} by resolution and WAVELENGTH_STDDEV
R_{free}
Resolution Bin Number

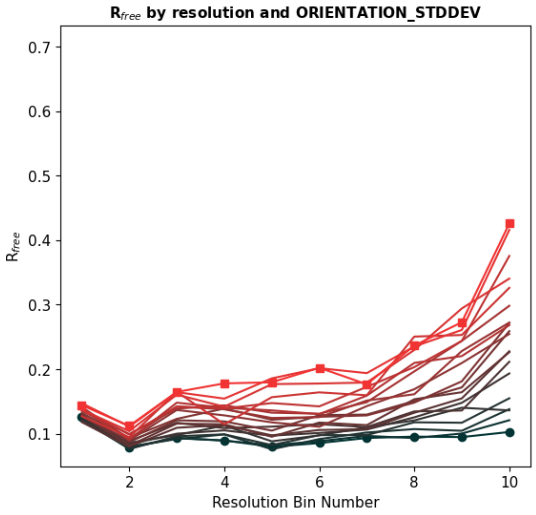

R_{work} by resolution and BEAM_STDDEV
0.0
0.0182
0.0364
0.0546
0.0728
0.0911
0.1093
0.1275
0.1457
0.1639
0.1822
0.2004
0.2186
0.2368
0.2550
0.2733
0.2915
0.3097
0.3279
0.3462
R_{work}
Resolution Bin Number

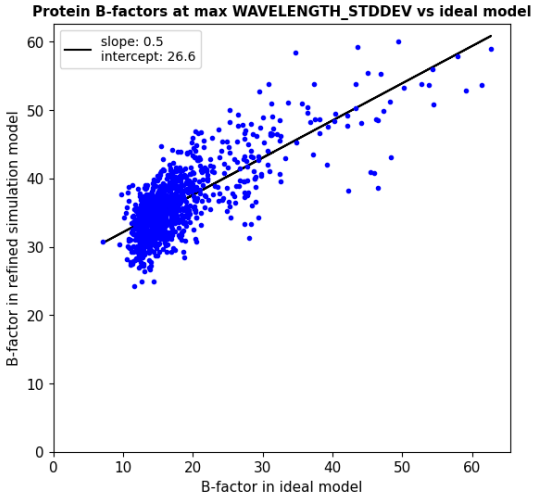

R_{free} by resolution and BEAM_STDDEV
R_{free}
Resolution Bin Number

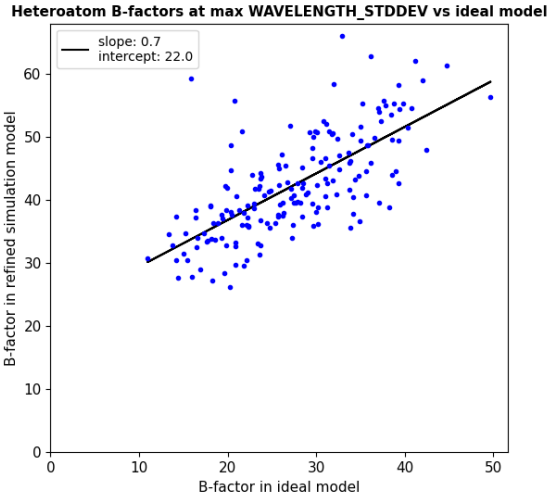

R_{work} by resolution and CELL_STDDEV
0.0
0.1132
0.2265
0.3398
0.4530
0.5663
0.6796
0.7929
0.9061
1.0194
1.1327
1.2460
1.3592
1.4725
1.5858
1.6991
1.8123
1.9256
2.0389
2.1522
R_{work}
Resolution Bin Number

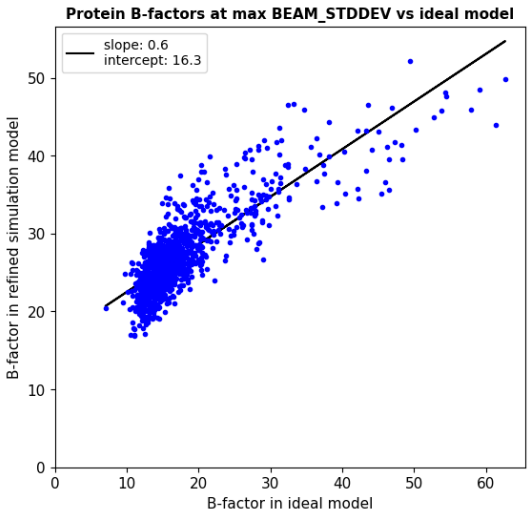

R_{free} by resolution and CELL_STDDEV
R_{free}
Resolution Bin Number

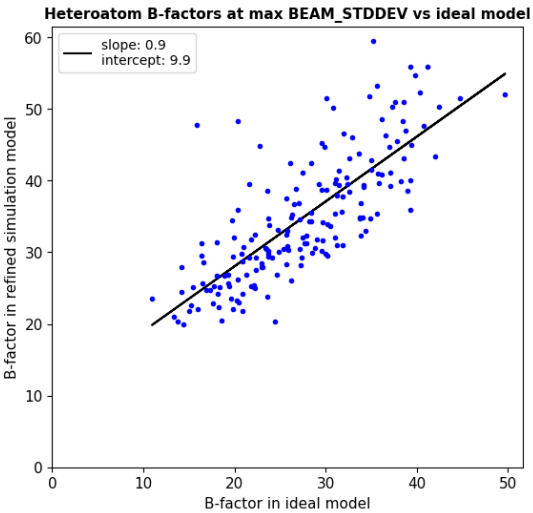

R_{work} by resolution and ORIENTATION_STDDEV
0.0
0.0468
0.0936
0.1405
0.1873
0.2342
0.2810
0.3279
0.3747
0.4216
0.4684
0.5153
0.5621
0.6090
0.6558
0.7027
0.7495
0.7964
0.8432
0.8901
R_{work}
Resolution Bin Number

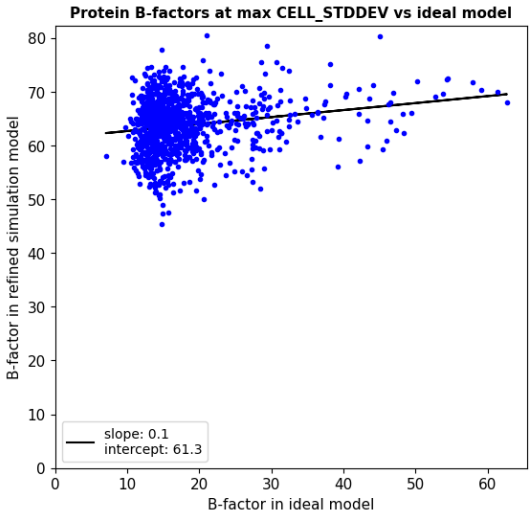

R_{free} by resolution and ORIENTATION_STDDEV
R_{free}
Resolution Bin Number

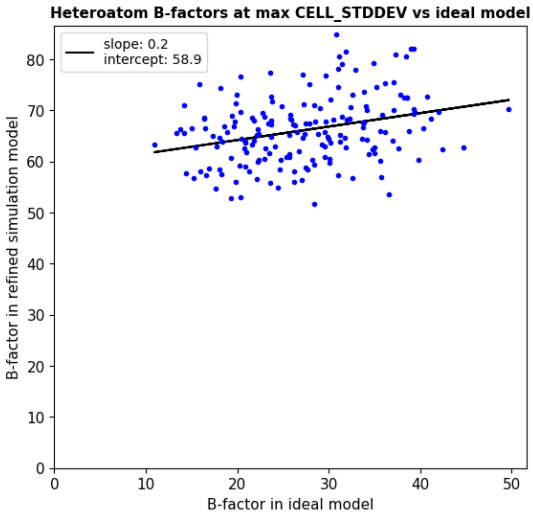

Protein B-factors at max WAVELENGTH_STDDEV vs ideal model
slope: 0.5
intercept: 26.6
B-factor in refined simulation model
B-factor in ideal model

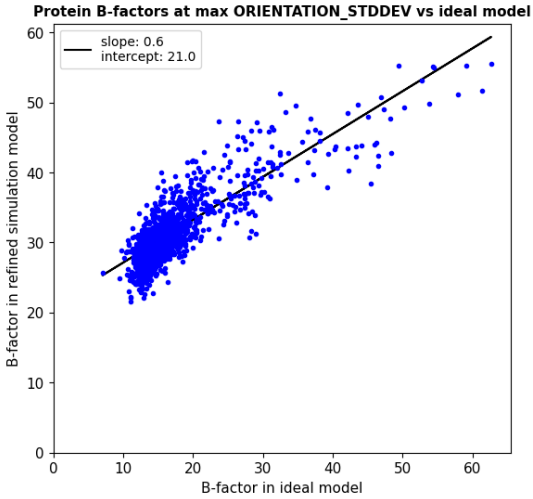

Heteroatom B-factors at max WAVELENGTH_STDDEV vs ideal model
slope: 0.7
intercept: 22.0
B-factor in refined simulation model
B-factor in ideal model

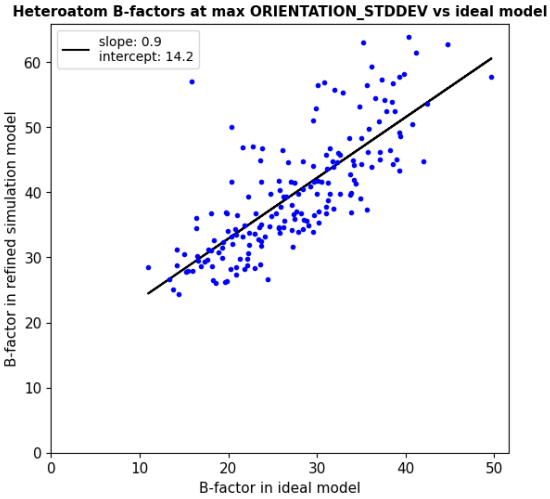

Protein B-factors at max BEAM_STDDEV vs ideal model
slope: 0.6
intercept: 16.3
B-factor in refined simulation model
B-factor in ideal model

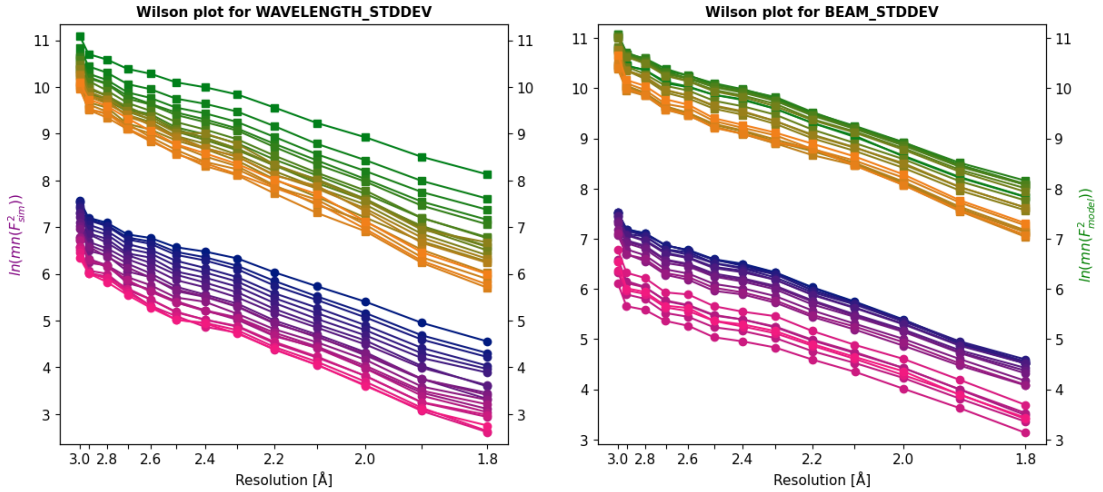

Heteroatom B-factors at max BEAM_STDDEV vs ideal model
slope: 0.9
intercept: 9.9
B-factor in refined simulation model
B-factor in ideal model


Protein B-factors at max CELL_STDDEV vs ideal model
B-factor in refined simulation model
B-factor in ideal model
slope: 0.1
intercept: 61.3

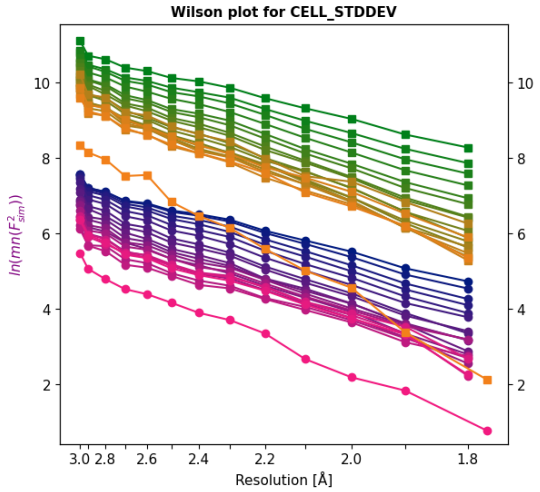

Heteroatom B-factors at max CELL_STDDEV vs ideal model
B-factor in refined simulation model
B-factor in ideal model
slope: 0.2
intercept: 58.9

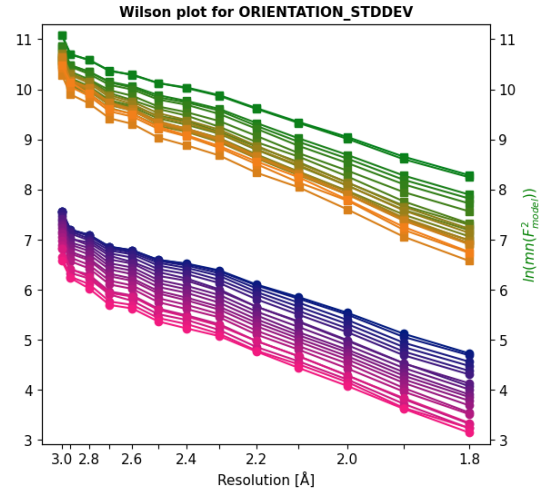

Protein B-factors at max ORIENTATION_STDDEV vs ideal model
B-factor in refined simulation model
B-factor in ideal model
slope: 0.6
intercept: 21.0

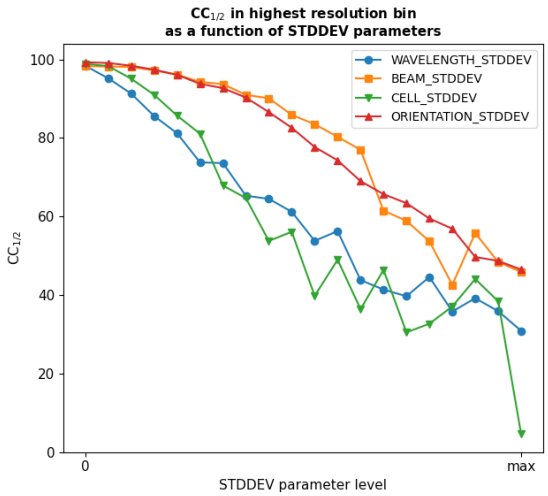

Heteroatom B-factors at max ORIENTATION_STDDEV vs ideal model
B-factor in refined simulation model
B-factor in ideal model
slope: 0.9
intercept: 14.2

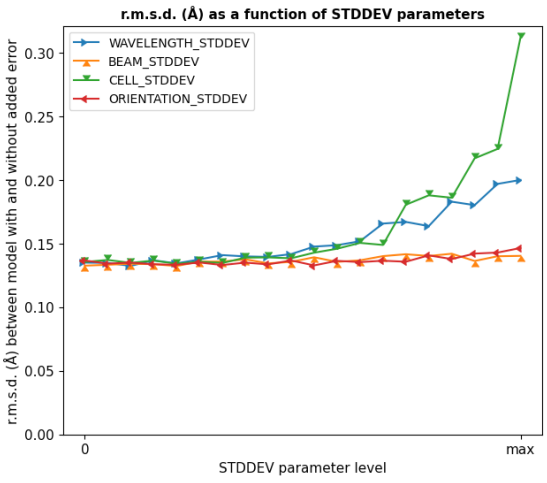

Wilson plot for WAVELENGTH_STDDEV
Resolution [Å]

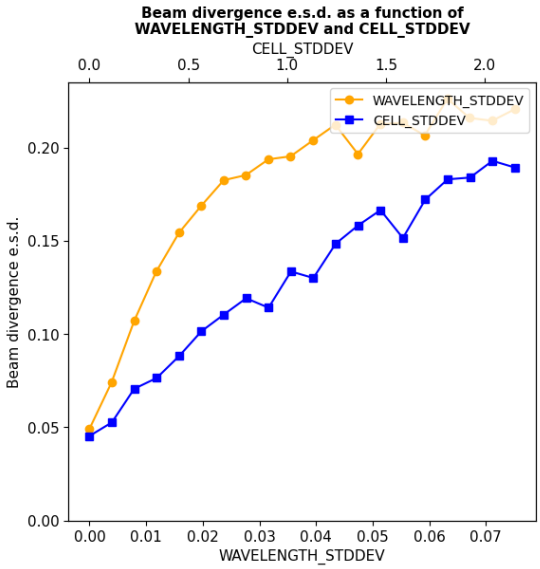
Wilson plot for BEAM_STDDEV
Resolution [Å]

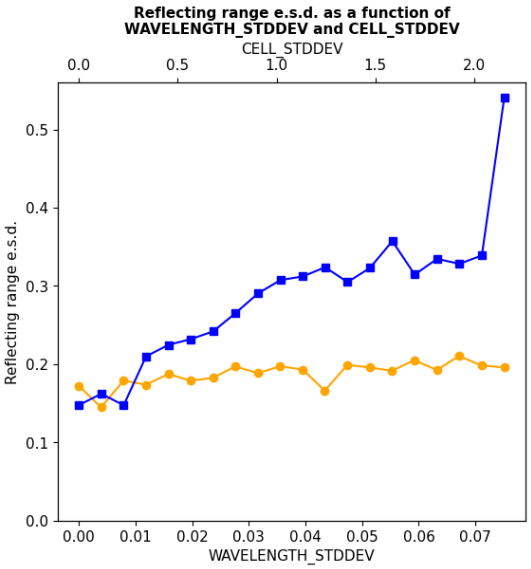

Wilson plot for CELL_STDDEV
Resolution [Å]

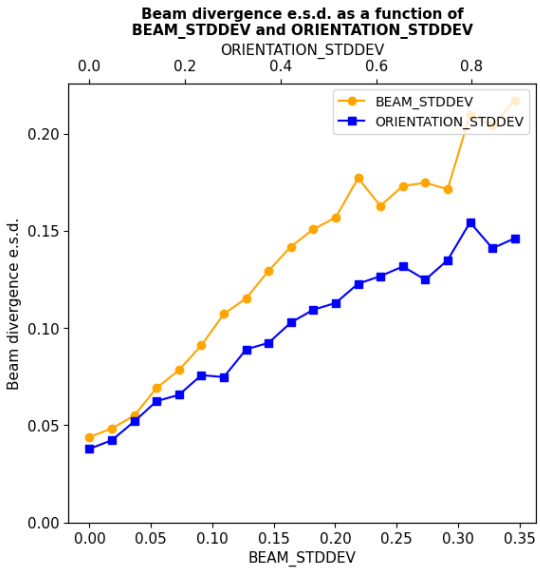

Wilson plot for ORIENTATION_STDDEV
Resolution [Å]

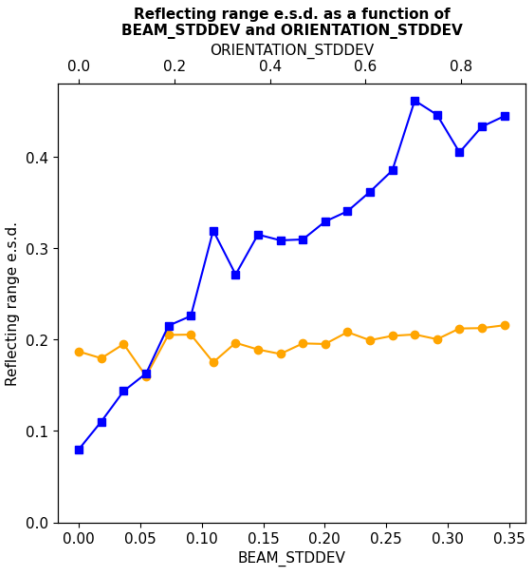

CC1/2 in highest resolution bin
as a function of STDDEV parameters
WAVELENGTH_STDDEV
BEAM_STDDEV
CELL_STDDEV
ORIENTATION_STDDEV
STDDEV parameter level

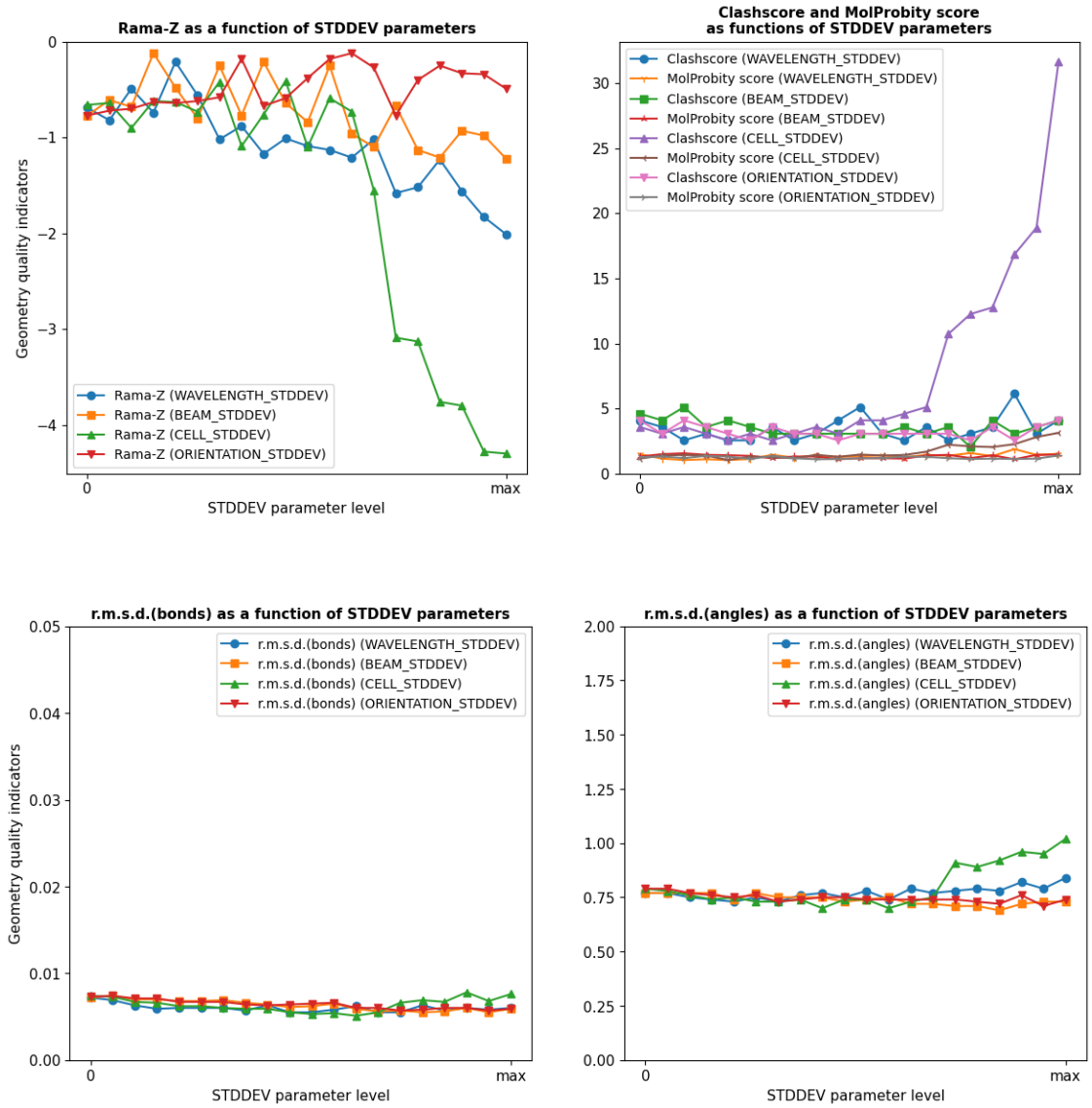

r.m.s.d. (Å) as a function of STDDEV parameters
WAVELENGTH_STDDEV
BEAM_STDDEV
CELL_STDDEV
ORIENTATION_STDDEV
r.m.s.d. (Å) between model with and without added error
STDDEV parameter level


Beam divergence e.s.d. as a function of
WAVELENGTH_STDDEV and CELL_STDDEV
CELL_STDDEV
WAVELENGTH_STDDEV
CELL_STDDEV
Beam divergence e.s.d.
WAVELENGTH_STDDEV


Reflecting range e.s.d. as a function of
WAVELENGTH_STDDEV and CELL_STDDEV
CELL_STDDEV
Reflecting range e.s.d.
WAVELENGTH_STDDEV


Beam divergence e.s.d. as a function of
BEAM_STDDEV and ORIENTATION_STDDEV
ORIENTATION_STDDEV
BEAM_STDDEV
ORIENTATION_STDDEV
Beam divergence e.s.d.
BEAM_STDDEV


Reflecting range e.s.d. as a function of
BEAM_STDDEV and ORIENTATION_STDDEV
ORIENTATION_STDDEV
Reflecting range e.s.d.
BEAM_STDDEV



Acknowledgements

I am grateful to Prof. Kay Diederichs for his guidance and attention throughout this project. I also appreciate the discussions with, and advice from my fellow group members Dr. Karsten Schäfer, Dr. Greta Assmann, Kathy Su, and Jill von Velsen.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., ... & Adams, P. D. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography*, *68*(4), 352-367.

Brünger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, *355*(6359), 472-475.

Brünger, A. T. (1997). Free R value: Cross-validation in crystallography. *Methods in enzymology,* 277, 366-396. Academic Press.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., ... & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*(1), 12-21.

Cowtan, K. (2008). Wilson plot. *Retrieved from http://www.ysbl.york.ac.uk/~cowtan/ccp4wiki/wiki60.html.*

Diederichs, K. (2009). Simulation of X-ray frames from macromolecular crystals using a ray-tracing approach. *Acta Crystallographica Section D: Biological Crystallography*, *65*(6), 535-542.

Diederichs, K. (2010). Quantifying instrument errors in macromolecular X-ray data sets. *Acta Crystallographica Section D: Biological Crystallography, 66(6),* 733-740.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography, 66(4),* 486-501.

Evans, P. R., & Murshudov, G. N. (2013). How good are my data and what is the resolution?. *Acta Crystallographica Section D: Biological Crystallography, 69(7), 1204-1214.*

GIMP Development Team. (2019). GIMP. *Retrieved from https://www.gimp.org*

GNU, P. (2007). *Free Software Foundation. Bash (4.2.46(2)-release)[Unix shell program].*

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362.

Hoffer, M. (2013). XDS-Viewer. *Retrieved from http://xds-viewer.sourceforge.net/*

Holton, J. M. (2019). Challenge data set for macromolecular multi-microcrystallography. *Acta Crystallographica Section D: Structural Biology*, *75*(2), 113-122.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering, 9(3),* 90-95.

ImageMagick Development Team (2021). ImageMagick. *Retrieved from https://imagemagick.org.*

IUCr Commission on Crystallographic Nomenclature (2017). R Factor. *Online Dictionary of Crystallography, diction*ary.iucr.org/R_factor.

Kabsch, W. (2010). Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D: Biological Crystallography*, *66*(2), 133-144.

Kabsch, W. (2010). XDS. *Acta Crystallographica Section D*, *66*, 125-132.

Karplus, P. A., & Diederichs, K. (2012). Linking crystallographic model and data quality. *Science*, *336*(6084), 1030-1033.

Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, *60*(12), 2256-2268.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., ... & Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology*, *75*(10), 861-877.

Nanao, M. H., Sheldrick, G. M., & Ravelli, R. B. (2005). Improving radiation-damage substructures for RIP. *Acta Crystallographica Section D: Biological Crystallography*, *61*(9), 1227-1237.

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347-352), 240-242.

Ramachandran, G. N., Ramakrishnan, C., Sasisekharan, V. (1963). "Stereochemistry of polypeptide chain configurations." *J. mol. Biol* 7, 95-99.

Rupp, B. (2009). Debye-Waller factor, atomic displacement, and B-factor. *Biomolecular crystallography: principles, practice, and application to structural biology.* Garland Science, 261-262.

Rupp, B. (2009). R-values and correlation coefficients. *Biomolecular crystallography: principles, practice, and application to structural biology.* Garland Science, 330.

Rupp, B. (2009). Wilson plots and initial scaling. *Biomolecular crystallography: principles, practice, and application to structural biology.* Garland Science, 355-357.

Sobolev, O. V., Afonine, P. V., Moriarty, N. W., Hekkelman, M. L., Joosten, R. P., Perrakis, A., & Adams, P. D. (2020). A global Ramachandran score identifies protein structures with unlikely stereochemistry. *BioRxiv 2020.03.26.010587*.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Van Rossum, G. (2020). The Python Library Reference, release 3.8. 2. *Python Software Foundation*, 36.

Weiss, M. S., Palm, G. J., & Hilgenfeld, R. (2000). Crystallization, structure solution and refinement of hen egg-white lysozyme at pH 8.0 in the presence of MPD. *Acta Crystallographica Section D: Biological Crystallography*, *56*(8), 952-958.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., ... & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, *27*(1), 293-315.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., ... & Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 235-242.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., ... & Richardson, D. C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology*, *285*(4), 1711-1733.